STAT 1793: Course Notes

Introduction to Probability and Statistics I

Dylan Spicker

2023-12-30

Table of contents

Preface

What are these notes?

These notes are being written to supplement the Winter 2024 offering of STAT 1793 at the University of New Brunswick on the Saint John campus. These are intended to provide the complete set of notes required to be successful in the course, along with a series of examples. Note that, if you are a student in the offering of STAT 1793, the specific material that we cover and focus on may be a subset of these notes: we will be crafting our path through them together, as a class, and as such, not all material will be equally relevant. It is important to check the course website and to listen in lecture to ensure that you are always aware of what is being emphasized, what will be covered on the assessments, and so forth.

Moreover, it is important to note that these notes are a work-in-progress. As of today (January 5, 2024), the material through to Chapter 5 has been made up-to-date. Further chapters will be added throughout term. Additionally, some of the completed chapters may be revised in an attempt to add clarity, fix mistakes or typos, and to better format or sequence the material. If at any point you have comments, questions, suggestions, or corrections to these notes, please do reach out! They are intended to be a helpful resource to ensure your success in the course, and to help your probability and statistics skills into the future.

One final note is that our coverage of concepts throughout STAT 1793 reflects my personal biases and proclivities towards what is important in a first course in probability and statistics. Compared to others I likely have a larger emphasis on probability, leaving concepts more tied to statistics to a second course (e.g., STAT 2793). This is reflected both in the ordering of material (we start at Probability not at Statistics), and in the weight that each topic will be given. It is my belief that you need to be able to fluently speak the language of probability in order to study statistics, and that is where we will put our focus. However, this does mean that if you have impressions or content from another offering of the course, they may be less well-suited to the specific range of topics we will cover. The material is all the same, and if they help you learn, by all means; it is however important to emphasize that what is highlighted will likely be different.

Using These Notes

These notes are best viewed online. You can access them from anywhere at https://dylanspicker.github.io/STAT1793-course-notes/index.html. Viewing online will allow you to take advantage of interactivity (such as through code blocks and solutions hiding), the search feature (visible in the left hand margin), as well as better overall formatting. There is a version of these notes in PDF format available. To download them, press the PDF icon button () next to the note title in the left hand bar. The PDF notes contain all the same content, and will be updated alongside the web notes. The formatting of the PDF document may be less visually appealing compared to the web notes. Though an effort will be made to ensure that everything remains legible, and well-formatted, the aesthetics of these notes will be a lower priority than those available on the web.

To successfully use these notes, I would recommend a several step procedure:

1. Prior to the lecture where the material will be covered, read through the sections that are to be covered. For examples, do not read the solution, and instead try to think of how the example can be solved (perhaps trying to solve it yourself).
2. Attend lecture and be engaged. Our lectures will follow along with the course notes rather closely, including doing as a class many of the example problems. As a result, this will reemphasize what you have read.
    - Note, by reading *before* you come to class you will be seeing the material for a second time, allowing you to better process it.
    - Additionally, this will allow you to better follow the parts that were confusing in the notes, and be more prepared to ask questions of me throughout the lecture.
3. Return to the notes after class to ensure that everything has made sense. At this point try solving the example problems directly without referencing the solutions at all.
    - Only once you've gotten a solution to the problem should you check if it is correct. When checking, only look as far as any mistakes that have been made, and then try the problem again.
    - Note: struggling with problems is a key component of learning in any mathematical discipline. It is only by struggling to understand how the pieces fit together that you will develop a proper understanding for yourself.
    - Looking at solutions will often undermine your own understanding. It is a much different process to look at a solution and understand why it is correct, than it is to determine a solution for yourself.
4. Ask questions, attend office hours, and send emails as is required to ensure you are fully successful in the course!

Throughout the notes there will be references to computer code in a programming language known as R. R is the most common language used for statistics by practicing statisticians, and

it is a great utility for many of the types of problems we will be considering in this course. You are not expected to be able to write R code throughout the course, however, there is plenty of material throughout the notes to ensure that you will be able to read and run R code. This will be required for solving certain problems, and will make the process of statistics far nicer to engage with. If you are using the web notes this code can all be run directly in the browser. If you are using the PDF notes, the code and output is still provided. In either case, I would suggest getting R running on your own device, if possible. It is free to download and install. I would also suggest installing R Studio, which is a program specifically designed for writing and running R code with a lot of nice features.

Some Mathematical Background

These notes (and this course more broadly) require high school mathematics to complete. There is no calculus specifically required. The assumption is that you will be comfortable manipulating mathematical expressions, solving basic equations, using a calculator, and so forth. The following are a handful of topics which will come up throughout the course that are assumed to be known, presented here in brief as a quick reference.

Logarithms and Exponent Rules

Recall that if $a^x = b$, then we can solve for x through the use of logarithms. Specifically, $x = \log_a(b)$, where we read this as "the logarithm with base a of b." The expression means that "x is the exponent we put on a to get b." In this course[1] we use the notation $\log(x)$ to represent the **natural logarithm**. The natural logarithm is $\log_e(x)$, where $e \approx 2.71828$ is Euler's number. The **exponential function** is given by $e^x = \exp(x)$. Which notation is used depends on the specific setting, but both are equivalent.

[1] And in most math courses beyond a certain point.

Recall further the following exponent and log rules:

$$\begin{aligned} a^b \times a^c &= a^{b+c} \\ a^b \times c^b &= (ac)^b \\ a^{-b} &= \frac{1}{a^b} \\ (a^b)^c &= a^{bc} \\ a^0 &= 1 \\ \log(ab) &= \log(a) + \log(b) \\ \log(a^b) &= b\log(a) \\ \log(1) &= 0 \\ \log(e) &= 1. \end{aligned}$$

Summation and Product Notation

If we wish to add up a series of numbers, $x_1, x_2, \ldots, x_n$ then we can use summation notation to record this. Specifically,

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i.$$

Here i is an indexing variable just used to keep track of which number we are currently working on. Note that the following quantities will hold

$$\begin{aligned} \sum_{i=1}^{n}(x_i + y_i) &= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i - y_i) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} cx_i &= c\sum_{i=1}^{n} x_i. \end{aligned}$$

We can also use a similar shorthand to discuss repeated products. That is, if we want to multiply $x_1, x_2, \ldots, x_n$, then we can write

$$x_1 \times x_2 \times \cdots \times x_n = \prod_{i=1}^{n} x_i.$$

Again, i is used as an indexing variable to keep track of what value is currently being multiplied.

The following results hold

$$\prod_{i=1}^{n} x = x^n$$
$$\prod_{i=1}^{n} b^{x_i} = b^{\sum_{i=1}^{n} x_i}$$
$$\prod_{i=1}^{n} cx_i = c^n \prod_{i=1}^{n} x_i.$$

Other Important Points

The following are some additional points that are useful to know, but which do not necessarily fit together nicely.

- For any two values x and y, we get $(x+y)^2 = x^2 + 2xy + y^2$.
- The integers are the set of all whole numbers, $\{0, \pm 1, \pm 2, \dots\}$. We will often denote the set of integers using $\mathbb{Z}$.
- The natural numbers are the set of all counting numbers, either $\{0, 1, 2, 3, \dots\}$ or else $\{1, 2, 3, 4, \cdots\}$ depending on the source. We often denote the set using $\mathbb{N}$.
- The real numbers are the set of all numbers that you think of when thinking of numbers. They include the integers, and any fractions, and anything that cannot be written as a fraction (like π and e) as well. We denote the real numbers as $\mathbb{R}$.
- Intervals of real numbers are denoted as (a, b) or $[a, b]$ (or a mixture, giving $[a, b)$ or $(a, b]$). A round bracket means that the endpoint is **not** included in the set, where a square bracket means that it is. That is, a value of x is in (a, b) if $a < x < b$, where it is in $[a, b]$ if $a \leq x \leq b$.
- We write "x is in the interval (a, b)" mathematically using the $\in$ symbol. Specifically, $x \in (a, b)$.
- Infinity (∞) is not a number. Rather, it is a statement regarding the lack of a bound. When we say that something is infinity, or infinite, we simply mean that it is larger than any of the natural numbers. If you were to be given any number, then infinity will be larger than that.
- **Limits:** note that, while limits are from calculus, they are useful for formally understanding a few concepts in class. The limit of a function, $f(x)$, is simply the value that the function approaches as x approaches a specific point. So, for instance, $\lim_{x \to a} f(x)$ is the value that $f(x)$ approaches as x approaches a. If $f(a)$ exists then it is simply $f(a)$. However, $f(a)$ may not exist, at which point, we look at where the function is tending to.
- **Infinite Limits:** Of specific importance to us are infinite limits. Specifically,

$$\lim_{x \to \infty} f(x)$$

captures the beahviour of the function $f(x)$ as x gets larger and larger. In many cases, $f(x)$ will approach some specific value, which we assign to be the limiting value of the function.

Additional Resources

The following are helpful resources which may be used to supplement the material in the class. If you have any suggested resources, please feel free to send them to me, and I will include them here.

1. Open Intro Statistics.
2. Probability Course.
3. Introduction to R.
4. Learning Statistics with R.

Part I

Part 1: Probability

1 Introduction to Probability

1.1 What is Probability?

At its core, statistics is the study of uncertainty. Uncertainty permeates the world around us, and in order to make sense of the world, we need to make sense of uncertainty. The language of uncertainty is **probability**. Probability is a concept which we all likely have some intuitive sense of. If there was a 90% probability of rain today, you likely considered grabbing an umbrella. You are not likely to wager your life savings on a game that has only a 1% probability of paying out. We have a sense that probability provides a measure of *likelihood*. Defining probability mathematically is a non-trivial task, and there have been many attempts to formalize it throughout history. While we will spend a good deal of time formalizing notions of probability in this course, we first pause to emphasize the familiarity with probability that you are likely starting with.

Suppose that two friends, Charles and Sadie, meet for coffee once a week. During their meetings they have wide-ranging, deep, philosophical conversations spanning across many important topics.[1] Beyond making progress on some of the most pressing issues of our time, Charles and Sadie each adore probability. As a result, at the end of each of their meetings, they play a game to decide who will pay. The game proceeds by having them flip a coin three times. If two or more heads come up Charles pays, and otherwise Sadie pays.

We can think about the strategy that they are using here and *feel* that this is going to be "fair". With two or more heads, Charles pays. With two or more tails, Sadie pays. There always has to be either two or more heads *or* two or more tails, and each is equally likely to come up[2]. The outcome of their game is uncertain before it begins, but we know that in the long run neither of the friends is going to be disadvantaged relative to the other. We can say that the probability that either of them pays is equal. It's 50-50. Everything is balanced.

Now imagine that one day, in the middle of their game, Charles gets a very important phone call[3] and leaves abruptly after the first coin has been tossed. The first coin toss showed a heads.

[1] For instance they ask "do we all really see green as the same colour?" or "why is it that 'q' comes as early in the alphabet as it does? it deserves to be with 'UVWXYZ'?"

[2] There has been some recent literature, see Bartoš et al. (2023), which may suggest that a coin is ever so slightly more likely to land on the same side it started on, perhaps undermining this assertion.

[3] Someone has just pointed out the irony in the fact that there is no synonym for synonym. Technically, there are the words *metonym* and *poecilonym* and *polyonym*, but these are rarely used and Charles would wager that there is a very high probability you have never seen them.

Sadie, recognizing the gravity of the phone call, pays for the both of them, while realizing that Charles was well on the way to having to pay.

Example 1.1 (Basic Probability Enumeration)**.** What is the probability that Sadie would have had to pay in the aforementioned scenario? That is, assuming that the first coin shows a head, what is the probability that at least two heads are shown on the first three coin tosses?

Solution

Sadie figures that any coin toss is equally likely to show heads or tails. because the first coin showed heads, then there are four possible sequences that could have shown up:

1. H, H, H;
2. H, H, T;
3. H, T, H;
4. H, T, T.

In three of these situations ((1) H, H, H, (2) H, H, T, and (3) H, T, H) there are two heads and so Charles would have to pay. In one of them there are two tails, and so Sadie would have to pay. As a result, Charles would have to pay in 3 of the 4 (with probability 0.75) and Sadie in 1 of the 4 (with probability 0.25).

Looking at Example 1.1, we can see that Sadie should likely have not paid. Only one out of every four times would Sadie have had to pay, given the first coin being heads. However, we can not be certain that, had all three tosses been observed, Sadie would *not* have paid. It is possible that we would have observed two tails, making her responsible for the bill. This possibility happens one time out of four, which is more likely than the probability of rolling a four on a six-sided die[4]. Fours are rolled on six-sided dice quite frequently[5], and so it is not all together unreasonable for her to have paid.

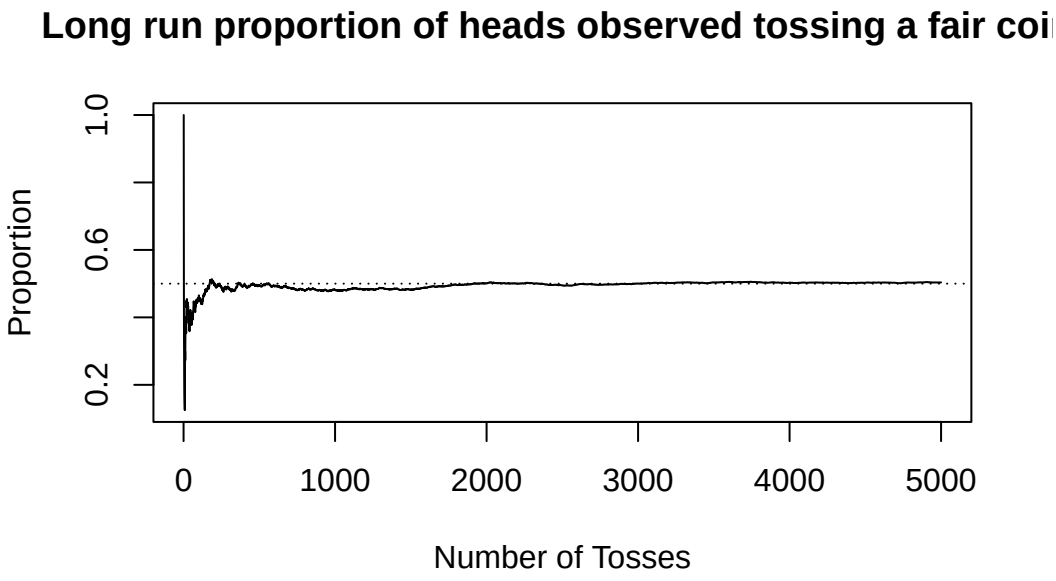
This seemingly simple concept is the core of probability. Probability serves as a method for quantifying uncertainty. It allows us to make numeric statements regarding the set of outcomes that we can observe, by quantifying the frequency with which we expect to observe those outcomes. Probability does not *remove* the uncertainty. We still need to flip the coin or roll the die to know what will happen. All probability gives us is a set of tools to quantify this uncertainty. These tools are critical for decision making in the face of the ever-present uncertainty around us.

[4]This event happens one time out of every six.
[5]Again, about one out of every six rolls

1.2 How to Interpret Probabilities (like a Frequentist)

We indicated that, intuitively, probability is a measure of the frequency with which a particular outcome occurs. This intuition can be codified exactly with the **Frequentist interpretation** of probability. According to the Frequentist interpretation (or frequentism, as it is often called), probabilities correspond to the proportion of time any event of interest is[6] actually occurs in the long run. For a Frequentist, you imagine yourself performing the experiment you are running over, and over, and over, and over, and over again. Each time you answer "did the event happen?" and you count up those occurrences. As you do this more and more and more, wherever that proportion lands corresponds to the probability.

Figure 1.1: This plot simulates the repeated tossing of a coin. The x-axis represents the number of coins being tossed, and the y-axis plots the proportion of times that heads has shown up cumulatively over the tosses thus far. We can see in the long run that this proportion tends towards 0.5.



To formalize this mathematically, we first define several important terms.

Definition 1.1 (Experiment)**.** Any action to be performed whose outcome is not (or cannot) be known with certainty, before it is performed.

[6]Be that flipping a heads, rolling a four, or observing rain on a given day.

Definition 1.2 (Event [Informally])**.** A specified result that may or may not occur when an experiment is performed.

Suppose that an experiment is able to be performed as many times as one likes, limited only by your boredom. If you take k_N to represent the number of times that the event of interest occurs when you perform the experiment N times, then a Frequentist would define the probability of the event as

$$\text{probability} = \lim_{N\to\infty} \frac{k_N}{N}.$$

This course does not assume that you have any familiarity with calculus, and yet, this definition relies on limits, a concept taken directly from calculus. However, we will not actually require the ability to work with limits for this course. Instead, when you see a statement of the form

$$\lim_{x\to\infty} f(x),$$

simply think "what is happening to the function $f(x)$ as x grows and grows (off to ∞)?"

In practice this means that, in order to interpret probabilities, we think about repeating an experiment many, many times over. As we do that, we observe the proportion of times that any particular outcome occurs, and take that to be the defining relation for probabilities. The reason that we say the probability of flipping a heads is 0.5 is because if we were to sit around and flip a coin[7] over, and over, and over again, then in the long-run we would observe a head[8] in 0.5 of cases.

Example 1.2 (Probability Interpretation)**.** How do we interpret the statement "the probability that Sadie would have had to pay, given a head on the first toss, is 0.25"? Recall that in the game, they toss three coins, and Sadie pays if two of them show tails.

Solution

This statement means that, if Sadie were to repeatedly be in the situation where one head has shown and there are two coins left to toss, then in 0.25 of these situations (in the limit, as this is repeated an infinite number of times) will end up showing two tails.

Many situations in the real world are not able to be run over and over again. Think about, for instance, the probability that a particular candidate wins in a particular election. There is uncertainty there, of course, but the election can only be run once. What then? There are several ways through these types of events.

First, we can rely on the **power of imagination**. There is nothing stopping us from envisioning the hypothetical possibility of running the election over, and over, and over, and over again. If we step outside of reality for a moment, we can ask "if we could play the day of the

[7] Our experiment.
[8] Our event.

election many, many, many times, what proportion of those days would end with the candidate being elected?" If we say that the candidate has a 75% chance of being elected, then we mean that in 0.75 of those imagined worlds, the candidate wins. It is crucial to stress that in our imagination here, we need to be thinking about the **exact same day** over and over again. We cannot imagine a different path leading to the election, different speeches being given in advance, or different opposition candidates. If we start from the same place, and play it out many times over, what happens in each of those worlds?

This repeated imagining is not for everyone. As a result, alternative proposals to the interpretation of probability have been made. Most popularly, the **Bayesian interpretation** has recently become prominent. To Bayesians, probability is a measure of subjective belief. To say that there is a 50% chance of a coin coming up is a statement about one's knowledge of the world. Typically, coins show up heads half the time, so that's our belief about heads. The Bayesian view, built around subjective confidence in the state of the world, can be formalized mathematically as well. A Bayesian considers the *prior evidence* that they have about the world[9] and combine this with current observations in order to update their subject beliefs, balancing these two sources of information.

Remark (Bayesian Probabilities and Belief Updating). Suppose that a Bayesian is flipping a coin. Before any flips have been made the Bayesian understandably believes that the coin will come up heads 50% of the time. However, when the coin starts to be flipped, the observations are a string of tails in a row.

After having flipped the coin five times, the individual has observed five tails. Of course, it is totally possible to flip a fair coin five times and see five tails, but there is a level of skepticism growing.

After 10 flips, the Bayesian has still not seen a head. At this point, the subjective belief is that there is likely something unfair about this coin. Even though the experiment started with a baseline assumption that the coin was fair the Bayesian no longer believes that the next flip will be a head.

As this goes on, you can imagine the Bayesian continuing to update their view of the world. To them, the probability is an evolving concept, capturing what was believed and what has been observed.

For the election example, the Bayesian interpretation is somewhat easier to think through. To say that a candidate has a 75% chance of winning the election means that "based on everything that has been observed, and any prior beliefs about the viability of the candidates, the subjective likelihood that the candidate wins the election is 0.75". If we disagree about prior beliefs, or have experienced different pieces of information, then we may disagree on the probability. That is okay.

[9] Any relevant evidence that has previously been collected.

In this course, we focus on the Frequentist interpretation. This is, in part, because Frequentist probability is an easier introduction to the concepts that are necessary for grasping uncertainty. Additionally, there is some research to suggest that Frequentist interpretations are fairly well understood by the general public [10]. However, it is important to know and recognize that there is a world beyond Frequentist Probability and Statistics, one which can be very powerful once it is unlocked. [11]

If probability measures the long term proportion of times that a particular event occurs, how can we go about computing probabilities? Do we require to perform an experiment over and over again? Fortunately, the answer is no. The tools of probability we will cover allow us to make concrete statements about the probabilities of events without the need for repeated experimental runs. However, before dismissing the idea of repeatedly running an experiment at face value, it is worth considering a tool we have at our disposal that renders this more possible than it has ever been: computers.

1.3 R Programming for Probability and Statistics

Throughout this course we will make use of a programming language for statistical computing, called R. Classically, introductory statistics courses involved heavy computation of particular quantities by hand. The use of a programming language (like R) frees us from the tedium of these calculations, allowing for a deeper focus on understanding, explanation, decision-making, and complex problem solving. This is **not** to say we will *never* do problems by hand[12], however, we will emphasize the use of statistical computing often. While you will not be expected to write R programs on your own, you will be expected to read simple scripts, make basic modifications, and to run code that is given to you.

Throughout these course notes, where relevant, R code will be provided to demonstrate the ideas being discussed. It may be useful to have R open alongside the notes to ensure that you can get the same results that are printed throughout. In this section we will cover some of the very basics of using R, and reading R code. If you are interested, there are plenty of resources to becoming a more proficient R programmer. This is a skill that will benefit you not only in this course, but in many courses to come, and far beyond your university training. If you have any programming in your background, R is a fairly simple language to learn; if not, R can be quite beginner friendly.

[10] See, for instance Gigerenzer et al. (2005), where they study how people interpret the statement "there is a 30% chance of rain tomorrow." Interestingly, most people can convert this into a frequency statement (3 out of 10, say), even if the specific meaning is sometimes lost. They conclude that there are other issues in attempting to understanding this statement, issues which we will address later on.

[11] More on this in later years, if you so desire! Come have a chat, if that sounds interesting.

[12] To the contrary, some amount of by-hand problem solving helps to solidify these concepts.

1.3.1 Basic Introduction to R Programming

When programming, the basic idea is that we are going to write instructions in a **script** which we will tell our computer to execute. These instructions are[13] performed one-by-one, from the top to the bottom of the script. We can have instructions which operate on their own, or which interact with previous (or future) instructions to add to the complexity. The trick with programming then is to determine which actions you need the computer to perform, in which order, to accomplish the task that you are setting out to do.

To begin, we may consider a very simple R program, one which uses the programming language as a basic calculator.

```
5 + 3 - (10*2) + 8^(25/3)
## [1] 33554420
```

Here, we ask the computer to perform some simple arithmetic operations. We use `+` for addition, `-` for subtraction, `*` for multiplication, `/` for division, and `^` for exponentiation. With the use of parentheses, any expressions relating to these basic operations can be performed. Note that here the result is simply output after it is computed. Try modifying the exact expressions being computed, allowing you to feel comfortable with these types of mathematical operations.

```
5 + 3 - (10*2) + 8^(25/3)
8 * 5
## [1] 33554420
## [1] 40
```

Here, we have two lines of math running with simple operations. Each is simply output when it is computed. These two results have no ability to interact with one another, and if we were to add more and more lines beneath, the same would continue to happen. If we want different commands to be able to interact with one another, we need a method for storing the results. To do so, we can define **variables**. In R, to define a variable, we use the syntax `variable_name <- variable_value`. We can choose *almost* anything that we want for the variable name[14] and the variable value can also be of many types.[15] The arrow between the two is the **assignment operator** and it simply tells R to assign the `variable_value` to be accessible from the `variable_name`.

[13] Typically. There are some exceptions to this, but if this is your first time programming, you need not worry about that!

[14] The variable name must start with a letter and can be a combination of letters, numbers, periods, and underscore.

[15] more on this later

```
my_5 <- 5
my_8 <- 8
my_5
my_8
## [1] 5
## [1] 8
```

In this code we assign the variable `my_5` to contain the value 5, and the variable `my_8` to contain the value 8. We can output these as expressions themselves, simply by typing the variable name. Simply outputting these variables is not of particular interest, however, we can use the variables in later statements by simply including the variable name in them.

```
my_5 <- 5
my_8 <- 8
my_5*my_8
## [1] 40
```

Here, instead of simply outputting the variables, we multiply them together. We could have used `5 * 8` in this case for the same result, however, we are afforded a lot more flexibility with this approach. Much of this flexibility comes from our capacity to *change* the values of variables over time. Consider the following script, and try to understand why the output is the way that it is.

```
my_5 <- 5
my_8 <- 8
my_5*my_8

my_5 <- 10
my_5*my_8
## [1] 40
## [1] 80
```

At the top point in the script, before the first `my_5*my_8` call, The variable `my_5` has the value `5`. However, after this is called, the value is updated to be `10`. Then, when we call `my_5*my_8` again, this is now simplified to `10 * 8`, giving the result. Perhaps more importantly, we can take the value of an expression and assign that to a variable itself.

```
my_5 <- 5
my_8 <- 8
result <- my_5*my_8
result
## [1] 40
```

Here, we define another variable. This time, `result` now contains the result of multiplying our previous two variables. Thus, when we output it, we get the same value. Take a moment to read through the following script, and try to understand what will happen at the end.

```
my_5 <- 5
my_8 <- 8
result <- my_5*my_8

my_5 <- result
my_8 <- my_5
result <- my_5 * my_8

result
```

The result is 1600. Why? We can read through this step-by-step in order to understand this. First we set our variables to have the values of `5` and `8`. Then, `result` is made to be the product of our two variables, which in this case is `40`. After that, we set the value of `my_5` to be the same as the value of `result`, which gives `my_5` equal to `40`. At this point, `result` equals `40`, `my_5` equals `40`, and `my_8` is equal to `8`. The next line updates `my_8` to be the same as the value of `my_5`, which we just clarified was `40`. As a result, all of the variables we have defined now take on the value of `40`. The final line before output, `result <- my_5 * my_8` updates the value of `result` to be the product of the two variables again, this time giving `40 * 40` which gives 1600.

1.3.2 Function Calls in R

Up until this point we have simply used numerical operations and variable assignment. While this allows R to serve as a very powerful calculator, we often want computers to do much more than arithmetic. As a result, we need to explore **functions** in R. A function is a piece of code which takes in various arguments and outputs some value (or values). Most of the way that we will use R in this course is through the use of function calls.[16] This is exactly analogous to a mathematical function: it is simply some rule which maps input to output. In fact, some of the most basic functions in R are functions which relate to mathematical functions.

```
x <- 10

exp(x)                                                    ①
sqrt(x)                                                   ②
```

[16] Note: when you begin to write programs for yourself, a lot of your time will be spent writing custom functions. If this is of interest to you, I suggest looking into programming more! For this course we will not need to define our own functions, except perhaps ones that will be defined for you.

```
log(x)                                                                    ③
## [1] 22026.47
## [1] 3.162278
## [1] 2.302585
```

① Computes the exponential function applied to `x`, that is $e^x = e^{10}$.
② Computes the square root of `x`, that is $\sqrt{x} = \sqrt{10}$.
③ Computes the natural logarithm of `x`, that is $\log(x) = \log(10)$.

The basic format of a function call will always be `function_name(param1, param2, ...)`. Each of these functions required only a single parameter, however, there are some functions which take more than one parameter. If we have decimal numbers, for instance, we may wish to round them. To do so, we can use the `round` function in R, which takes two parameters: the number we wish to round, and how many digits we wish to keep.

```
unrounded_value <- 3.141592
rounded_value <- round(unrounded_value, digits = 3)

rounded_value
## [1] 3.142
```

There are a few things to note about this sequence of function calls. First, note that we assign the output of a function to a new variable. This behaves exactly as we saw above with simple numeric calculations. Next, consider that the output of the function[17] has a value of 3.142. That is: we rounded the value to 3 decimal points, exactly as would be expected. The final part to note is that the second parameter passed to the function call is **named**. That is, instead of simply calling `round(unrounded_value, 3)`, we have `round(unrounded_value, digits = 3)`. If you had made the first call this would have worked perfectly.[18] However, R also provides the ability to pass in parameter names alongside the parameter values with the syntax `function_name(param_name = param_value, ...)`. The benefit to doing this is two-fold. First, it is easier to read what is happening, especially for function calls that you have never seen before. Second, it removes the need to have the parameters ordered correctly. It is best practice to **always** include parameter names where you can.

Now, you may be wondering "how do you know what the parameter names should be?" The names are built-in to the different functions that you are working with and so overtime you will become quite familiar with them. However, at any point you can also run the command `?function_name`, replacing `function_name` with the name of a function you are interested in, to open documentation about that function. There you will see not only the names of the different parameters, but useful information regarding what the function does, examples of how to call it, and so forth.

[17] Which we have stored in the variable `rounded_value`
[18] Try it out to convince yourself!

```
?round
```

When we run this code, we are given *a lot* of information. We can see the function name, the details, how it's called, and so forth. In the **Arguments** section we get a list of all of the arguments we can pass to the function, along with a description of them. In this case we see that `round` can take a parameter called `x` for the number to be rounded, and `digits` (which we have previously seen).

```
unrounded_value <- 3.141592
rounded_value <- round(digits = 3, x = unrounded_value)

rounded_value
## [1] 3.142
```

This code produces the exact same output, where now both parameters are named. Specifically, we could read this function call as "round the value of `x` to have `digits` decimal points." If you had instead written `round(3, unrounded_value)`, you would get the value `3`, since now we are rounding `3` to `3.141592` decimal points.

1.3.3 Moving Beyond Numeric Data

So far, everything that we have looked at has been numeric data. We have seen integers, and decimals. You can have negative results, say by taking `my_var <- -5`. And while numbers are frequently useful, we will require further types of data to write useful computer programs. For this course we will focus on three additional data types: textual data which (referred to as **strings**), true and false binary data (referred to as **booleans**), and lists of the same data type (referred to as **vectors**). They will behave in much the same way as numeric data, with different functions and techniques which can be applied to them.

To define a string of text, we simply encapsulate the text that we are interested in within quotation marks (either single `'` or double `"` quotation marks will work).

```
first_string <- "This is a string."
second_string <- 'This is also a string.'

first_string
## [1] "This is a string."
```

Two commonly used functions which rely on strings are `paste` and `print`. Each will take in strings as input, and they do not need to be named. The function `paste` can take in as many strings as you would like. It will "paste" together all the strings provided, creating a longer

string out of these. The function `print` will display the string that is passed as output. Until now we have been running these programs in a way where all calls are displayed as output: this will not always be the case, and so print can come in handy there.

```
my_greeting <- "Hello! Welcome to R programming,"
my_name <- "Dylan"

combined_string <- paste(my_greeting, my_name, "!")
print(combined_string)
## [1] "Hello! Welcome to R programming, Dylan !"
```

Note that `paste` has several additional options which can be investigated in the documentation for `paste`. This is only the simplest use case for it. In general, strings are particularly helpful when we wish to have output from the computer that will be human readable. Where strings are largely for humans, booleans are largely for computers.

Much of what computer programming entails is checking whether certain conditions hold, and then taking different actions depending on what is found. In order to do this, the computer needs a way to represent true and false statements. In R, these are codified with the values `TRUE` and `FALSE`. Note, the capital letters here make a difference. You cannot use `true` or `True` or any other combination thereof. More important than simply being able to specify the values `TRUE` and `FALSE` directly is the ability to detect whether certain statements are `TRUE` or `FALSE`. For this we require comparison operators.

If we think of mathematical comparisons we can state whether two things are equal, or not equal, and whether one thing is less than (or equal to) or greater than (or equal to) another. We can run all of these same checks in R.

- To check whether two quantities are equal you use `quantity_1 == quantity_2`. This statement will be `TRUE` if `quantity_1` and `quantity_2` are exactly the same, and will be `FALSE` otherwise.
- To check whether two quantities are not equal, you can use `quantity_1 != quantity_2`. This statement will be `TRUE` if the quantities differ from one another.
- To check whether one quantity is larger than another, you can use `quantity_1 > quantity_2`. If you want to know whether it is greater than *or* equal to, you can use `quantity_1 >= quantity_2`.
- To check whether one quantity is smaller than another, you can use `quantity_1 < quantity_2`. If you want to know whether it is less than *or* equal to, you can use `quantity_1 <= quantity_2`.

```
5 == 5 # TRUE
5 == 6 # FALSE
```

```r
5 != 5 # FALSE
5 != 6 # TRUE

5 > 6 # FALSE
5 > 5 # FALSE

5 >= 6 # FALSE
5 >= 5 # TRUE

5 < 6 # TRUE
5 < 5 # FALSE

5 <= 6 # TRUE
5 <= 5 # TRUE
## [1] TRUE
## [1] FALSE
## [1] FALSE
## [1] TRUE
## [1] FALSE
## [1] FALSE
## [1] FALSE
## [1] TRUE
## [1] TRUE
## [1] FALSE
## [1] TRUE
## [1] TRUE
```

There are two key points to note beyond this. First, we will of course not normally compare two constants to one another. We already know that `5==5` and so we would not need to check it. We can, however, plug-in variables, perhaps with unknown values, and have the same types of statements being made. Second, the checks for equality and inequality also work with other data types (like strings).

```r
string1 <- "STRING1"
string2 <- "STRING1"
string3 <- "string1"
string4 <- "5"
string5 <- "5.3421"
num1 <- 5
num2 <- 1
num3 <- 0
bool1 <- TRUE
```

```r
bool2 <- FALSE

string1 == string2 # TRUE
string1 == string3 # FALSE
string2 != string3 # TRUE
string1 != string4 # TRUE

string4 == num1    # TRUE
string1 == num2    # FALSE
num1 == string5    # FALSE
num2 == bool1      # TRUE
bool2 == num3      # TRUE
bool2 == num2      # FALSE
## [1] TRUE
## [1] FALSE
## [1] TRUE
## [1] TRUE
## [1] TRUE
## [1] FALSE
## [1] FALSE
## [1] TRUE
## [1] TRUE
## [1] FALSE
```

The final checks may be slightly odd. Here we are comparing across different types of data. When we do this R will automatically try to convert from one type to the other. With strings and numbers this is not too challenging. If they can be converted nicely between types, then they are and the values are compared. Otherwise, R will conclude they are not equal by default. For booleans, it is important to note that `TRUE == 1` and `FALSE == 0`. We will often use these values interchangeably.

The final data type that we will consider are vectors. Vectors store many different values, of the same type, in a single object. Thus, we may have a vector of numeric data, or a vector of strings, or a vector of booleans. The vectors will always contain the same type throughout, but they are stored in a single object (and as such, for instance, can be stored in a single variable). To define a vector we call `c(...)`, where the `...` contains the set of objects we want to store in the vector. The `c` stands for **c**oncatenate, as we are *concatenating* together the set of items into a single container.

```r
v1 <- c("vector", "of", "strings")
v2 <- c(3, 1, 4, 1, 5)
v3 <- c(TRUE, TRUE, FALSE, TRUE)
```

```
v4 <- c(1==2, 2==2, 3==4)
v1
v2
v3
v4
## [1] "vector"  "of"      "strings"
## [1] 3 1 4 1 5
## [1]  TRUE  TRUE FALSE  TRUE
## [1] FALSE  TRUE FALSE
```

We see that each of these vectors holds one type of object. Vectors can be of arbitrary and different lengths. It is also possible to combine multiple vectors *of the same type* into one, by using the `c` function again.

```
v1 <- c(3,1,4,1,5)
v2 <- c(9,2,6)
v3 <- c(5)
num1 <- 3

combined_v1 <- c(v1, v2)
combined_v2 <- c(combined_v1, v3)
combined_v3 <- c(combined_v2, num1)

combined_v3
##  [1] 3 1 4 1 5 9 2 6 5 3
```

Here we combine different numeric vectors together. We also show, when forming `combined_v3`, how numeric vectors can have single items added onto them. That is, if you have a single number, it can be treated as a vector with one element in it. This becomes very useful when building up vectors within code. In addition to combining multiple vectors together, we can also select elements out of a vector. To do this, we use a set of square brackets after the vector's name, with a number within those square brackets specifying the **index** of the vector we are interested in. The index is simply the element position starting at 1[19] and running to the length of the vector. We can include a vector of indices to select multiple elements at once.

```
alpha_vector <- c("A","B","C","D","E","F","G","H","I",
                  "J","K","L","M","N","O","P","Q","R",
```

[19] If you have programmed in the past there is a good chance the language you have learned is "0 indexed" rather than "1 indexed". In R, all vectors start at position 1 and count up, which is not the case in many languages. Be careful of this.

```
                  "S","T","U","V","W","X","Y","Z")

alpha_vector[4]    # D
alpha_vector[25]   # Y
alpha_vector[12]   # L
alpha_vector[1]    # A
alpha_vector[14]   # N

my_name <- alpha_vector[c(4,25,12,1,14)]

my_name
## [1] "D"
## [1] "Y"
## [1] "L"
## [1] "A"
## [1] "N"
## [1] "D" "Y" "L" "A" "N"
```

Note that, each element is selectable individually, giving a single item of that type (in this case, strings). If you select multiple of the elements together, it will create a vector of that type (in this case, a string vector). Note that in addition to selecting elements in this way by their indices, you can also update the elements in the same way.

```
cur_year <- c(2,0,2,3)
cur_year

# After Midnight on December 31
cur_year[4] <- 4
cur_year
## [1] 2 0 2 3
## [1] 2 0 2 4
```

In this example we are changing the last element of the vector. Sometimes we may not know how long the vector actually is, if for instance, it is being built-up as our code runs. If we ever want to check the length of a vector, we can simply call the function `length` which takes as input only one vector, and outputs the numeric value of its length.

```
v1 <- c(1,2,3,4,5)
length(v1)
## [1] 5
```

1.3.4 Program Control Flow

We have seen different types of data, different ways of manipulating data, functions, and variables so far. In order to bring all of these concepts together into useful programs we need some way to control the flow of our programs. We have seen that, by default, programs execute from the top until the bottom. However, it will often be the case that we want want to have certain code running only if certain conditions hold, or that we want to repeat some piece of code many times over. To accomplish these tasks we require **control flow statements**. We will consider only two types of control flow statements now, which will serve well enough to read most of what needs to be read for this course.

The first type of statement is the **conditional statement**. Conditional statements execute only when certain conditions hold. The simplest conditional statement is an **if** statement. The format to define an if statement if `if (condition){ ... }`, where `condition` is some logical condition to be evaluated. If the condition is `TRUE` then the code contained in { … } is evaluated. If the condition is `FALSE` it is not.

```
my_number <- 5

if(my_number > 0) {
    print("My number is a positive.")
}

if(my_number < 0) {
    print("My number is a negative.")
}
## [1] "My number is a positive."
```

In this case, these simple conditional statement check to see whether the number we entered is larger than zero and whether the number we entered is smaller than zero, respectively. When run, notice that only one of the statements is executed.[20] In this case, we know that only one (or neither) of these statements can be true. When that is the case it may make sense to make use of `else` clauses in our conditional logic.

```
my_number <- -5

if(my_number > 0) {
    print("My number is a positive.")
} else {
    print("My number is not a positive.")
}
## [1] "My number is not a positive."
```

[20] In fact, if `my_number` is set to `0` then none of the statements are executed.

Here, *if* the number is greater than `0`, then we run the first block of code, *otherwise* we run the second block of code. Thus, whenever a positive number is entered, we see "My number is a positive", and whenever a non-positive number is entered, we see "My number is not a positive." We can extend `else` blocks to be `else if` blocks, where further conditions can be specified.

```
my_number <- 100

if(my_number > 50) {
    print("My number is a large positive value.")
} else if(my_number > 0){
    print("My number is a small positive value.")
} else if(my_number < -50) {
    print("My number is a large negative value.")
} else if(my_number < 0) {
    print("My number is a small negative value.")
} else {
    print("My number is zero.")
}
## [1] "My number is a large positive value."
```

In these case we can simply pass through each conditional statement in order. First, is the number larger than `50`? If so, print the statement, otherwise we check the next condition, is the number greater than `0`? Note that if we are checking this condition we *know* that the number is less than or equal to `50` since it failed the first check. We continue through the rest of this procedure down until the last else block. This block is run only when all of the other conditions fail: that is, our value is not larger than `50`, or larger than `0`, or smaller than `-50`, or smaller than `0`. The only value that satisfies this is `0` itself.

Sometimes, we wish to check compound conditions. That is, we want to know whether multiple conditions hold, or perhaps, whether at least one of many conditions hold. These statements can be converted into "and" and "or" statements, respectively. To denote "and" statements we use `&&` and to denote "or" statements we use `||`. Thus, the check `my_val > 0 && my_val < 100` returns true only if `my_val` is both above `0` **and** below 100. The check `my_val < -50 || my_val > 50` returns true whenever **either** `my_val` is less than `-50` **or** `my_val` is greater than `50`.

```
my_number <- 5

if(my_number < 50 && my_number > 10) {
    print("A moderate, positive number.")
} else if(my_number != 5 && my_number <= 10 && my_number >= 0) {
    print("A positive value which is not 5.")
```

```
} else if(my_number == 5 || my_number < 0) {
    print("Either 5 or a negative value.")
}
## [1] "Either 5 or a negative value."
```

Conditional statements can grow to be very complex, however, with these rules you can read through them top-to-bottom, substituting for "and" and "or" where necessary. It is also possible, where required, to place one conditional statement inside the code block for another, and to combine them with any of the other techniques that we have learned thus far.

The final piece of control flow that we will consider for now is the `for` loop. The idea with a `for` loop is that we want to repeat the same action either a certain number of times, or for every item in a set of items. To do so, we use the syntax `for(x in vector){...}`, where the code in `...` will be performed once for every single item in the vector. Within the code block specified by `...`, the value `x` will take on the current value in the loop.

```
for(x in c(1,2,3)){
    print(x)
}
## [1] 1
## [1] 2
## [1] 3
```

Notice that three values are printed, in order, `1` then `2` then `3`. The loop code is run three different times, one for each element in the list. Each time the loop is running the next value from the list gets assigned to the variable `x`. The first time it runs it gets the first element, and so forth. As a result, we can use these values in our calculations in whatever way we need to.

```
for(x in c(1,2,3)){
    x_sq <- x^2
    print(paste("The square of", x, 'is', x_sq))
}
## [1] "The square of 1 is 1"
## [1] "The square of 2 is 4"
## [1] "The square of 3 is 9"
```

Whenever we are trying to form a numeric vector with consecutive elements, as we are in `c(1,2,3)`, we can make this easier on ourselves by simply specifying the upper and lower bounds of the range, separated by a colon. That is `c(1,2,3) == 1:3`. This is often useful when specify a loop as we very often want to repeat something a set number of times. Note

that we do not ever *need* to use the value of the looping variable. Sometimes, we just want things to repeat, and so the loop is a convenient way to do that.

```
times_to_loop <- 5

for(x in 1:times_to_loop){
    print(paste("This will get printed", times_to_loop, "times."))
}
## [1] "This will get printed 5 times."
## [1] "This will get printed 5 times."
## [1] "This will get printed 5 times."
## [1] "This will get printed 5 times."
## [1] "This will get printed 5 times."
```

1.3.5 Reading Through a More Complex R Program

Take a moment to read through the following R program and try to understand what is happening exactly. There are comments throughout which will assist in the parsing of the script. We have seen comments up until now, without drawing explicit attention to them. In R, anything placed after a **#** on the line is considered a comment. The programming language ignores these and so they are only there to help other individuals who may be reading through. It is good practice to comment your code to help others, and also to help yourself whenever you return to it in the future. In the course notes code will typically be commented. If you are reading the PDF version of the notes often these comments will be annotations beside the code (numbering certain lines) with the comments provided below, for the sake of legibility.

Note, this combines everything that we have learned, and it is entirely understandable if it takes some time to process. Fortunately, you can always try running the script yourself, and playing with different components of it. Remember, if you do not know what a function call does, you can use the documentation[21] and try playing with it some yourself. To help with the interpretation here, note that this is an implementation of the game that Charles and Sadie have been playing.

```
# Define some variables which dictate how the game runs.
player1 <- "Charles"
player2 <- "Sadie"
num_of_flips <- 3
flip_option1 <- "H"
flip_option2 <- "T"

# Begin with the game setup
```

[21]The internet is also a wonderful resource, one which even very experienced developers make frequent use of.

```r
num_to_win <- round(x = num_of_flips / 2, 0)
flip_results <- c()
flip_options <- c(flip_option1, flip_option2)
total_1 <- 0
total_2 <- 0
player_name <- "" # This is an 'empty string'

# Start Playing the Game
for(game_round in 1:num_of_flips) {
    print(paste("Now starting round", game_round))

    # Select a flip of the coin, using the sample function.
    # See more details with ?sample
    flip_result <- sample(x = flip_options, size = 1)

    # See who benefits from this flip
    # Then set player_name to be the player who benefits from the flip,
    # and update their score variable.
    if (flip_result == flip_option1) {
        player_name <- player1
        total_1 <- total_1 + 1
    } else if (flip_result == flip_option2) {
        player_name <- player2
        total_2 <- total_2 + 1
    }

    print(paste("The flip was a", flip_result, "which benefits", player_name))

    # Check to see if either player has won at this point.
    if(total_1 >= num_to_win) {
        print(paste(player1, "has scored enough points to win."))
    } else if(total_2 >= num_to_win){
        print(paste(player2, "has scored enough points to win."))
    }
}

# The game is over
# Select the winner based on who scored above the threshold
# and then print out the results.
winner <- player1

if(total_2 >= num_to_win) {
```

```
    winner <- player2
}

# Separate these statements into multiple lines to ensure the text is
# not too wide.
print(paste("After flipping the coin", num_of_flips, "times,", player1,
            "scored a total of", total_1, "points"))
print(paste("while", player2, "scored a total of", total_2, "points."))
print(paste("As a result", winner, "won the game and will not have to pay!"))
## [1] "Now starting round 1"
## [1] "The flip was a H which benefits Charles"
## [1] "Now starting round 2"
## [1] "The flip was a T which benefits Sadie"
## [1] "Now starting round 3"
## [1] "The flip was a T which benefits Sadie"
## [1] "Sadie has scored enough points to win."
## [1] "After flipping the coin 3 times, Charles scored a total of 1 points"
## [1] "while Sadie scored a total of 2 points."
## [1] "As a result Sadie won the game and will not have to pay!"
```

1.3.6 R Programming for Probability Interpretations

Recall that the motivation for the discussion of R was the Frequentist interpretation of probability. One task that computers are very effective at is repeatedly performing some action. As a result, we can use computers to mimic the idea of repeatedly performing an experiment. Consider the simple case of flipping a coin over and over again.

We can use `sample(x, size)` as a function to select `size` realizations from the set contained in `x`. Thus, if we take `sample(x = c("H","T"), size=1)` we can view this as flipping a coin one time. If we use the loop structure we talked before, then we can simulate the experience of repeatedly flipping a coin. Consider the following R code. Note, any time that we are doing something which is randomized in R (such as drawing random samples) we also will make use of the `set.seed()` function. This function takes in an integer value as an argument, and by providing the *same* integer value we can make sure to always get the same random numbers generated.[22] This helps to ensure the repeatability of any R analysis, and it is good practice to do. To see what happens without seeding, try modifying the following code without a seed, and running it several times. Then, set the seed (to any number you like) and do the same process.

[22] Technically, we cannot use a computer to generate random numbers. We can only generate *pseudo random* numbers, which are close enough for most purposes.

```r
set.seed(3141592)                                                    ①

number_of_runs <- 1000                                               ②
tosses <- c()                                                        ③

for(idx in 1:number_of_runs){                                        ④
  toss <- sample(x = c("H","T"), size = 1)                           ⑤
  if(toss == "H"){                                                   ⑥
    tosses <- c(tosses, 1)
  } else {
    tosses <- c(tosses, 0)
  }
}

mean(tosses)                                                         ⑦
## [1] 0.522
```

① A seed ensures that the random numbers generated by the program are always the same. This helps to be able to reproduce our work.
② This is how many times we want to repeat the experiment.
③ This is where we are going to store the results of our tosses. It creates an empty list for us.
④ Here we are going to loop over the experiments, one for each run.
⑤ This is our coin toss. We are going to sample 1 from either 'H' or 'T'
⑥ If the coin toss is heads, then we add a 1 to the list. Otherwise, we add a zero to the list.
⑦ Return the mean of all of the tosses.

It is worth adjusting some of the parameters within the simulation, and seeing what happens. What if you ran the experiment only 5 times? Ten thousand times? What if instead of counting the number of heads, we wanted to count the number of tails? What if we wanted to count the number of times that a six-sided die rolled a 4? All of these settings can be investigated with simple modifications to the provided script.

References

2 The Mathematical Foundations of Statistical Experiments

2.1 The Sample Space and Events

While we gave the mathematical formulation for the Frequentist interpretation of probability, we will typically require a more detailed mathematical model to work with probabilities. We want a description, framed in terms of mathematical objects, which will allow us to work out the probabilities of interest. In general, to form such a probability model we need both a list of all possible outcomes that the experiment can produce, as well as the probabilities of these outcomes.

We call the list of outcomes that can occur from an experiment the **sample space** of the experiment. The sample space is denoted $\mathcal{S}$, and is defined as the set of all possible outcomes from the experiment. For instance, if the experiment is flipping a coin we have $\mathcal{S} = \{\mathrm{H}, \mathrm{T}\}$. If the experiment is rolling a six-sided die then $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

Definition 2.1 (Sample Space)**.** The sample space of a statistical experiment is the set of all possible outcomes that can be realized from that experiment. The sample space is typically denoted $\mathcal{S}$, or with similar script letters.

Example 2.1 (Enumerating Sample Spaces)**.** Write down the complete sample space, $\mathcal{S}$ for the game that Sadie and Charles play, based on flipping and observing a coin three times in sequence.

Solution

For Sadie and Charles their experiment involves tossing a coin three times in sequence. As a result each outcome is a three-dimensional list of values, given for instance by $(\mathrm{H}, \mathrm{H}, \mathrm{H})$. As a result, we can write down the full sample space as

$$\mathcal{S} = \{(\mathrm{H}, \mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H}), (\mathrm{H}, \mathrm{T}, \mathrm{T}), (\mathrm{T}, \mathrm{H}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T}, \mathrm{T})\}.$$

With the sample space formally defined, we can revisit Definition 1.2, and formally define the concept of an event.

Definition 2.2 (Event)**.** An event is any collection of outcomes from a sample space for a statistical experiment. Mathematically, an event, E, is a subset of $\mathcal{S}$, and we write $E \subset \mathcal{S}$.

Take for instance the experiment of a single coin. In this case, $E_1 = \{\mathrm{H}\}$, $E_2 = \{\mathrm{T}\}$, and $E_3 = \{\mathrm{H}, \mathrm{T}\}$ are examples of possible events. Here, E_1 corresponds to the event that a head is observed, E_2 corresponds to the event that a tail is observed, and E_3 corresponds to the event that either a tails or a heads was observed. Note that for each event we have $E_1 \subset \mathcal{S}$, $E_2 \subset \mathcal{S}$, and $E_3 \subseteq \mathcal{S}$.

Example 2.2 (Basic Event Listing)**.** List several events from the game that Charles and Sadie are playing. Indicate why these are events.

Solution

Recall that an event is any subset of the sample space. In Example 2.1 we define $\mathcal{S}$ for this game. As a result we can take sets which contain any combinations of these elements. For instance $E_1 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H})\}$, or $E_2 = \{(\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H})\}$, or

$$E_3 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H}), (\mathrm{H}, \mathrm{T}, \mathrm{T}), (\mathrm{T}, \mathrm{H}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T}, \mathrm{T})\}.$$

These are all events since $E_1 \subset \mathcal{S}$, $E_2 \subset \mathcal{S}$, and $E_3 \subset \mathcal{S}$.

Example 2.3 (Event Identification)**.** Is "Charles has to pay" an event from the game that Charles and Sadie are playing? Why?

Solution

No. "Charles has to pay" is not an event as it is not a subset of the sample space. This could plausibly be seen as a real-world description of a possible event, but it is not *itself* an event. [1]

Example 2.4 (Defining Events from Real-World Descriptions)**.** What event corresponds to the description "Charles has to pay" in the game that Charles and Sadie are playing? Recall that they flip a coin three times, and Charles will pay if at least two heads come up, while Sadie will pay if at least two tails come up.

Solution

Charles will have to pay whenever there are two or more heads. As a result we can

[1] As time goes on we will become less strict about this language. When speaking to a statistician, they would understand "Charles has to pay" as an event that *can* occur based on the defined sample space, by simply transforming it into the language of the sample space. However, the distinction is important to make: events are always subsets of the sample space. Once this is second nature, it is a rule that can be loosened, as the knowledge can always be fallen back on when needed. Simply put: you need to know the rules in order to break them!

enumerate the possible outcomes that leads to Charles paying. We have

1. (H, H, H);
2. (T, H, H);
3. (H, T, H);
4. (H, H, T).

Any other outcome will have fewer than two heads, and as a result, Charles will not have to pay. Thus, to form an event, we consider the set with each of these outcomes in it. This gives

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

While the events $E_1 = \{\mathrm{H}\}$ and $E_2 = \{\mathrm{T}\}$ each correspond to a simple outcome from the sample space, $E_3 = \{\mathrm{H}, \mathrm{T}\}$ corresponds to a combined event. We call direct outcomes **simple events** and more complex outcomes like E_3 **compound events**.

Definition 2.3 (Simple Event)**.** A simple event is any event which corresponds to exactly one outcome from the sample space. A simple event only has one way of occurring. The size of the set for a simple event will be 1. The sample space, in turn, is made up of a collection of simple events.

Definition 2.4 (Compound Event)**.** A compound event is any event which corresponds to more than one outcome from the sample space. A compound event can occur in multiple different ways. The size of the set for the compound event will be greater than 1.

If we consider rolling a six-sided die, then an example of a simple event is that a four shows up, denoted $\{4\}$. A compound event could be that an even number is rolled, $\{2, 4, 6\}$, or that a number greater than or equal to four is rolled, $\{4, 5, 6\}$.

Example 2.5 (Identifying Simple and Compound Events)**.** List an example of (at least) one simple and one compound event from the game that Charles and Sadie are playing.

Solution

An example of a simple event would be $E_1 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H})\}$ since it is comprised of exactly one outcome. If three heads are rolled, this event occurs, there is no other way for it to occur. An example of a compound event would be

$$E_2 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{T}, \mathrm{H}), (\mathrm{H}, \mathrm{H}, \mathrm{T})\}.$$

Here there are four outcomes that correspond to this event, and if any of those outcomes are observed the event occurs.

We say that an event "occurs" if any of the outcomes comprising the event occur. As a result we can have more than one event occurring as the result of a run of a statistical experimental. Suppose that we are rolling a fair, six-sided die. Consider the events "an even number was rolled" and "a number greater than or equal to four was rolled." If a four or a six are rolled, both of these events happen simultaneously. Our goal when working with probability will be to assign probability values to different events. We will talk about how likely, or unlikely, events of interest are, given the underlying statistical experiment.

Above, we defined E_3 to be equal to $\mathcal{S}$. As a result, we can say that $\mathcal{S}$ *is* an event since $\mathcal{S} \subseteq \mathcal{S}$. This is the event that any outcome is observed, which is certain to happen. Since it is certain to happen, we know it happens with probability 1. There is another "special" event which is important to consider. We call this the *null event*. Denote $\emptyset$, the null event is an event that corresponds to "nothing in the sample space". We know that every time an experiment is run something in the sample space occurs, and so the null event is assigned probability zero.

Definition 2.5 (Null Event)**.** The null event, denoted $\emptyset$ or $\{\}$, is an event from a statistical experiment which corresponds to nothing within the sample space. The null event has probability zero, and it is impossible to observe. Note that, no matter the sample space, $\emptyset \subset \mathcal{S}$.

2.2 Set Operations for Event Manipulation

Ultimately, we think of all events as being sets. These sets are subsets of the sample space, and can contain single or multiple outcomes. Every quantity that we are interested in can be expressed as some set of outcomes of interest. In building up these sets it is common to construct through the use of "and", "or", and "not" statements. That is, we may say that our event occurs if some outcome **OR** another outcome occurs, or perhaps our outcomes occurs if some outcome does **NOT** occur.[2]

Consider the example of drawing cards from a standard 52-card deck. In such a deck there are 13 card ranks, and four card suits, with one of each combination present. If we draw a single card we can think of the outcomes of the experiment as being any of the 52 possible combinations of rank and suit. We are often interested in an event such as "the card is red", which is the same as saying "the card is a heart **or** the card is a diamond." Perhaps we want to know whether the card was an ace through ten, this is the same as saying "the card is **not** a Jack **or** a Queen **or** a King." If we are interested in the event that the ace of spades was drawn, this can be expressed by saying that "the card was a spade **and** the card was an ace."

As you begin to pay attention to the linguistic representation of events that we use, you will notice more and more the use of these words to form compound events in particular. As a

[2]While both *or* and *not* language is likely clear from the examples we have seen so far, *and* language may be slightly less obvious. While we will explore this in more depth shortly, note that you could not have two simple events occurring simultaneously. If E_1 and E_2 are both simple events, then you can have E_1 **or** E_2, and you can have **not** E_1, but you cannot have E_1 **and** E_2. This is *not* true for compound events.

result, we give each of them a mathematical operation which allow us to quickly and compactly express these quantities in notation.

Example 2.6 (Description of Events)**.** Describe "Charles has to pay", based on the game Charles and Sadie are playing, using language revolving around "or", "and", and "not" each. That is, describe observing at least two heads, on three flips of a coin, one time using "or", one time using "and", and one time using "not".[3]

Solution

There are plausibly many possible ways of doing so. Consider the following:

1. **OR**: Two heads are observed OR three heads are observed.
2. **AND**: Not two tails are observed AND not three tails are observed.
3. **NOT**: Not more than one tails are observed.

We define mathematical operations to encapsulate the use of **or**, **and**, and **not**. These operations apply to any mathematical sets, whether they refer to events or not.

Definition 2.6 (Union)**.** The union encodes the use of "or" in reference to two or more sets. Formally, with two sets A and B, the union of A and B is the set of all elements that are contained in A, or B, or both A and B. We write $A \cup B$ and read that as A union B. When we wish to take the union of many sets, $A_1, A_2, \dots, A_n$, we write this as

$$\bigcup_{i=1}^{n} A_i.$$

Definition 2.7 (Intersection)**.** The intersection captures the use of "and" in reference to two or more sets. Formally, the intersection of two sets, A, and B, is the set that contains all elements that belong to both A and B. We write $A \cap B$, and say "A intersect B." When we wish to take the union of many sets, $A_1, A_2, \dots, A_n$, we write

$$\bigcap_{i=1}^{n} A_i.$$

Definition 2.8 (Complement)**.** The complement makes formal the concept of "not." The complement of a set is the set of all elements which occur in the sample space but are not in the given set. We write this as A^C and say "A complement." When dealing with a sample space, $\mathcal{S}$, the complement of A is the set of all elements in $\mathcal{S}$ that are not in A.

[3]It may be helpful to notice that you can mix and match these terms to your hearts content!

Example 2.7 (Basic Set Operations)**.** For the game being played by Charles and Sadie, take $E_1 = \{(\mathrm{H},\mathrm{H},\mathrm{H})\}$, $E_2 = \{(\mathrm{H},\mathrm{H},\mathrm{H}),(\mathrm{T},\mathrm{H},\mathrm{H}),(\mathrm{H},\mathrm{T},\mathrm{H})\}$, and $E_3 = \{(\mathrm{H},\mathrm{H},\mathrm{T})\}$. Express the following events.

a. $E_1 \cup E_2$;
b. $E_1 \cap E_2$;
c. E_2^C;
d. $E_2 \cap E_3$;
e. $E_1 \cup E_2 \cup E_3$.

Solution

Directly from definitions we can write down each of the following sets:

a.
$$E_1 \cup E_2 = \{(\mathrm{H},\mathrm{H},\mathrm{H}),(\mathrm{T},\mathrm{H},\mathrm{H}),(\mathrm{H},\mathrm{T},\mathrm{H})\} = E_2.$$
As a result, the union of E_1 and E_2 is simply E_2.[4]

b.
$$E_1 \cap E_2 = \{(\mathrm{H},\mathrm{H},\mathrm{H})\} = E_1.$$
As a result, the intersection of E_1 and E_2 is simply E_1.[5]

c.
$$E_2^C = \{(\mathrm{H},\mathrm{H},\mathrm{T}),(\mathrm{H},\mathrm{T},\mathrm{T}),(\mathrm{T},\mathrm{H},\mathrm{T}),(\mathrm{T},\mathrm{T},\mathrm{H}),(\mathrm{T},\mathrm{T},\mathrm{T})\}.$$

d. For $E_2 \cap E_3$ note that they share no elements. As a result, the intersection will be empty since there are no elements common to both of them. This gives $E_2 \cap E_3 = \emptyset$.

e.
$$E_1 \cup E_2 \cup E_3 = \{(\mathrm{H},\mathrm{H},\mathrm{H}),(\mathrm{T},\mathrm{H},\mathrm{H}),(\mathrm{H},\mathrm{T},\mathrm{H}),(\mathrm{H},\mathrm{H},\mathrm{T})\}.$$

Definition 2.9 (Disjoint Events)**.** Two events, E_1 and E_2 are said to be disjoint whenever their intersection is the null event. That is, if $E_1 \cap E_2 = \emptyset$ then E_1 and E_2 are disjoint events.

These concepts allow us to more compactly express sets of interest, and in particular, will be quite useful when it comes to assigning probability. The more times you work with the set operations, the more familiar they will become, and as a result, practice is always useful. Considering rolling a 6-sided die, and take A to be the event that a 6 is rolled, B to be the event that the roll was at least 5, C to be the event that the roll was less than 4, and D to be the event that the roll was odd.

- If we consider D^C this is the event that the roll was even;
- $A \cup C$ is the event that a 6 was rolled or that a number less than 4 was rolled, which is to say anything other than a 4 or a 5;

[5]Note that whenever we have two events, A and B, with $A \subset B$, then $A \cup B = B$.
[5]Note that whenever we have two events, A and B, with $A \subset B$ then $A \cap B = A$.

- If we take $A \cup B$ then this will be the same as B, and $A \cap B$ will be A.
- If we take the event $A \cap C$, notice that no outcomes satisfy both conditions, and so $A \cap C = \emptyset$.
- We can also join together multiple operations. $D^C \cap C$ gives us even numbers than less than 4, which is to say the outcome 2.
- Similarly, $(A \cap B)^C$ would represent the event that a number less than 6 is rolled.

Example 2.8 (Set Operations with Decks of Cards)**.** Charles and Sadie are tiring of flipping their coin, and so they wish to start using decks of cards sometimes instead. Before they formalize a game based on decks of cards, they want to make sure that they are both very comfortable working with these. Suppose that the sample space is defined to be the set of 52 standard cards that may be drawn on a single draw. Describe how set operations can be used to form events corresponding to:

a. A red card is observed.
b. Any card between an ace and a ten is observed.
c. The ace of spades is observed.

Solution

First, we define several events. Note, these can be defined in shorthand to prevent needing to write out many different cards. we take D to be the event that a diamond is observed, we take H to be the event that a heart is observed, take S to be the event that a spade is observed[6] and then take A to be the event that an ace is observed, J to be the event that a Jack is observed, Q to be the event that a Queen is observed, and K to be the event that a King is observed[7] then we can use unions, intersections, and complements to express the previously mentioned scenarios.

a. To represent outcomes corresponding to "the card is red", we can use $D \cup H$.
b. To represent outcomes corresponding to "an ace through ten", we can use $(J \cup Q \cup K)^C$.
c. To represent the outcome "the ace of spades", we may use $A \cap S$.

Working with these basic set operations should eventually become second nature. There are often very many ways of expressing the same event using these different operations, and finding the most useful method of representing a particular event can often be the key to solving challenging probability questions. The first step in making sure that these tools are available to you is in ensuring that the basic operations are fully understood, and this comes via practice. Remember, unions represent "ors", intersections represent "ands", and complements represent "nots".

[7]Note, these three are compound events with 13 different outcomes contained within them.
[7]Compound events with four different options.

2.2.1 Using R To Represent Sample Spaces and Events and Performing Set Operations

We have seen how R can encode sets of elements using vectors. For instance, we may take `sample_space <- 1:6` to represent the sample space of rolling a six-sided die. We can form events by taking subsets of the relevant quantities, selecting via indices. Fortunately, there are also all of the basic set operations implemented in R. We can use `union(x, y)` to perform the union of `x` and `y`, `intersect(x, y)` to perform the intersection of `x` and `y`, and `setdiff(x = sample_space, y)` to perform the complement of `y` (assuming that `sample_space` contains the full sample space).[8]

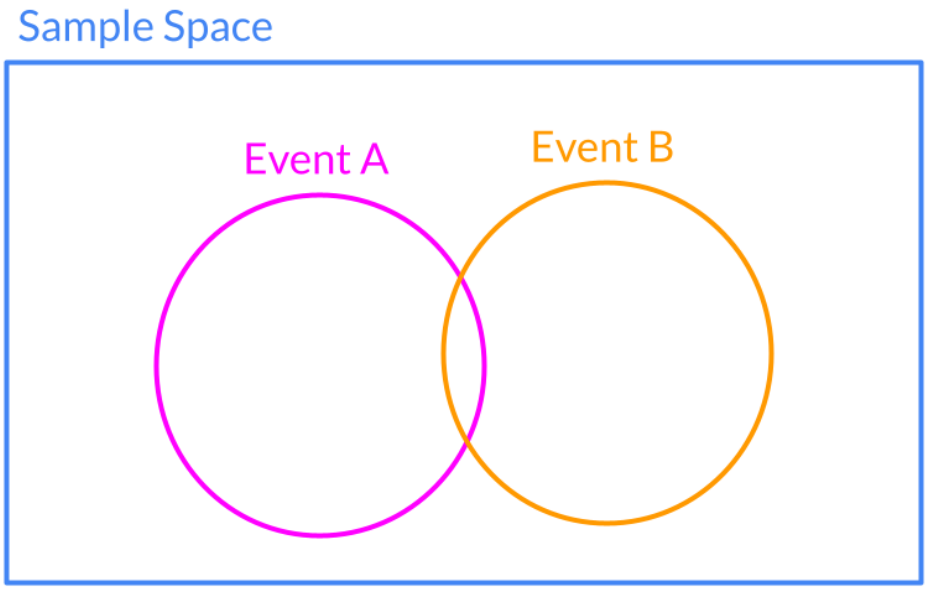
```
# Define the Sample Space of Rolling a 20 Sided Die
sample_space <- 1:20

# Define some Events
E1 <- sample_space[2]
E2 <- sample_space[c(1, 3, 5, 7)]
E3 <- sample_space[c(1, 2, 4, 8)]

# Consider Set Operations
union(x = E1, y = E2) # E1 union E2 = {1, 2, 3, 5, 7}
union(x = E1, y = E3) # E1 union E3 = {1, 2, 4, 8}

intersect(x = E1, y = E3) # E1 intersect E3 = {2}
intersect(x = E1, y = E2) # E1 intersect E2 = {}

setdiff(x = sample_space, y = E1) # E1 complement

# (E2 union E3) complement
setdiff(x = sample_space, y = union(E2, E3))
## [1] 2 1 3 5 7
## [1] 2 1 4 8
## [1] 2
## integer(0)
##  [1]  1  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  [1]  6  9 10 11 12 13 14 15 16 17 18 19 20
```

[8] R does not implement complements directly, and instead implements the set difference operation. The set difference function, `setdiff(x, y)` returns the set of all elements in `x` which are not in `y`, a sort of subtracting of sets. The complement of a set is defined to be $A^C = \text{setdiff}(\mathcal{S}, A)$, indicating why this works!

2.3 Venn Diagrams

The sample space is partitioned into outcomes, and the outcomes can be grouped together into events. These events are sets and can be manipulated via basic set operations. Sometimes it is convenient to represent this process graphically through the use of Venn diagrams. In a Venn diagram, the sample space is represented by a rectangle with the possible outcomes placed inside, and events are drawn inside of this as circles containing the relevant outcomes.

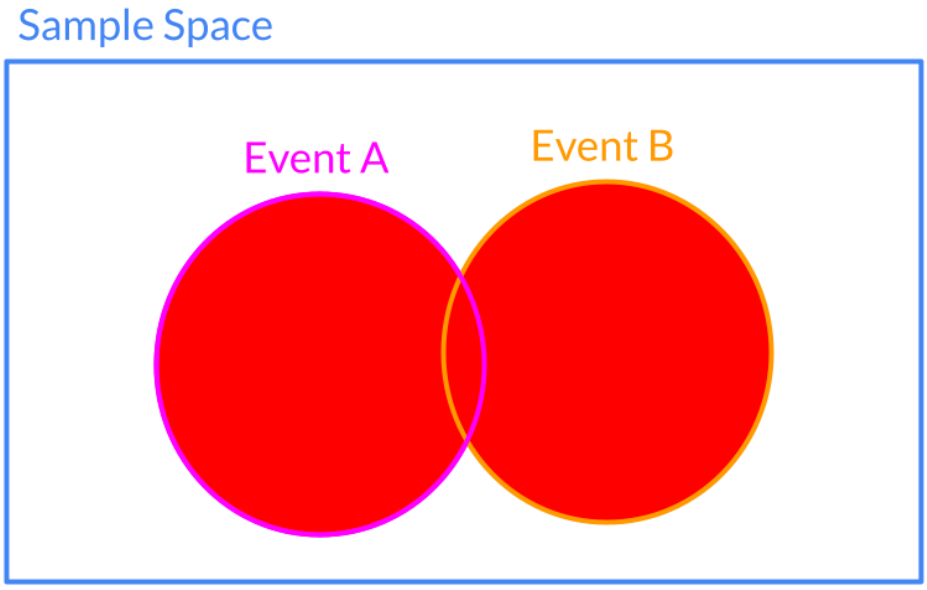
Figure 2.1: A basic Venn diagram, representing the sample space and two different events. In practice, the sample space would have the possible outcomes written into the rectangle, and the circled events would end up containing the relevant outcomes for those events.



On the Venn diagram then, the overlap between circles represents their intersection, the combined area of two (or more) circles represents their union, and everything outside of a given circle represents the complement of a set. This can be a fairly useful method for representing sample spaces, and for visualizing the basic set operations that we use to manipulate events inside the sample spaces. A word of caution: Venn diagrams are useful tools, but they are not suitable as proofs themselves. It is possible to convince yourself of false truths if the wrong diagrams are used, and as a result, Venn diagrams should be thought of as aids to understanding, rather than asa rigorous tool in and of themselves.[9]

[9]This is a general principle in mathematics. Coming up with one example that makes something *seem* true

Figure 2.2: **Union:** The union of events A and B is shaded here in red. The union of two sets is all of the contents of both sets, including the overlap between the two.



does not form an argument demonstrating that it *is* true. Venn Diagrams should largely be thought of as specific examples of the underlying phenomena, which are great if you're a visual learner!

Figure 2.3: **Intersection:** The intersection of events A and B is shaded here in red. The intersection of two sets is all of the content shared by both sets, given by the overlapping area of the two circles.

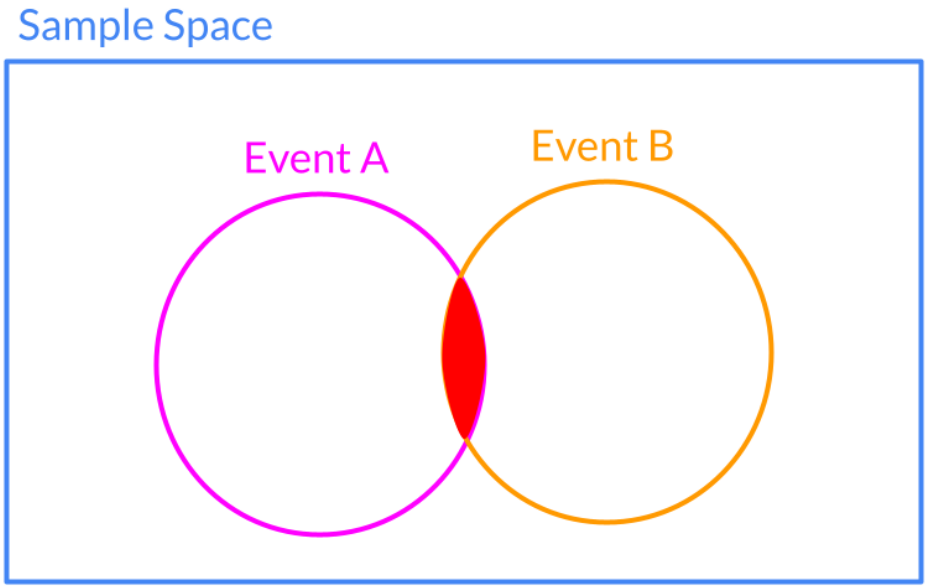
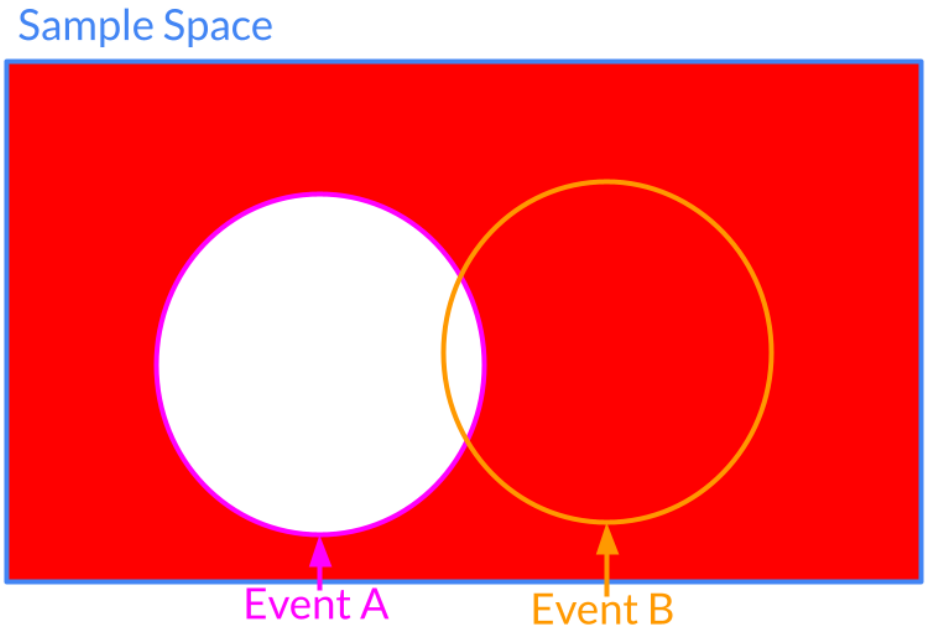


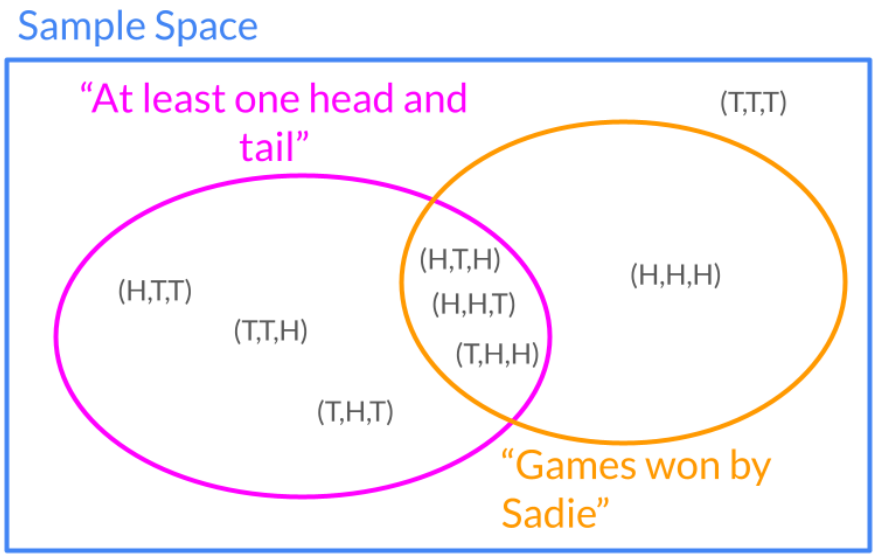
Figure 2.4: **Complement:** The complement of event A is shaded here in red. The complement of a sets is all of area inside of the sample space, not inside of the set. Here we show the complement of Event A, though Event B would be similar.



Example 2.9 (Venn Diagram with Defined Events)**.** Draw a Venn diagram representing the original game that Charles and Sadie played. On the diagram draw the events corresponding to "At least one head **and** one tail are observed", and "Sadie won the game". Recall that three coins are tossed, and Sadie wins if at least two of them show heads.

Solution

The sample space contains the eight possible options. Only $(\mathrm{T}, \mathrm{T}, \mathrm{T})$ does not belong to at least one of the events. Both events share $(\mathrm{H}, \mathrm{H}, \mathrm{T})$, $(\mathrm{H}, \mathrm{T}, \mathrm{H})$, and $(\mathrm{T}, \mathrm{H}, \mathrm{H})$.



Sample spaces, events, and the manipulation of these quantities forms a critical component of understanding probability models. In particular, they describe the complete set of occurrences in a statistical experiment that we could be interested in assigning probability values to. To formalize a probability model, however, we also need some rule for assigning probability values.

Exercises

Exercise 2.1. For each of the following experiments, describe the relevant sample space and identify one possible event of interest.

a. The quality control inspection of smartphone screens from a manufacturing process.
b. Monitoring the ongoing structural integrity of a newly built bridge.
c. A clinical trial studying the effectiveness of a new drug.

d. Epidemiological monitoring of a disease outbreak.
e. Dealing a hand of black jack.
f. Observing the launch conditions for a rocket launch.
g. Debugging in software development.
h. Playing the lottery.

Exercise 2.2. A card is drawn at random from an ordinary deck of 52 playing cards. Let A be the event that a king is drawn, and B the event that a club is drawn. In words, describe the following events.

a. $A \cup B$;
b. $A \cap B$;
c. $A \cup B^C$;
d. $A^C \cup B^C$;
e. $(A \cap B) \cup (A \cap B^C)$.

Exercise 2.3. Suppose that $\mathcal{S} = \{\phi, \lambda, \Delta, \mu\}$. List all possible events from the corresponding experiment.

Exercise 2.4. Suppose an experiment is run which generates realizations from positive integers. Take A to be the event $\{1, 5, 31, 56, 101\}$, $B = \{22, 56, 5, 103, 87\}$, $C = \{41, 13, 7, 101, 48\}$, and D to be the event that the number is odd. Identify (write down or describe) each of the following events.

a. D^C
b. $A \cap B$
c. $C \cup A$
d. $C \cap D$
e. $(A \cup B) \cup (C \cup D)$
f. $A \cap D^C$

Exercise 2.5. Suppose that a 20 sided die is rolled. The events of interest are: A the outcome is a multiple of 4, and B the outcome is a multiple of 5.

a. Draw a Venn Diagram representing the sample space and events.
b. Identify the event $A \cup B$. What does this correspond to in words?
c. Identify the event $A \cap B^C$. What does this correspond to in words?
d. How would you denote the event "neither a multiple of 4 nor 5.'' using this notation?

Exercise 2.6. Suppose that two indistinguishable coins are flipped. The events of interest are: A exactly two heads are seen, and B at least one head is seen.

a. Draw a Venn Diagram representing the sample space and events.

b. Describe every possible outcome with respect to the identified events.
c. Give an event, in terms of the number of heads observed, which is equivalent to the sample space.

Exercise 2.7. A cinema has 12 screens, numbered 1 through 12. Before opening, an employee checks to ensure that the projectors are correctly calibrated.

Let A be the event that all the screens are correctly calibrated, B be the event that the third screen is not correctly calibrated, C be the event that exactly one screen is not correctly calibrated, and D be the event that 5 and 8 are correctly calibrated.

Which of the following pairs of events are disjoint?

a. A and B.
b. B and D.
c. C and D.
d. B and C.

3 The Core Concepts of Probability

3.1 Assigning Probabilities (and The Equally Likely Outcome Model)

There are a plethora of ways to assign probabilities to different events. At the most basic level any rule that maps from the space of possible events to real numbers between 0 and 1 can be used as rules for probability assignment. That is, probability assignment is simply a set of rules which says "for this event assign this probability."

Example 3.1 (Coin Toss Probabilities)**.** Suppose that the fair coin used by Charles and Sadie is tossed one time. Write down the probability assignments relating to this experiment.

Solution

In this case we have $\mathcal{S} = \{\text{H}, \text{T}\}$. Thus, the possible events for which we need to assign probabilities are $\emptyset$, $\{\text{H}\}$, $\{\text{T}\}$, and $\{\text{H}, \text{T}\} = \mathcal{S}$. For any probability model we have $P(\emptyset) = 0$ and $P(\mathcal{S}) = 1$. When we say that a coin is "fair" we are saying that $P(\text{T}) = P(\text{H})$, and since these are the only two possible outcomes in the sample space, we must have that they each have probability 0.5.

Not every assignment of probability values is going to be valid. Suppose, for instance, that we have a six-sided die, each side labelled with a number from one to six. If I told you that there was a probability of 0.5 that it comes up 1, 0.5 that it comes up 2, 0.5 that it comes up 3, 0.5 that it comes up 4, 0.5 that it comes up 5, and 0.5 that it comes up 6, you would probably call me a liar.[1] If, as we have previously seen, probabilities represent the long run proportion of time that a particular event is observed, we cannot have 6 different outcomes each occurring in half of all cases.

Beyond the requirements that we impose on what constitutes a "valid" probability rule, we have another concern: scalability. It is perfectly acceptable to indicate that in an experiment with 3 outcomes, the first has a probability of 0.25, the second of 0.3, and the third of 0.45. What if the experiment has 100 possible outcomes? Or 1000? It quickly becomes apparent that enumerating the probabilities of each event in the sample space is an efficient way of assigning probabilities in practice. A core focus of our study of probability will be finding techniques

[1] Or else conclude that I was mistaken and maybe should not be teaching probability.

that allow us to efficiently encode probability information into manageable objects. Once we have done this we will be in a position where we can manipulate these (comparatively) simple mathematical quantities in order to make statements and conclusions about any of the events of interest, even if they have never been explicitly outlined as having an assigned probability.

While we will consider myriad methods for accomplishing these goals throughout our study of probability, we begin with a very useful model which simplifies probability assignment, without any added complexity, and creates a solid foundation for us to explore the properties of probability models. We start by considering **equally likely outcomes.** As the name suggests, the probability model considering equally likely outcomes assigns an equal probability to every possible outcome of the experiment. This is a probability model that we are already distinctly familiar with: flipping a coin, rolling a die, or drawing a card are all examples of experiments which rely on the equally likely outcomes framework.

Remark (Statisticians and Urn Models). In statistics and probability courses and books you will often have instructors or authors using fairly simple models to illustrate probability concepts. There will often be questions relating to coin tosses, and dice, and decks of cards, and everyone's favourite: urns. It will very frequently be the case that a statistics question will state that there is an urn with some combination of coloured balls within it, from which you will be selecting some number either with or without replacement. The frequency of these types of examples and questions often feels disconnected from the refrain that "uncertainty is all around us" and that "statistics is relevant to every aspect of our world!"[2] Why is it that we seldom see questions or examples that are directly tied to these wide spread applications of the lessons and techniques being taught?

In part these simple experiments are cleaner to handle than "real world" situations. We can easily assume that a die is fair and that takes care of any unsuspecting wrinkles that will necessarily come along with the "real world". This is not dissimilar to working under the assumption of frictionless surfaces in introductory physics, or assuming that human beings are rational in economics. Another key point is that most of us have deep familiarity with dice, and coins, and cards.[3] The same is not going to be true of stories that are derived from different use cases in the real world. A final important point, and this will be something we see in depth in the coming chapters, is that from a statistical point of view: there is no difference. Once we have the tools to work with these quantities, we have the tools to work with any of the quantities. This actually distinguishes the use of these types of examples in statistics and probability from those for other subjects: at no point is anything that we are learning incorrect, or overly simple - we are just focusing on the raw probabilistic nature of the phenomenon. As a result, we will continue to see these simple models in these notes. I would encourage you,

[2] One of the most famous quotes from a statistician was a thought shared by John Tukey, stating "The best thing about being a statistician is that you get to play in everyone's backyard." This is a common refrain, and one rooted in truth. Statistics is everywhere, across every field of human inquiry, and can help us make sense of everything from the trivial to the deeply important.

[3] This does not help to explain why we use urns so much, of course. When was the last time any of us drew a ball from an urn?

whenever possible, to hold a topic in mind that matters more to you and start trying to draw the parallels between rolling dice, and whatever it is that you may care about.

Why urns, specifically? Well, whether it be coin flipping or dice rolling or card selection, we can model this equivalently using an urn (with 2, 6, and 52 items, respectively). The urn becomes more flexible to *exactly* dictate what the probability of any selection will be, which is a useful way of moving from equally likely models (each ball is equally likely to be selected) to arbitrary models (we can have however many identical balls in the urn as we would like).

If we have an experiment with a sample space $\mathcal{S}$ which has $|\mathcal{S}| = k$ total elements[4], then each element of the sample space occurs with probability $\frac{1}{k}$. In the case of the coin toss example, $\mathcal{S} = \{\text{H}, \text{T}\}$, and so $k = 2$ and each outcome occurs with probability $\frac{1}{2}$. In the case of drawing a card at random, there are 52 different outcomes, and so $k = 52$, and the probability of drawing any particular card is $\frac{1}{52}$.

It is critically important to recognize that the equal probability model assigns equal likelihood to the possible outcomes of an experiment, not the possible events of interest. It will not be the case that all events have the same probability. To make this concrete, consider the events A "the ace of spades is drawn" and B "any spade is drawn". It is clear that B happens more frequently than A, even though we have said that this is an experiment with equally likely outcomes. Remember: an outcome is an observation from a single experimental run, an event is any collection of these possible outcomes.

A core goal is then bridging the gap between the probability of an outcome[5] and the probability of an event. In order to do so, we will next consider the rules of probability, introducing properties that are required for valid probability assignments, and the techniques for manipulating probabilities to calculate the probabilities of quantities of interest.

3.1.1 Using R for the Equally Likely Probability Model

In the previous chapter we saw how we can codify sample spaces and events using vectors in R. In the introduction we actually saw how can sample from a sample space using the equally likely outcome framework. Specifically, an application of the `sample` function will draw a set number of values from a sample space, giving each value an equal probability to be drawn.[6]

```
# Define the Sample Space of Rolling a 20 Sided Die
sample_space <- 1:20

# Recall that whenever we wish to perform an experiment in R with
```

[4]Note that, when we have a set, using the absolute value symbols $|\cdot|$ stands for the **cardinality** of the set. Cardinality is just a fancy way of saying the size or the number of elements that the set has in it.

[5]A quantity which in the equally likely outcome framework, we know exactly.

[6]The `sample` function can also be used without equally likely events by specifying a vector of probabilities, however, this is a less common use case.

```r
# randomness, we should call set.seed
set.seed(31415)

# The sample function takes three main parameters:
#   x: the sample space
#   size: the number of items to draw
#   replace: a boolean representing whether the draws should be
#            with replacement or not.
one_roll <- sample(x = sample_space, size = 1)
ten_rolls_with_replacement <- sample(x = sample_space,
                                     size = 10,
                                     replace = TRUE)
ten_rolls_without_replacement <- sample(x = sample_space,
                                        size = 10,
                                        replace = FALSE)

one_roll
ten_rolls_with_replacement
ten_rolls_without_replacement
## [1] 2
##  [1] 19 17 14  3  5 12  2 15  9  3
##  [1] 18  8 16  9  7 20  2 19 10 13
```

3.2 The Axioms of Probability

We have previously seen that not every probability assignment can be valid. For instance, assigning 0.5 probability to each outcome on a die leads to a nonsensical scenario. With just a little imagination, we can conjure equally nonsensical scenarios in other ways. For instance, it would make very little sense to discuss the probability of an event being a negative value. What would it mean for an event to occur in a negative proportion of experimental runs? Alternatively, we can consider two events that are nested in one another: say event A is that we draw the ace of spades, and event B is that we draw any spade. Every single time that A happens, we know that B also happens. But there are ways that B can occur where A does not.[7] If I told you the probability of A was 0.5 and the probability of B was 0.2, this would violate our base instincts. How can it be more likely to draw the ace of spades than it would be to draw any spade at all?[8]

[7] For instance, the Queen of spades being drawn.

[8] This is actually a scenario where our instincts may lead us awry in some situations. Consider the following from Kahneman and Tversky (1972): Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable? (a) Linda is a bank teller, or (b)

Often in mathematics when we have an intuitive set of rules[9] that particular quantities must obey, we work to add formality through defining properties of these concepts. To this end, we can define the key properties that probabilities must obey in order to be well-defined, valid probabilities. With three fairly basic properties, we can completely specify what must be true in order for a set of probabilities to be "valid", and to in turn match with our intuitions.

The Axioms of Probability

1. **Unitary:** Every valid set of probabilities must assign a probability of 1 to the full sample space. That is, $P(\mathcal{S}) = 1$. This is an intuitive requirement as every time the experiment is run we observe an outcome in the sample space. As a result, in every experimental run the event $\mathcal{S}$ occurs.
2. **Non-negative:** We require that every probability is non-negative. We can have probabilities of 0, but we can never have a probability less than zero. Again, this is sensible[10] but is important to include in our formalization. Specifically, for every event E, we must have $P(E) \geq 0$.
3. **Additivity:** the final property requires slightly more parsing on first pass. Suppose that we define a sequence of events, $E_1, E_2, E_3, ...$ such that no two events have any overlap. That is, $E_j \cap E_\ell = \emptyset$ for all $\ell \neq j$. Then, the final property we require for probabilities is that
$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i).$$
That is, the probability of the union of disjoint events is the summation of the probability of these events.

It is worth dwelling slightly on axiom 3. Consider the case of drawing a card at random from a deck of 52 cards. Using the equally likely outcome model for probability we know that the probability that any card is drawn is given by $\frac{1}{52}$. If I were to ask "what is the probability you draw that ace of spades?" under this model you can respond, immediately, with $\frac{1}{52}$. Now, if I were to ask "what is the probability that you draw the ace of spades or the two of spades?" then intuitively you likely figure that this will be $\frac{2}{52}$. Note that the event E_1, "draw the ace of spades" and the event E_2 "draw the two of spades", are disjoint events. Moreover, recall that the union is the "or" and so $E_1 \cup E_2$ is the same as E_1 or E_2. Taken together then,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

The axiom of additivity simply extends this intuition to an arbitrary number of events.

Linda is a bank teller and is active in the feminist movement. A majority of respondents rate (b) as being more probable, even though (a) is contained in (b).

[9] These rules, which we call "properties" are formally known as "axioms".

[10] What would it mean to have a negative probability? It is perhaps a more interesting question than it seems at first glance. It is a topic that has come up in some pretty strange places and, while it is not presently sensible to call them "probabilities" in a traditional sense, there are interesting results which follow.

Example 3.2 (Basic Additivity)**.** Still unsure of how best to go about using cards to replace their coin game, Charles and Sadie are considering various different events and trying to understand their probabilistic behaviour. They take S, C, H, and D to be the events that a spade, club, heart, or diamond are drawn from a standard deck of cards, respectively. Further, they take C_j to be the event that a card with denomination j is drawn (j ranging from ace with 1 through King with 13). If they consider the union of any two (or more) of these events when can they leverage properties of additivity? When can't they?

Solution

In order to use the properties of additivity it is required that the two events are disjoint. Note that taking any two (or more) of S, C, H, and D will lead to disjoint events. There is no way to draw a card which has two suits on it at once. Similarly, taking any two (or more) of C_j will lead to disjoint events. However, mixing any of the suited events (S, C, H, and D) with any C_j will not be disjoint.
Consider $S \cap C_1$. The ace of spades is in S since it is a spade and it is in C_1 since it is an ace. As a result, $S \cap C_1 = \{\text{Ace of Spades}\}$. Because of this we are not able to say that $P(A \cup C_1) = P(A) + P(C_1)$. However, we can say that

$$P(S \cup C \cup H \cup D) = P(S) + P(C) + P(H) + P(D),$$

and could do the same with any subset of these sets. Similarly, we can take

$$P\left(\bigcup_{j=1}^{13} C_j\right) = \sum_{j=1}^{13} P(C_j),$$

or any of the subsets there.

These three axioms fully define valid probabilities. Any mechanism that assigns probability values to events which conforms to these rules will assign valid probabilities. While it may seem counterintuitive that such basic rules fully define our notion of a probability, these rules readily give rise to many other properties that are indispensible when working with probabilities.

3.3 Secondary Properties of Probabilities

Using the previously indicated axioms of probability we are able to derive many useful **secondary properties**. These properties will frequently be used to actually compute different probabilities, and are helpful to become familiar with. All of the following properties follow directly from the axioms, though, some are more clear than others. For the following we take E and $E_1, E_2, E_3, ...$ to be arbitrary events on some well defined sample space.

1. $P(E^C) = 1 - P(E)$, and equivalently, $P(E) = 1 - P(E^C)$.

2. $P(\emptyset) = 0$.
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
4. $$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3).$$
5. If $E_1 \subset E_2$ then $P(E_1) \leq P(E_2)$.

Proofs of the Secondary Properties of Probability

It may be instructive to see how these properties are derived. Doing so generates added familiarity with manipulating probability expressions and helps to encourage deeper understanding.

1. Note that, for any event E, by definition we have $E \cup E^C = \mathcal{S}$ and $E \cap E^C = \emptyset$. As a result, we can apply **additivity** to the sets E and E^C giving $P(E \cup E^C) = P(E) + P(E^C)$. However, since $E \cup E^C = \mathcal{S}$, then we know that $P(E \cup E^C) = P(\mathcal{S}) = 1$ by the **unitary** property. Taken together this tells us that $1 = P(E) + P(E^C)$, and rearranging gives $P(E^C) = 1 - P(E)$, or $P(E) = 1 - P(E^C)$, as required.

2. We know that $\mathcal{S}^C = \emptyset$. Using secondary property (1), $P(E) = 1 - P(E^C)$. Taking $E = \emptyset$ gives $P(\emptyset) = 1 - P(\mathcal{S}) = 1 - 1 = 0$, by the **unitary** property.

3. Here note that $E_1 \cup E_2$ can be written as $E_1 \cup E_2'$ where $E_2' = E_2 \cap E_1^C$. That is, E_2' contains the outcomes from E_2 which were not shared by E_1. Then $E_1 \cap E_2' = \emptyset$ so we can write $P(E_1 \cup E_2) = P(E_1 \cup E_2') = P(E_1) + P(E_2')$, by **additivity**. Now, if we define $E_2^* = E_2 \cap E_1$ then $E_2 = E_2' \cup E_2^*$, and $E_2' \cap E_2^* = \emptyset$. Thus, $P(E_2) = P(E_2' \cup E_2^*) = P(E_2') + P(E_2^*)$. Rearranging this gives $P(E_2') = P(E_2) - P(E_2^*)$, and we know that $P(E_2^*) = P(E_1 \cap E_2)$. Thus, plugging into what we found before we get
$$P(E_1 \cup E_2) = P(E_1) + P(E_2') = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

4. This follows exactly from the argument for (3). To see, first consider $E_2 \cup E_3$ to be an event itself, say E_4. Then we can apply the above result to $E_1 \cup E_4$. And then we need only repeat the process for the remaining terms.

5. We can rewrite E_2 as $E_1 \cup (E_2 \cap E_1^C)$. These two sets are disjoint, since the one is E_1 and the other must contain only elements in E_1^C. Then, $P(E_2) = P(E_1) + P(E_2 \cap E_1^C)$ by **additivity**. Then, through the **non-negative** property we know that $P(E_2 \cap E_1^C) \geq 0$, and so rearranging we have $P(E_1) = P(E_2) - P(E_2 \cap E_1^C) \leq P(E_2)$.

These properties are immensely useful when computing probabilities. In fact, these secondary properties will be used with more frequency than the basic axioms when manipulating proba-

bilities in practice. It is worth building comfort with these properties, early and often, as they will assist in manipulating all probability expressions in the future.

While these properties hold in general for all probability models, it is instructive to focus on the equal probability model to begin building familiarity with probability. These properties allow us to take events – whether compound or simple – and combine, rewrite, and manipulate expressions to assist in the handling of the computations. Eventually, however, we require the ability to assign numerical values to these probabilities.

Consider a simple event, A. Recall that a simple event is defined as a possible outcome of an experiment, and so in this case, A corresponds directly to an event that may be observed. If our sample space is k elements large, then $P(A) = \frac{1}{k}$ in this framework. For instance, if A is the event that a two is rolled on a six-sided fair die, then $P(A) = \frac{1}{6}$.

Now, suppose that a compound event is defined, B. By definition, a compound event can be expressed as a set of possible outcomes from the experiment. Suppose that we enumerate these possible events as $b_1, b_2, \ldots, b_\ell$. Then we know that B occurs if any of $b_1, b_2, \ldots, b_\ell$ occur. Each b_j are elements of the sample space and correspond to possible outcomes of the experiment. As a result, we know that $P(b_j) = \frac{1}{k}$, based on the equal probability assumption. Now, if we take any two distinct events, say b_i and b_j, we know that they must be disjoint: $b_i \cap b_j = \emptyset$. This is because in an experiment run only one outcome can occur. Moreover, we can say that $B = b_1 \cup b_2 \cup \cdots \cup b_\ell$.

Using the axioms of probability outlined above we therefore know that $P(B) = \sum_{j=1}^{\ell} P(b_j) = \sum_{j=1}^{\ell} \frac{1}{k} = \frac{\ell}{k}$. This holds in general for any compound event in this setting. If we take B to be the event that an even number is rolled on a six-sided die, then we would have b_1 is the event that a two is rolled, b_2 is the event that a four is rolled, and b_3 is the event that a six is rolled. There are three such events, and so the probability that an even number is rolled must be $\frac{3}{6} = 0.5$, which matches our intuition.

If we consider what this process is doing at its core, we can reframe the calculation as counting up the number of ways that event can happen and dividing by the total number of events. In our previous discussion, there were ℓ ways of B occurring, a total of k outcomes, and so the probability becomes $\frac{\ell}{k}$. In the equal probability model, this will always be the case. The probability of any event A occurring is given by

$$P(A) = \frac{N_A}{k},$$

where N_A is the number of unique ways that A can occur. In other words, N_A is the size of the set A, $|A|$.

As a result of this, computing probabilities largely relies on the counting of possible outcomes corresponding to different events. If we can determine $N_A = |A|$, and the count of the total number of occurrences, k, then we can determine the probability of A. This study of counting is known as **combinatorics**, and it is where we will turn our attention next.

Example 3.3 (Unmatched Six-Sided Dice)**.** Charles and Sadie, not all together content with the progress through decks of cards, are considering games with dice. Suppose that they have two, fair, six-sided dice. They are interested in the probability that the two dice show different numbers when they are rolled. What is this probability?

Solution

Here, the key is to realize that the probability is easier to solve when considering the complement rather than the event itself. Notably, rolling two, fair, six-sided dice gives a total of 36 possible outcomes. Of these, exactly 6 have equal numbers showing on both dice. Thus, the probability that the two dice show the **same** number is going to be $\frac{6}{36} = \frac{1}{6}$. Then, using the fact that $P(E) = 1 - P(E^C)$, and that taking E to refer to the event where the two dice show different numbers, then E^C refers to the event that the two dice show the *same* number. As a result, the probability that we want is $P(E) = 1 - \frac{1}{6} = \frac{5}{6}$.

Example 3.4 (Unmatched Arbitrary Dice)**.** Charles and Sadie, working from their intrigue about dice, have decided that instead of using two six-sided dice, they wish to take two dice of possibly different sizes. Suppose that the first die has d_1 sides and the second has d_2 sides, and that both dice are otherwise fair. They are interested in the probability that the two dice show different numbers when they are rolled. What is this probability?

Solution

This problem is conceptually no different from Example 3.3. There will be a total of $d_1 \times d_2$ possible combinations of the two dice to be rolled.[11] Of these, the dice will match in $\min\{d_1, d_2\}$ events.[12] Then, with the same complement trick discussed above, we get

$$P(E) = 1 - P(E^C) = 1 - \frac{\min\{d_1, d_2\}}{d_1 \times d_2}.$$

3.4 Combinatorics

3.4.1 The Product Rule

Fundamentally, counting is a matter of assessing the size of a collection of items. Sometimes, this is very straightforward. If you want to count the number of students in a classroom, you start at 1 and enumerate upwards through the integers. To count the number of days

[12]Note: if this is not yet clear to you, that's okay! In the very next section we begin to discuss how to count the possible combinations in these types of scenarios.

[12]Suppose that $d_1 = 2$ and $d_2 = 4$. Then here, the dice can match when they show either 1 or 2, but if the second die shows 3 or 4 there is no possibility of having a match at all.

until the next Holiday, you do the same thing. If you really need to sleep, perhaps you will count imaginary sleep until you drift off. There is not much to this type of counting, and it is certainly deeply familiar to you all. However, it is also quite limited in its utility.

Imagine that you are interested in determining how many possible ways there are of arranging a deck of 52 cards. You could of course arrange them in a particular order, then count each of those. That would take a tremendous amount of time, so perhaps instead of using an actual deck you just write down the combinations. Still, each combination is going to be 52 cards long, and keeping track of that all will be a tremendous challenge. This seems like an approachable question, and yet, it illustrates how complicated (and large) these types of "counting" problems can become very quickly.

Fortunately for us there are some strategies for simplifying these problems down, some of which you are likely already familiar with. Think about trying to form an outfit where you have 4 different sweaters, 3 pairs of pants, and 2 options for your shoes. Suppose that any combination of these will work well. How many total outfits are there? Well, if you have already picked your sweater and pants, then there are going to be 2 different outfits using these: one with each of the pairs of shoes. This is true for each possible sweater-pant combination, and so we can count 2 for each one these. In other words, to get the total number of outfits we multiply the number of sweater-pant combinations by the number of shoe options. The same rationale can be applied to count the total number of sweater-pant combinations. For each sweater, there are 3 pairs of possible pants, and so to get the total number we can take 3 for each possible sweater, or in other words, 3×4. Taken together then we have $4 \times 3 \times 2 = 24$ total possible outfits.

Another way of framing this is that we have to make three sequential decisions: which of the 4 sweaters, which of the 3 pants, and which of the 2 shoes are to be worn? When we do this we multiply through the number of alternatives at each decision point to get the total number of combinations. This is known as the **product rule for counting**.

Definition 3.1 (Product Rule for Counting)**.** The product rule for counting states that, when there are a sequence of k decisions to be made, and for each decision $j = 1, \ldots, k$, there are n_j options, then the total number of combinations will be

$$N = n_1 \times n_2 \times \cdots \times n_k.$$

Example 3.5 (Counting Coffee Orders)**.** When Charles and Sadie are out for coffee, Sadie enjoys ordering the same thing each time: a black coffee and vegan chocolate chip cookie. Charles, on the other hand, has decided to work through the entire menu of the local coffee shop, each day ordering a drink, with one add-in, and a snack. If there are 10 different drinks, 8 possible add-ins, and 12 different snacks, how many trips to the coffee shop will it take until Charles has tried it all?

Solution

This necessitates an application of the product rule for counting. Specifically, we can view this as three sequential decisions, where the first decision is which drink (with $n_1 = 10$), the second decision is which add-in (with $n_2 = 8$), and the third decision is which snack (with $n_3 = 12$). Taking the product gives the total number of combinations as $10 \times 8 \times 12 = 960$. As a result, it will take 960 visits (assuming that nothing on the menu changes!) to try all combinations.

Example 3.6 (Sequence of Dice Rolls)**.** Charles and Sadie have been enjoying playing with dice, but they lost one of the two they had. As a result, they are trying to come up with games revolving around rolling a single die. They decide to try a game called "six is lava", where they roll a single six-sided die 10 times in a row. If they get 1 or more sixes, they lose the game. They are not sure if 10 is the correct number of rolls to use. What is the probability that they lose on any given set of 10 rolls of the die in this game?

Solution

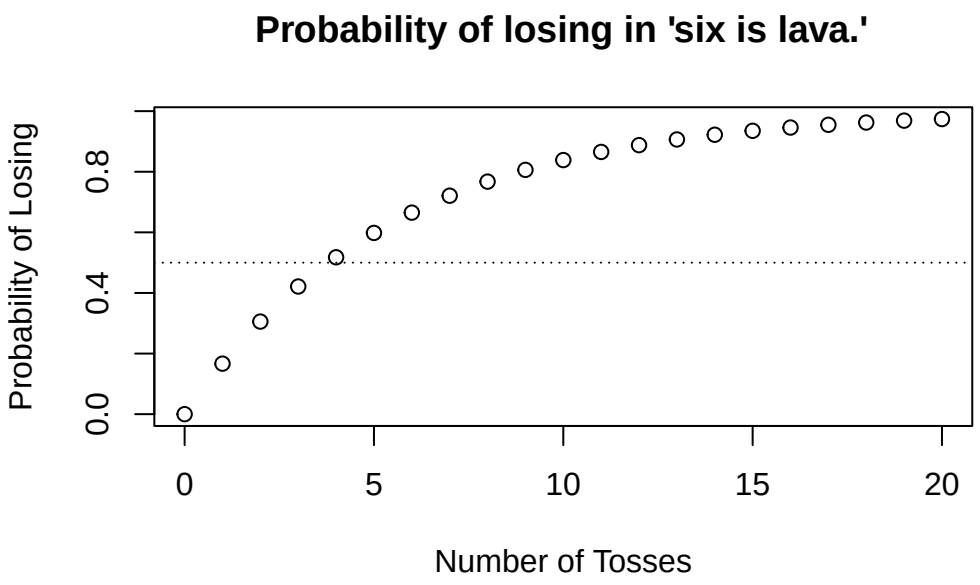
Once again this is a scenario where using the complement simplifies the problem. If we asked "what is the probability that no 6's are rolled, on 10 rolls of the die" then we can count the number of possibilities through an application of the product rule. In particular, there are going to be 5 options which are not 6 at each possible step. We can view this as have $n_1 = n_2 = \cdots = n_{10} = 5$. Thus, the total number of ways of rolling **no** sixes is $5 \times 5 \times 5 \times \cdots \times 5 = 5^{10}$.
Essentially the same process can be used to count the total number of possible rolls, replacing 5 with 6 to get the denominator. This means that there are 6^{10} total sequences of 10 rolls, and 5^{10} which contain no sixes. As a result, taking E to represent the probability that we observe no sixes on 10 rolls of the die, we would get

$$P(E) = \frac{5^{10}}{6^{10}}.$$

The question asks for E^C and so we take

$$P(E^C) = 1 - P(E) = 1 - \left(\frac{5}{6}\right)^{10}.$$

This is approximately 0.8385.
Note that if instead of 10 flips they had n flips, the probability would be $1 - (5/6)^n$. We can plot this over various values of n, to see how the number of flips impacts this probability. The probability of 0.5 is marked and we can see that taking 4 tosses gives an ever so slightly greater than 0.5 probability of losing (0.5177).



3.4.2 Tree Diagrams

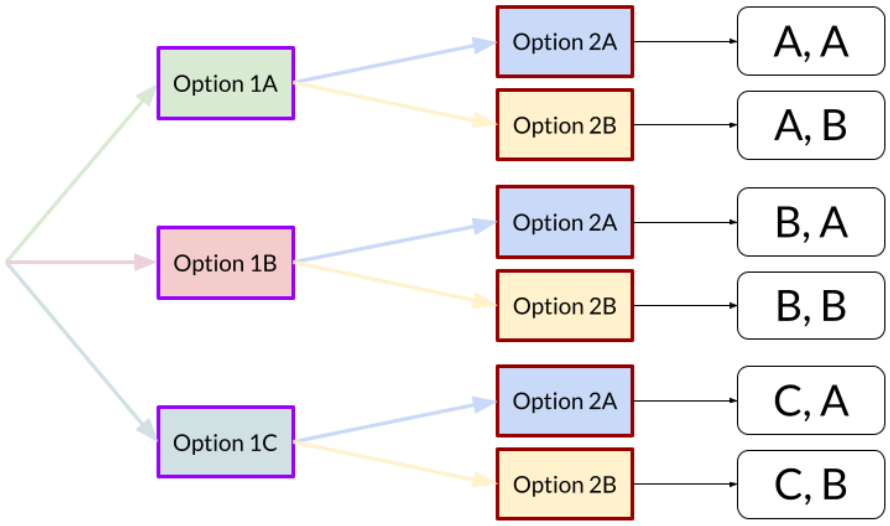
Sometimes it is helpful to express counting rules graphically. To do so we rely on tree diagrams.

Definition 3.2 (Tree Diagram)**.** A tree diagram is a graphical representation for the product rule of counting. Specifically, a tree diagram puts each of the decisions in sequence, and draws a branch for each separate option, starting from the branches drawn at the previous decision step.

To draw a tree diagram, you start with the first choice, drawing one branch for each of the n_1 alternatives, labelling each. Then, at the second choice, you do the same process at the end of each of the branches you drew for choice 1, this time drawing n_2 branches there (so you will have just drawn $n_1 \times n_2$ branches). Then for each of those you draw the n_3 further branches, and so on and so forth until the end.

If you want to know the total number of choices, you simply count the end points at the very end of the diagram. Each branch corresponds to a single option. To determine which combination of choices it corresponds to, you simply read off the branch labels at each branch you take. If you want to know how many possible combinations come with certain options selected, you can look at only those branches which are downstream from the choices that you care about.

Figure 3.1: A generic tree diagram. Here the first choice has three different options and the second choice has two. We can see the six total combinations, labelled on the right of the diagram, and can trace the choices required to get there.



While tree diagrams can be quite useful for visualizing a problem, they often grow to be overly complex. As a result, we need to fall back on the numerical representation afforded to us through the product rule for counting. Counting problems, in general, can very quickly become tremendously large and complex. For this reason, we have several tools to assist us in reducing this complexity based on common types of problems that we would like to count.

3.4.3 The Factorial

The first useful tool for simplifying these problems is the **factorial**. The factorial of an integer, denoted by $x!$ is given by the product of all integers from x to 1.[13] That is,

$$x! = x(x-1)(x-2)\cdots(2)(1).$$

If we consider the product rule for counting then note that if $n_1 = 1$, $n_2 = 2$, … , $n_k = k$, then the total number of options is $k \times (k-1) \times \cdots \times 1 = k!$. The most common reason that this comes up is when we want to order a collection of items. Suppose that you have 10 books that you want to place on a shelf. You can view this as making 10 sequential decisions: what book

[13] Factorials are exciting because they always look like they are shouting!

goes first, second, third, and so on. There are 10 options for the first book, then 9 for the second (any except for the first one), and then 8 for the third (any except for the first 2). This continues down to the last book, and so we conclude that there are $10 \times 9 \times 8 \times \cdots \times 1 = 10!$ ways of arranging these books.

Example 3.7 (Seating in a (Full) Coffee Shop)**.** One day Charles and Sadie walk into the coffee shop and find that it is completely full. There are ten seats and ten people sitting in them. They are disappointed that they do not have room to sit themselves, however, they are never ones to pass up an interesting probability question.

a. How many different ways could these ten people have sat in these ten seats?
b. If there are ten drinks that have been made, and one is to be passed out to each seat, how many different ways can these ten people sit in these ten seats, with each of these ten drinks?
c. Alongside the ten drinks, there are ten snacks to be served up as well. How many different ways can these ten people sit in these ten seats, with each of these ten drinks, and each of these ten snacks?

Solution

a. Here, we can think about lining up the ten seats in a row, each number 1 through 10. Then, we want to place one patron into each of the seats. This is no different from ordering 10 books, and so the total is 10! which is $3,628,800$.

b. Note that we still have to make the seat choices from part (a), so there are 10! ways of getting the 10 people sat in the 10 chairs. Once there, we can think of handing out the drinks to each of the numbered combinations of person-chair. This is no different from passing out the people as well, giving 10! ways of doing this. We use the product rule to combine these two choices, with $10! \times 10!$ total combinations of people-chair-drink. This is $13,168,189,440,000$.

c. Extending the same logic before, there are $10! \times 10!$ ways of getting each person sat in a chair with a drink. Then, there will be an additional 10! ways of passing out the snacks to these people. Taken together this gives $10! \times 10! \times 10!$ which is $47,784,725,839,872,000,000$.[14]

[14]It is somewhat interesting to note how large these values get relatively quickly. This is a comparatively small question: only 10 people with 10 drinks and 10 snacks. If you consider any counting problem with a larger number of items, these problems quickly grow to be intensely complex. For instance, a count of my main home book collection reveals 376 of them. To order these would give 376! possible orderings. In decimal representation this is

$$4.992244775852435618292576458782762114148884082811840265632\cdots \times 10^{806}.$$

This is an 807 digit long number. This is an incomprehensibly large number. This number is 8×10^{726} **times larger than** the number of atoms in the universe. That is, if every atom in the universe were given

Remark (0!). Depending on how factorials are thought of, some trouble can come up around a quantity like 0!. On one hand, if we view factorials as multiplying each number between n and 1 together then $0! = 0 \times 1$ and we get $0! = 0$. On the other hand, if we view factorials as counting the number of ways which we can order a set of n items, then 0! is the number of ways we can order 0 items, which is 1.[15] So, which is it?

We take 0! to be equal to 1. The ordering argument is perhaps the most convincing. However, if you are algebraically minded you may wonder how we get around the tricky issue of using our algebraic definition. The key insight is to not define $n!$ as the product of the numbers from n to 1, but rather, to define

$$(n-1)! = \frac{n!}{n},$$

and specify that $1! = 1$. Then in this case we get all of the usual requirements for how we have discussed factorials, but we also get that $0! = \frac{(0+1)!}{(0+1)} = 1$. As a result, we will take $0! = 1$.[16]

3.4.4 Permutations and Combinations

Sometimes, we want to order items from a collection, but we want to only take a subset of these times. That is, suppose that you have 20 books, only 9 of them will fit on the self, and you want to know "how many ways can you put 10 books on the shelf, in order, from your collection of 20?" Using the product rule of counting for this directly, we recognize that there are 20 options for the first, then 19, then 18, and so on until there are 12 choices for the 10th book to place. We can write this out in a seemingly strange way.

$$\frac{(20)(19)(18)(17)(16)(15)(14)(13)(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}$$
$$= \frac{(20)(19)(18)(17)(16)(15)(14)(13)(12)\cancel{(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}}{\cancel{(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}}$$

This expression is 20! divided by 11!, and gives the same as our argument from the product rule for counting directly. This is a more general result than our example with books would suggest. If we have n items, and we want to choose k of them taking into account the order those choices, it will always be $n!$ divided by $(n-k)!$. We call this a permutation.

---

some arrangements of books to hold onto, they would need to hold 8×10^{726} of them in order for all of the arrangements to be held. I point this out because combinatorics *explodes* in this way. Even simple problems grow out of hand very, very quickly. This is where comfort with the algebraic tools is required, rather than a reliance on intuition. There is simply no way to have intuition regarding the scope of these numbers, at least, not without a lot of practice.

[15] Imagine I am placing books on my shelf. With 3 books there are $3! = 6$ ways my shelf can look at the end. With 2 books there are $2! = 2$ ways my shelf can look at the end. With 1 book there are $1! = 1$ ways my shelf can look at the end. With 0 books there are $0! = 1$ ways my shelf can look at the end.

[16] Note, this does not help us with the factorials of negative numbers, nor of fractional numbers. Factorials *can* be extended to these in sensible ways, but these are not for combinatorial purposes and are no longer "factorials" exactly.

Definition 3.3 (Permutations)**.** If we wish to select k items from a collection of n items, where the ordering of these selections matters, then the total number is referred to as a permutation. Mathematically,

$$P_{n,k} = \frac{n!}{(n-k)!}.$$

Permutations arise when we select ordered subsets from a collection. We often, in combinatorial problems, talk about ordering, though sometimes what we mean by this is slightly more abstract. Suppose that you want to form a committee with 5 different people, each of which occupies a different role: the president, vice president, treasurer, note taker, and critic. If there are 30 people to select for this then there are $P_{30,5}$ total possible committees that can be formed. While there is not a sequential order here, we talk about this as being "ordered" since we can differentiate between the five roles. Instead of labelling them with their names, we could label them 1 through 5 and make the ordering more explicit.

Example 3.8 (Seating in a (Not Full) Coffee Shop)**.** Still haunted by that time when the coffee shop was full, Sadie and Charles enter the coffee shop at a later date and find that, including themselves, there are only 7 patrons in the store, and still the 10 seats to choose from.

a. How many different ways can the 7 people sit in the 10 different chairs?
b. If there are 10 drinks on the menu, how many different ways can each person choose a chair and a drink?
c. If the coffee shop can make only one of each drink, how does the previous total change?

Solution

a. In this case we are looking to order subsets from a total collection. If we line up the 7 patrons we then need to select 7 chairs to go with them, and keep these ordered. This is simply

$$P_{10,7} = \frac{10!}{3!} = 604800.$$

b. In this part of the question we know that the 7 people have $P_{10,7}$ ways of sitting into seats, we can view this as decision one. Then, for each of these 7 people there is a decision of what drink they will order. For these 7 decisions, $n_2 = \cdots = n_8 = 10$. As a result, we get the total number is

$$P_{10,7} \times 10 \times 10 \times \cdots \times 10 = P_{10,7} \times 10^7 = 6,048,000,000,000.$$

c. This setting, like part (b) starts with a first decision involving $P_{10,7}$ choices. Then, instead of there being 7 more decisions with 10 choices each, we can either view this as 7 more choices with a descending number of options 10 for the first, then 9,

and so on, or we can view this as a single choice where we need to select 7 **ordered** options from the 10 drinks available. This gives

$$P_{10,7} \times P_{10,7} = 365,783,040,000,$$

total choices.

Factorials compute the number of orderings for a set of objects, and permutations compute the number of ordered subsets from a collection of objects. What about when we do not wish to differentiate the order of subsets? Suppose that you still need to form a 5 person committee, but you do not have explicit roles for the different members of the committee. Here we cannot use a permutation directly, as we know that this takes into account the order.

To determine the number of unordered subsets, we will consider a different approach for taking ordered subsets. Suppose that we formulate the ordered committee as a two step procedure. First, we select 5 people without concern for their order. Then we choose which order they will have. If M represents the number of unordered sets of 5 from this population, the product rule for counting tells us that the total number of ordered committees will be $M \times 5!$, since there are 5! arrangements of the 5 people. Thus, we can write this down as

$$P_{30,5} = \frac{30!}{25!} = M \times 5! \implies M = \frac{30!}{25!5!}.$$

This will be true far more broadly than our committee example. If we want to select k items from a collection of n, we will have $n!$ divided by the product of $k!$ and $(n-k)!$. We refer to these as combinations.

Definition 3.4 (Combinations)**.** If we wish to select k items from a collection of n items, where the ordering of these selections does not matter, then the total number is referred to as a combination. Mathematically,

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

We read $\binom{n}{k}$ as "n choose k", which translates to "select k items from a population of n total options, without concern for their order."

To summarize: factorials allow us to order a complete collection, permutations allow us to select a subset with consideration of the ordering, and combinations allow us to select a subset from the collection without regard to the order. These three techniques can be used in combination with the product rule for counting to allow us to have very complex total summations.

Example 3.9 (Changing the Seating in the Coffee Shop)**.** Some nights, the coffee shop hosts local music acts. Because of the added equipment, the coffee shop owners only keep out the number of seats that are going to be required based on the number of tickets that were sold.

a. If there are 8 tickets sold, how many different combinations of the 10 chairs can get left out?
b. Suppose that only 6 people end up showing up. How many different ways can the 6 people sit in the 8 chairs that are being selected from the 10 total possibilities?

Solution

a. Here, there is no ordering for the 8 chairs that are to be selected. As a result, we are simply looking for how many chairs can be selected from a group of 10 of them. This is
$$\binom{10}{8} = \frac{10!}{8! \cdot 2!} = 45.$$

b. With the first choice having 45 possible options, the second choice that needs to be made is how 6 people sit into 8 chairs. Here the ordering *does* matter, since the chairs are distinguishable. As a result, this is a permutations question, with there being a total of $P_{8,6} = \frac{8!}{2!}$ ways of having the people select their seats. In total then there are
$$\binom{10}{8} \times P_{8,6} = \frac{10!}{8! \cdot 2!} \cdot \frac{8!}{2!} = 907,200.$$

3.4.5 Less Common Counting Techniques

While most of the problems we address will revolve around permutations and combinations (with heavy use of the product rule), there are additional techniques which are important to know (and recognize when to use). In particular, combinations and permutations each assume that we are sampling from our set **without replacement**. That is, each time you select an item, it is removed from the population. These are the most common situations in these combinatorial problems, however, there are *some* situations which arise where we need to count the number of ordered or unordered subsets *with* replacement.

3.4.5.1 Ordered Subsets with Replacement

Consider, for instance, forming a password using only lowercase numbers and letters. If you decide on a fixed length for the password, then there are going to be 36 choices at each decision point, and you want to take an ordered subset of these. This is forming an **ordered subset with replacement**, and to count how many different ways there are of doing this, we can simply use the product rule. That is, you have 36 choices at each decision point, and so there are $36 \times 36 \times \cdots \times 36 = 36^k$ total decisions, where k is the number of items to select.

In general, if you have n total items and you want to make an ordered set of k of these items **with replacement** you will have n^k total ways of doing this.

3.4.5.2 Unordered Subsets with Replacement

Forming unordered sets with replacement is slightly less intuitive. Consider rolling k dice which are not distinguishable from one another. We know that there are 6 total sides that can show up on each of these dice, but how many different combinations of numbers can show up overall? If the dice can be distinguished we would say that there are 6^k possible ways of doing this. However, some of these combinations are going to be equivalent in the unordered world. Take the simple case of $k = 2$. Here we have the following possibilities:

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

This gives a total of 21 possible combinations, rather than 36.

In general, if we want to find the way of selecting k elements *with replacement* from a total of n, then the number of ways of doing this will be

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

In our example this gives $\binom{6+2-1}{2} = 21$.

3.4.5.3 Permutations with Identical Objects

Finally, it is worth understanding how to handle identical objects in combinatorial problems. Suppose that, of the 10 books that we wish to place on a shelf, we have 2 copies of one of them, 3 copies of another one, and the other 5 have one copy each. Supposing that there is no way to tell these identical objects apart, how many ways can we arrange the bookshelf?

First, if we pretend that all of the items are actually able to be differentiated then there are 10! ways of placing these books. Now, in any of these permutations, had we swapped the order of the first book (with 2 copies) the ordering would have been indistinguishable. As a result, for every ordering of the 8 other books, we counted that permutation twice (when it should have only been counted once!). So to address the two repeated copies we need to take 10!/2. Now, a similar argument is going to hold for the book with 3 repeated copies. However, instead of there being 2 permutations which are identical, there are going to be $3! = 6$ permutations which are identical. This is because we can reorder the 3 copies of the book in anyway we

choose, and still wind up with the same overall permutation.[17] As a result, the total number is going to be

$$\frac{10!}{2! \cdot 3!} = 302400.$$

We can see this same result through an alternative construction. First, we select which of the 10 slots should have the first book. We do not care about the order, and so there are $\binom{10}{2}$ ways of doing this. Next, we can select which of the 8 remaining slots should have the second book. Like the first one there will be $\binom{8}{3}$ ways of placing these. Now, there are 5 slots remaining, and 5 books to place, so as a result, we can order those in 5! different ways, and then slot them into the remaining places in order. This gives, in total

$$\binom{10}{2}\binom{8}{3}(5!) = \frac{10!}{2!\cancel{8!}} \cdot \frac{\cancel{8!}}{3!\cancel{5!}} \cdot \cancel{5!} = \frac{10!}{2!3!}.$$

To generalize this, if we want to order n elements, such that there are k distinguishable elements with n_1 of the first type, n_2 of the second, and so forth until n_k of the last type ($n = n_1 + n_2 + \cdots + n_k$), then the total number of orderings will be

$$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}.$$

3.5 From Combinatorics to Probability

While combinatorics is a field of study on its own, with many intriguing tools and developments surrounding the enumeration of objects, for the purposes of simple probability models these tools will suffice. Ultimately, we care about counting since in the equal probability model, the probability of any event can be determined by counting the number of ways that the event can occur and dividing by the total number of outcomes that are possible. That is, we use these tools to derive N_A, the total number of ways that A can occur, and N, the total number of experimental outcomes, and then we conclude that

$$P(A) = \frac{N_A}{N}.$$

Example 3.10 (Poker Hand Counts)**.** During one of their conversations, Charles and Sadie were remarking how they never really played poker. As they understand it, in poker you are dealt a hand of 5 cards and you want to use these 5 cards to try to match certain sets of cards, some of which are more rare than others. Charles and Sadie start to get hung-up on discussions regarding "straights" and "flushes".

A straight is any sequence of 5 cards in ascending order (where aces can be low, or high). For instance, $7, 8, 9, 10, \text{J}$ of any suit. A flush, is any set of 5 cards belonging to the same suit. Charles just *feels* that straights have to be more rare than flushes.

[17] In fact, the reason that there are 2 ways of doing this with the book with 2 copies is since $2! = 2$.

a. How many different straights are there from a standard deck of cards?
b. How many different flushes are there from a standard deck of cards?
c. If dealt 5 cards at random, what is the probability of a flush? What is the probability of a straight?
d. A straight flush occurs when you have 5 cards in order, of the same suit. What is the probability of a straight flush?
e. If straight flushes were not counted as flushes, and not counted as straights, how do the probabilities of either hand change?

Solution

a. A straight necessitates drawing five cards in order, with each of any suit. Just as with the straight flush, there are 10 possible starting values for the straight. Once we have selected the starting value, then for each of the five cards we can pick any of the four suits, resulting in 4 choices each. That gives

$$N_A = 10 \times 4 \times 4 \times 4 \times 4 \times 4 = 10 \times 4^5 = 10240.$$

b. A flush necessitates drawing **any** five cards from the same suit. If we had a suit fixed, there would be $\binom{13}{5}$ ways of doing this, since we do not care about ordering. If we think about first choosing the suit, we have 4 ways of doing that, resulting in

$$N_A = 4 \times \binom{13}{5} = 5148.$$

c. To find the probabilities of each of these, we need to know the total number of 5 card hands. We do not consider order, and so $N = \binom{52}{5} = 2598960$. Then, the probability is simply the number of combinations (calculated above) divided by the number of hands. This gives, for straights,

$$P(A) = \frac{10240}{2598960} = \frac{128}{32487} \approx 0.00394,$$

and for flushes,

$$P(A) = \frac{5148}{2598960} = \frac{33}{16660} \approx 0.00198.$$

As a result, we see that straights are roughly twice as common as flushes are.[18]

d. Note that to form a straight flush, we first have to fix a suit. There are $\binom{4}{1} = 4$ total ways of doing this. Next, we need to pick which starting value we will use. Once a card has been selected as a starting value, the remaining cards are fixed. The start value ranges from A through to 10. Correspondingly, we have

$$N_A = \binom{4}{1}\binom{10}{1} = 4 \times 10 = 40.$$

As a result, we get that

$$P(A) = \frac{40}{2598960} = \frac{1}{64974}.$$

e. From (d) exactly 40 of the straights and 40 of the flushes are also straight flushes. We can thus remove 40 from the totals of each of these, giving $\frac{10200}{2598960}$ for straights and $\frac{5108}{2598960}$ for flushes.

Exercises

Exercise 3.1. The probability that a smartphone breaks during the first year of use is 0.19. What is the probability that it does not break during the first year?

Exercise 3.2. Assuming that A and B are disjoint, which of the following statements are always true.

a. $P(A \cup B) = 0$.
b. $P(A \cap B) = 0$.
c. $P(A \cup B) = P(A \cap B)$.
d. $P(A \cup B) = P(A) + P(B)$.

Exercise 3.3. A ball is drawn at random from a box containing 6 red balls, 4 white balls, and 5 blue balls. Determine the probabilities associated with the following events.

a. The ball is red.
b. The ball is white.
c. The ball is blue.
d. The ball is not red.
e. The ball is red or white.

[18]This is *still* a deeply unintuitive result to me. To me it feels like it should be harder to get a nice ordering of cards all in a row than it is to find ones of the same suit. And yet, it is about twice as likely to have the straights. Now, if you think about this deeply, it makes a lot of sense: there is a lot more leeway in selecting the straight than the flush. However, my brain refuses to accept this as intuitive. This is something that can often occur in probability questions where the true results can be far from what we would expect. An interesting question is how long does a straight have to be to be less likely than a flush of the same length. Note that the number of ways of choosing a flush will always be $4 \times \binom{13}{k}$, and the number of ways of forming a straight will always be 10×4^k. If we consider k from 1 to 13, we can take the ratios of these to see just how much more (or less) likely a straight is. Doing this gives: k=1 gives 0.769. k=2 gives 0.513. k=3 gives 0.559. k=4 gives 0.895. k=5 gives 1.989. k=6 gives 5.967. k=7 gives 23.869. k=8 gives 127.304. k=9 gives 916.587. k=10 gives 9165.874. k=11 gives 134432.821. k=12 gives 3226387.692. k=13 gives 167772160. Up to (and including) 4 cards the straight is indeed harder to achieve. However, this does not last long!

Exercise 3.4. A survey of high schoolers were asked about where they most prefer to spend their time when they are at home. The results were:

- Bedroom: 0.37;
- Living Room: 0.26;
- Den: 0.22;
- Basement: 0.12;
- Kitchen: 0.02;
- Bathroom: 0.01.

a. What is the probability that a student prefers being in their living room or den?
b. What is the probability that a student does not prefer being in their bedroom?

Exercise 3.5. Let S be the event that a randomly selected college student has taken any statistics course, and C be the event that a randomly selected college student has taken any chemistry course. Suppose $P(S) = 0.5$, $P(C) = 0.15$, and $P(S \cap C) = 0.12$.

a. Find the probability that the student has taken either a statistics or chemistry course.
b. Find the probability that a student has taken neither chemistry nor statistics.
c. Find the probability that a student has taken statistics but not chemistry.

Exercise 3.6. Let M be the event that a student passes their math exam, and let H be the event that a student passes their history exam. Suppose that $P(M) = 0.85$, $P(H) = 0.95$, and $P(M \cup H) = 0.99$.

a. Find the probability that the student passes both math and history.
b. Find the probability that the student passes neither math, nor history.
c. Find the probability that the student passes math, but not history.

Exercise 3.7. A medical clinic screening for a particular disease classifies its patients into three health risk categories: low risk, moderate risk, and high risk. Suppose that 70% of the patients are categorized low risk, 20% are categorized as moderate risk, and 10% are categorized as high risk. Suppose a patient is randomly selected.

a. What is the probability that the patient is categorized as low risk?
b. What is the probability that the patient is not categorized high risk?

Exercise 3.8. Two different designs are being considered for a particular mechanical system. For each, determine the probability that the system continues to function.

a. The system contains two components, A and B. The system will continue to function so long as either A or B functions. The probability that A functions is 0.95, the probability that B functions is 0.90, and the probability that the both function is 0.88.

b. The system contains two components, A and B. The system will continue to function only if both A and B function. The probability that A functions is 0.98, the probability that B functions is 0.95, and the probability that at least one is functioning is 0.99.

Exercise 3.9. A hotel chain is concerned about the cleanliness of their rooms, so they send two inspectors to check on this. Inspector A visually inspects 1000 rooms and found 37 of them to be inadequately cleaned. Inspector B visually inspects the same rooms found 43 of them to be lacking sufficient cleanliness. A total of 948 rooms were found to be clean by both inspectors.

a. Find the probability that a randomly selected room has an issue that was found by at least one of the two inspectors.
b. Find the probability that a randomly selected room has an issue that was found by both inspectors.
c. Find the probability that a randomly selected room has an issue that was found by inspector A but not by inspector B.

Exercise 3.10. Let A and B be events with probabilities $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{3}$. Show that

$$\frac{1}{12} \leq P(A \cap B) \leq \frac{1}{3}.$$

Exercise 3.11. Find the probability of a 4 turning up at least once in two tosses of a fair die.

Exercise 3.12. It is required to place 5 hardcovers and 4 paperbacks onto a bookshelf, so that each of the paperbacks are between two hardcovers. How many ways can this be done?

Exercise 3.13. A committee of eight people must choose a president, a vice-president, and a secretary. In how many ways can this be done?

Exercise 3.14.

a. One drawer in a dresser contains 8 blue socks and 6 white socks. Another drawer contains 4 blue socks and 2 white socks. One sock is chosen from each drawer. What is the probability that they match?
b. A third drawer contains 6 red socks, 4 green socks, and 2 black socks. Two socks are selected at random. What is the probability that they match.

Exercise 3.15. A company has hired 15 new employees, and must assign 6 to the day shift, 5 to the graveyard shift, and 4 to the night shift. In how many different ways can the shifts be made?

Exercise 3.16. A group of 10 people have gotten together to play basketball. They will begin by dividing themselves into two teams of 5 players each. One team will wear red and the other blue. How many ways can this be done?

Exercise 3.17. A factory produces a total of $15,000$ candies a day, of which 3% are defective. Find the probability that, out of 500 candies chosen at random on a single day, 12 are defective.

Exercise 3.18. A shelf has 6 mathematics books and 4 physics books. If these were randomly ordered, what is the probability that 3 particular mathematics books will be together.

Exercise 3.19. Suppose that car license plates consist of three letters followed by three numbers.

a. How many different license plates can be made?
b. How many different license plates can be made in which no letter or number appears more than once?
c. A license plate is chosen at random. What is the probability that no letter or number appears more than once?

Exercise 3.20. A game has individuals forming random "words" (strings of letters, whether they have a meaning or not). On a particular turn, one player must form a word which has 3 different vowels, and 4 different consonants, where there are 7 total consonants and 5 total vowels to choose from. How many possibilities are there for a valid word?

Exercise 3.21. Six pairs of socks come in sets: there are two pairs of red socks, two pairs of white socks, and two pairs with stripes on them. If the socks are placed randomly into pairs, find the probability that no sock ends up in a pair with the same pattern.

References

4 Probabilities with More than One Event

4.1 Marginal and Joint Probabilities

Up until this point we have primarily focused on assigning probabilities to particular events. If we have some event of interest, A, then $P(A)$ is the probability that A occurs in any manner. If we are using our equally likely probability model, then $P(A) = \frac{N_A}{N}$. This is the probability of the event A where nothing else is known at all. If we smooth over anything which could alter the likelihood, if we have no additional information, if we want the best guess for the likelihood of occurrence in a vacuum, this is the probability of interest. We refer to such quantities as marginal probabilities.

Definition 4.1 (Marginal Probability)**.** The marginal probability of a single event, A, is the probability that the event happens without considering any information. This is simply the probability that the event happens, denoted $P(A)$.

It is useful to specifically call these marginal probabilities to differentiate them from probabilities which depend on two or more events. Specifically, we can think about taking the intersection of two events, say $A \cap B$. In words this is the event that A occurs **and** B occurs. Until this point we have thought about solving for probabilities related to intersections as a two-step procedure: first, find an event $C = A \cap B$, and then second solve for the marginal probability of C. While this is a useful technique for calculation, sometimes it is more useful to think of the probability of the intersection as the probability of A and B occurring simultaneously. We call this the joint probability of A and B.

Definition 4.2 (Joint Probability)**.** The joint probability of two events, A and B, is given by the probability of their intersection. That is, $P(A \cap B)$. This is sometimes denoted $P(A, B)$ and corresponds to the probability that both A and B occur. The joint probability of more than two events extends in the same way. Suppose there exists a sequence of events, $A_1, \ldots, A_n$, then the joint probability of these events is

$$P(A_1, A_2, \ldots, A_n) = P\left(\bigcap_{i=1}^{n} A_i\right).$$

4.2 What are Conditional Probabilities?

Marginal probabilities are often of interest, and frequently are the best tool for summarizing the overall state of the world, or our knowledge regarding the state of the world. Joint probabilities can be useful when working with compound events, or thinking of complex outcomes that we wish to consider. However, there are many scenarios which are not covered by either the joint or marginal probabilities we have seen. In practice we know that sometimes information that we have will change our understanding of the probability of an event. Suppose the event A corresponds to the event that it snows tomorrow, in some particular city. It is possible to think about how often it snows on average, and report a value related to that as $P(A)$. Now, what if we know that it is currently the middle of summer? In this case, while $P(A)$ does not shrink to 0, it becomes far less likely than if we did not have that information. Similarly, if we know that it is winter, the likelihood that it snows tomorrow increases. In order to formally capture this we can introduce the idea of conditional probability.

Definition 4.3 (Conditional Probability)**.** The conditional probability of an event A given an event B, is the probability that A occurs assuming that we know that B occurs. The conditional probability takes into account additional information codified through the occurrence of an additional event. We write this quantity as $P(A|B)$, and will read this as "the probability of A given B."

Unlike in the case of marginal probabilities, conditional probabilities allow us to *condition* on extra pieces of information. Instead of asking "what is the probability of this event", we instead ask, "given that we know this piece of information, what is the probability of this event?" Joint probabilities ask "what is the probability that both of these events occur simultaneously" which is distinct in that we do not have any additional information about the state of the world when working with joint probabilities. The subtle distinction becomes quite powerful, both in terms of manipulating and working with probabilities, but also in terms of expressing the correct events of interest for ourselves.

To make use of conditional probabilities, we will think of the process of *conditioning* on one or more events. We will talk of the probability of A conditional on B, where A and B are two events of interest. Intuitively, this it the probability of A happening, supposing that we know that B has already happened.

Example 4.1 (Six from the Sum)**.** Charles and Sadie are playing a new game using two dice. In the game they each take turns rolling the two dice into a container that the other player cannot see. They add up the dice and then report this sum to the other player. The other player then has to guess whether or not there is a six showing.

a. On one round, Charles does not hear the sum that Sadie reports as he was not paying attention. Sadie is strict and insists that she will not repeat herself. What should Charles guess?

b. Determine a strategy which optimizes the likelihood that the guessing player will be correct.

Solution

a. In this case we want the **marginal probability** of a six showing up on the roll of two dice. Take E to represent the event wherein at least one six is showing on the roll of two dice. Thus, E^C is the event that no sixes are showing. There are $5 \times 5 = 25$ (using the product rule for counting) ways of *not* rolling a 6, meaning that $\$P(E^C) = \frac{25}{36} \implies P(E) = 1 - \frac{25}{36} = \frac{11}{36}$. Thus, Charles should guess that there is no 6 as the probability is only $11/36 \approx 0.31$.

b. Here we wish to determine conditional probabilities. We take E to be the event that at least one 6 is showing, and then S to be a variable representing the sum of the two dice. For $s = 1, \ldots, 12$ we wish to find $P(E|\{S = s\})$. Notice that for $s = 2, \ldots, 6$ $P(E|S = s) = 0$. If the sum is 2 we know that there could not have been a 6. Moreover, for $s = 11, 12$ we know that $P(E|S = s) = 1$ since the only way to form 11 is a five and a six, and the only way to form 12 is to have two sixes. This leaves $s = 7, \ldots, 10$ to check. The following table gives the set of values, the possible combinations to reach the value, and then the combinations that end up involving a 6.

Value	Combinations	Involving 6
7	(1,6) (2,5) (3, 4) (4, 3) (5, 2) (6, 1)	(1, 6) (6, 1)
8	(2,6) (3,5) (4, 4) (5, 3) (6, 2)	(2, 6) (6, 2)
9	(3,6) (4,5) (5, 4) (6, 3)	(2, 6) (6, 2)
10	(4, 6) (5, 5) (6, 4)	(4, 6) (6, 4)

Referencing from this table we can read off the following probabilities

$$P(E|S = 7) = \frac{2}{6} = \frac{1}{3}$$
$$P(E|S = 8) = \frac{2}{5}$$
$$P(E|S = 9) = \frac{2}{4}$$
$$P(E|S = 10) = \frac{2}{3}.$$

As a result, the best strategy is to guess "yes" when a 10, 11, or 12 is rolled and to be indifferent to the guess when a 9 is rolled. Otherwise, guess "no."

Recall that A and B, as events, are merely subsets of the sample space, $\mathcal{S}$. Each item in either

A or B is one of the possible outcomes from the experiment or process that we are observing. Suppose that we know that B has occurred. What this means is that, one of the outcomes in the set B was the observed outcome from the experiment. Now, if we want to know $P(A|B)$, we want to know the probability, working from the assumption that B has happened, that A also happens. That is, knowing that B has happened, what is the probability that A and B both happen.

The event that A and B both happen is denoted by the intersection, $A \cap B$. This corresponds to the set of events inside the set B which also belong to the set A. Now, instead of considering the joint probability directly, we need to acknowledge that for $A|B$, only the events in B were possible. That is, instead of being divided by the whole space, we can only divide by the space of B. In some sense, we can view conditioning on B as treating B as though it is the full sample space, and finding probabilities within that. In general, B will be smaller than $\mathcal{S}$, and so $P(B) < 1$. Instead of the conditional probability being "out of" 1, it will instead be "out of" $P(B)$, which gives $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Computing Conditional Probabilities

For an event A, and an event B with probability $P(B) > 0$, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

To make this more clear, let's consider a simple example. Suppose that we take A to be the event that a 2 is rolled on a fair, six-sided die, and B to be the event that an even number was rolled. This is an equal probability model, and so each outcome gets $\frac{1}{6}$ probability. The original sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, the event A is $\{2\}$, and the event B is $\{2, 4, 6\}$. In order for both A and B to occur, we note that we need $A \cap B = \{2\}$. If we know that B has occurred, then we know that either a 2, 4, or 6 has been rolled, with equal probability for each. Thus, intuitively, we can view B as the new sample space, and say that rolling a 2 has a $\frac{1}{3}$ probability, given that there are 3 outcomes and 1 of them is the event of interest. Another way to consider this is to note that $P(B) = \frac{1}{2}$, and so we need to scale each event by $\frac{1}{1/2}$ in order to make sure that the total probability of our reduced sample space equals 1. Then $P(A \cap B) = \frac{1}{6}$, so

$$P(A|B) = \frac{1/6}{1/2} = \frac{1}{3}.$$

Suppose that, instead of a fair die, it was weighted so that 6 comes up more frequently than the other options. Consider the probability of observing a six to be 0.5, with the other five values each coming up with probability 0.10. If A and B are the same events as above, then $P(B) = 0.7$. If we know that B has occurred then the new sample space is $\{2, 4, 6\}$ where $P(2) = \frac{0.1}{0.7} = \frac{1}{7}$, $P(4) = \frac{0.1}{0.7}$, and $P(6) = \frac{0.5}{0.7} = \frac{5}{7}$. Note that these three probabilities sum to 1 still, which constitutes a valid probability model, and so $P(A|B) = \frac{1}{7}$.

Example 4.2 (Six from the Sum - Revisited)**.** Sadie sees the solution worked out for the best strategy in the dice game above, but is having trouble understanding it in the context of conditional probability more broadly. For the sums $7, 8, 9$, and 10, describe the process of limiting the sample space to get the correct conditional probability, and use the formula to show that the probabilities worked out in Example 4.1 are correct.

Solution

The relevant table of solutions is copied from above.

Value	Combinations	Involving 6
7	(1,6) (2,5) (3, 4) (4, 3) (5, 2) (6, 1)	(1, 6) (6, 1)
8	(2,6) (3,5) (4, 4) (5, 3) (6, 2)	(2, 6) (6, 2)
9	(3,6) (4,5) (5, 4) (6, 3)	(2, 6) (6, 2)
10	(4, 6) (5, 5) (6, 4)	(4, 6) (6, 4)

Here, given a sum, we **know** that one of the events listed under "combinations" has occurred. As a result, once we have conditioned on what the sum is, we are able to treat the "combinations" column as the full sample space of possible outcomes. Each of these in this case is equally likely, and so we can divide the number which contain a 6, by the total number in the reduced sample space.
To do this algebraically, we first note that

$$\begin{aligned} P(S=7) &= \frac{6}{36} \\ P(S=8) &= \frac{5}{36} \\ P(S=9) &= \frac{4}{36} \\ P(S=10) &= \frac{3}{36}. \end{aligned}$$

Now, suppose we consider the event $E \cap \{S=7\}$. This is the set $\{(1,6),(6,1)\}$ and so $P(E \cap \{S=7\}) = 2/36$. In fact, the same holds true for all of the joint events. As a result, we get

$$\begin{aligned} P(E|S=7) &= \frac{\frac{2}{36}}{\frac{6}{36}} = \frac{2}{6} \\ P(E|S=8) &= \frac{\frac{2}{36}}{\frac{5}{36}} = \frac{2}{5} \\ P(E|S=9) &= \frac{\frac{2}{36}}{\frac{4}{36}} = \frac{2}{4} \\ P(E|S=10) &= \frac{\frac{2}{36}}{\frac{3}{36}} = \frac{2}{3}. \end{aligned}$$

This corresponds exactly to the values we found before.

Sometimes, we wish to condition on more than one event. To do so, the same process extends naturally. For instance, suppose we want to know the probability of A given B and C. This would be written

$$P(A|B,C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A,B,C)}{P(B,C)}.$$

Moving beyond two events occurs in the expected way.

4.3 Using Conditional Probabilities

Conditional probability is a mechanism for capturing our knowledge of the world, and using that to update our sense of the uncertainties at play. For instance, suppose that we are interested in drawing a random card from a deck of 52, and we want to know the probability that it is a heart. Without any additional knowledge, the probability of this event is $\frac{1}{4}$. Now, suppose that you know that it is a red card. In this case, we now know that it is either a heart or a diamond, and there are equal numbers of each, meaning that the new probability is 0.5. We can work this out directly

$$P(\text{Heart}|\text{Red}) = \frac{P(\text{Heart}, \text{Red})}{P(\text{Red})} = \frac{P(\text{Heart})}{1/2} = \frac{1/4}{1/2} = 0.5.$$

Suppose instead that we had been told that the card was an ace. Here we now know that there are four possible outcomes that correspond to an ace, and only one of these is a heart, meaning the probability is $\frac{1}{4}$. In this case, $P(A|B) = P(A)$, and our beliefs did not update.

What if instead we had considered the second event to be "the card was a spade." In this case if we want to know $P(A|B)$ then, given a spade being drawn, we know that the probability of drawing a heart is 0.

Example 4.3 (Charles's Mismatched Urn Mishap)**.** Charles's love of probability prompts a spontaneous decision: buy an urn and some coloured balls to fill it up with. Unfortunately, the supplier of the balls misunderstood the request and sent over an assortment of different shapes rather than just spheres. There are some spheres, some cubes, some pyramids, and some cones. There are also five different colours present, red, blue, green, yellow, and black. Charles is slightly dismayed as, when reaching into the urn, it is very easy to feel what shape you are pulling out before you see the object, and the distribution of colour-shape combinations is not even. Despite the dismay, Charles shows Sadie, who is deeply excited, explaining how this mismatched urn is perfect for understanding conditional probabilities deeply. The distribution of shapes and colours is presented in the following table.

Colour	**Sphere**	**Cube**	**Pyramid**	**Cone**
Red	2	3	2	0
Blue	1	0	0	6
Green	2	2	1	2
Yellow	0	4	2	1
Black	3	1	1	2

a. What is the probability of drawing each colour, if items are selected at random, without knowledge of the shape?
b. Assuming that the shape is known, what is the probability of selecting each colour?

Solution

a. Note that there are $2+3+2=7$ red, $1+6=7$ blue, $2+2+1+2=7$ green, $4+2+1=7$ yellow, and $3+1+1+2=7$ black objects. As a result, each colour is going to be equally likely, with a probability of 0.2.

b. Take S, C, P, Cn to be the events that a sphere, cube, pyramid, or cone are drawn, respectively. Take R, B, G, Y, Bk to be the events that a red, blue, green, yellow, or black object is drawn, respectively. We want the probability of each colour, given the corresponding shape. There are a total of 20 conditional probabilities to solve for here. We walk through, in full, the first conditional probability calculation. Afterwards, the same process follows to get the (provided) answer.

First note that there are 8 spheres, 10 cubes, 6 pyramids, and 11 cones. Thus, the marginal probabilities are $P(S)=8/35$, $P(C)=10/35$, $P(P)=6/35$, and $P(Cn)=11/35$. Note further that the joint probability between any colour-shape combination is going to be given by the number of that combination that exist, divided by 35. Thus, suppose we want $P(R|S)$. There are 2 red spheres, so $P(R\cap S)=2/35$. The marginal probability $P(S)=8/35$, so taken together this gives

$$P(R|S)=\frac{2/35}{8/35}=\frac{2}{8}=\frac{1}{4}.$$

Applying the same process gives the following probabilities, reported in the table for convenience.

Colour	**Sphere**	**Cube**	**Pyramid**	**Cone**
Red	$\frac{2}{8}=\frac{1}{4}$	$\frac{3}{10}$	$\frac{1}{3}$	0
Blue	$\frac{1}{8}$	0	0	$\frac{6}{11}$
Green	$\frac{2}{8}=\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{2}{11}$
Yellow	0	$\frac{2}{5}$	$\frac{1}{3}$	$\frac{1}{11}$
Black	$\frac{3}{8}$	$\frac{1}{10}$	$\frac{1}{6}$	$\frac{2}{11}$

4.3.1 The Multiplication Rule

While sometimes we will want to work out the conditional probability using our knowledge of the joint and marginal probabilities, there are other times where it is easier to determine the conditional probability directly. In these settings we may wish to understand the marginal or joint probabilities. That is, we may know $P(A|B)$, but we want to make statements regarding $P(A)$ or $P(A,B)$.

To do so, we can simply rearrange the defining relationship of conditional probability, to solve

for the quantities of interest. Because of the importance of this procedure, we actually give this mostly straightforward rearrangement a special name.

Multiplication Rule

The multiplication rule states that, for two events A and B where $P(B) > 0$,

$$P(A \cap B) = P(A|B)P(B).$$

Note that, by multiplying both sides of the definition of $P(A|B)$ by $P(B)$ gives the result. In words it states that we can solve for the joint probability of A and B by multiplying the conditional probability of A given B, by the marginal probability of B. This is symmetric in A and B so that

$$P(A \cap B) = P(B|A)P(A).$$

This is useful as sometimes it is easier to determine B given A.

Example 4.4 (Package Delivery Times)**.** To thank Sadie for the help in seeing the silver lining with the urn mishap, Charles decides to order a small gift online, sending it direct to Sadie. Unfortunately, the website does not list which delivery company each package is sent out with. After some sleuthing, Charles determines that there are two different companies that it may have been sent with. Looking at online reviews it appears as though company A is late 75% of the time while company B is late 15% of the time.

If the store that Charles ordered from sends out 10% of packages with company A, and the rest with company B, which is more likely: that the package is late and was sent with A, or that the package was late and sent with B?

Solution

We will take L to correspond to the events where the package is late, A to the events where the package was sent with A and B to be the events where the package was sent with B. We wish to compare $P(L, A)$ and $P(L, B)$. We know that $P(A) = 0.1$ and $P(B) = 0.9$, and we know that $P(L|A) = 0.75$ and $P(L|B) = 0.15$. As a result, we can use the multiplication rule to find the joint probabilities. First, $P(L, A) = P(L|A)P(A) = (0.75)(0.1) = 0.075$. Additionally, $P(L, B) = P(L|B)P(B) = (0.15)(0.9) = 0.135$. As a result it is more likely to have the package late and sent with B then the package late and sent with A.[1]

[1] Note, this is another result which seems to (on the surface) defy expectations. We will see this again later on in this chapter in a slightly different context. It seems strange that company A is more likely to be late, and yet, we are more likely to see a late package that is sent by company B then a late package that is sent by company A. The reason for this is that company B is used much more frequently than company A, which overcomes the added likelihood of company A being late.

4.3.2 Partitions and the Law of Total Probability

While the multiplication rule gives us the capacity to solve for joint probabilities, often we wish to make statements regarding marginal probabilities. Fortunately, we can extend this process outlined in the multiplication rule to solve directly for marginal probabilities as well. To do so, we first introduce the concept of a partition.

Definition 4.4 (Partition)**.** A partition is a collection of sets which divide up the sample space such that all of the sets are disjoint from one another, and the sample space is given by the union of all of the sets. That is, $A_1, \ldots, A_n$ is a partition of $\mathcal{S}$ if:

1. $A_i \cap A_j = \emptyset$ for all $i \neq j$, and
2. $$\bigcup_{i=1}^{n} A_i = \mathcal{S}.$$

For instance, if the sample space were all the positive integers, we could partition this space into all the even numbers as one set and all the odd numbers as a second. We could also partition this into the set of numbers which are less than 10, the set of numbers that are greater than 10, and then 10. In both examples we have sets whose union forms the full sample space with no overlap. Note that we could not partition the set into multiples of 2 and multiples of 3, since (i) not all values are contained between these two sets,[2] and (ii) there is overlap between these two sets.[3]

Example 4.5 (Partitions of the Coin Game)**.** Charles and Sadie are thinking back with fondness to their original coin flipping game, where they would toss a coin three times in a row. If two or more heads showed up, Charles would pay. Otherwise, Sadie would.

To help them reminisce, Write down three different partitions of the sample space, each with a different number of partitioning sets. Describe your partitions using words.

Solution

There are many possible partitions to write down. The following are examples.

1. We can partition the space into games where Charles pays and games where Sadie pays. This gives

$$B_1 = \{(\mathrm{H},\mathrm{H},\mathrm{H}), (\mathrm{H},\mathrm{H},\mathrm{T}), (\mathrm{H},\mathrm{T},\mathrm{H}), (\mathrm{T},\mathrm{H},\mathrm{H})\}$$
$$B_2 = \{(\mathrm{T},\mathrm{T},\mathrm{T}), (\mathrm{T},\mathrm{T},\mathrm{H}), (\mathrm{T},\mathrm{H},\mathrm{T}), (\mathrm{H},\mathrm{T},\mathrm{T})\}.$$

[2] For instance, 5 is neither a multiple of 2 nor of 3.
[3] For instance, 6 is a multiple of both 2 and 3.

2. We can partition the space into those with 0, 1, 2, or 3 heads. This gives

$$B_1 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H})\}$$
$$B_2 = \{(\mathrm{T}, \mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{T})\}$$
$$B_3 = \{(\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{H})\}$$
$$B_4 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H})\}.$$

3. We can partition the space into the number of times the sequence switches between heads and tails. There will be either 0 switches, 1 switch, or 2 switches. This gives

$$B_1 = \{(\mathrm{H}, \mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{H}, \mathrm{H})\}$$
$$B_2 = \{(\mathrm{T}, \mathrm{T}, \mathrm{H}), (\mathrm{H}, \mathrm{T}, \mathrm{T}), (\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H}, \mathrm{H})\}$$
$$B_3 = \{(\mathrm{T}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H})\}.$$

Partitions allow us to move from discussions regarding the joint probability of events to the marginal probability of an event. Suppose that we have a partition given by $B_1, B_2, ...$ This means that our full sample space can be cut up into these various non-overlapping sets, and every single outcome belongs to exactly one of them. Now, suppose we are interested in some other event A. We can ask: how can A occur, in terms of the events $B_1, B_2, ...$? Since every single event in the sample space belongs to exactly one of our partitioning sets, then it **must** be the case that every single event in A belongs to exactly one of our partitioning sets. This means that if we consider $A \cap B_j$, for all j, then every single event in A must belong to exactly one of these. In other words, it must be the case that

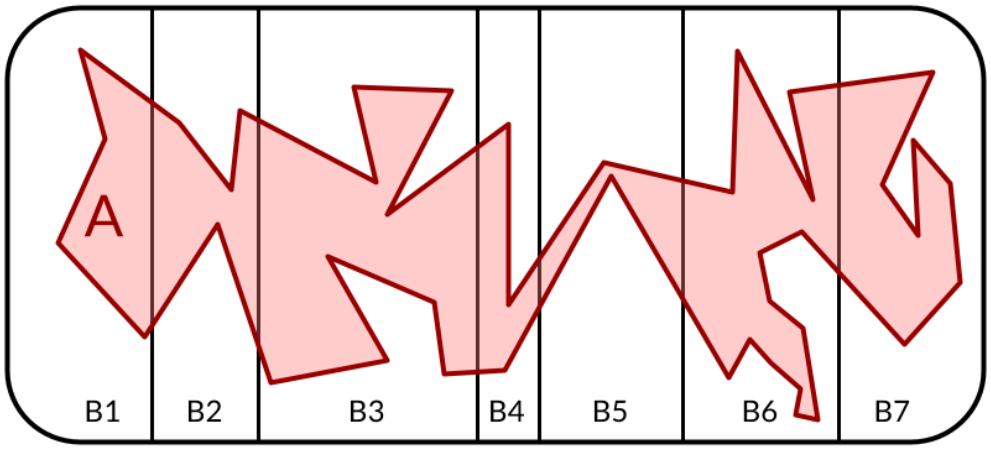
$$A = \bigcup_j A \cap B_j,$$

for any partition $B_1, B_2, ...$ Moreover, every single $A \cap B_j$ is disjoint from every other $A \cap B_\ell$, whenever $\ell \neq j$. This means that we can use the axiom of additivity to give

$$P(A) = P\left(\bigcup_j A \cap B_j\right) = \sum_j P(A \cap B_j).$$

In other words: the marginal probability of A can be found by summing over **all** joint probabilities between A and sets that form a partition. This argument gives the law of total probability.

Figure 4.1: This graphic shows the argument that summing over the joint probabilities between an event and a partition gives the full marginal probability. Note that $B_1, \ldots, B_7$ forms a partition of the space where every possible outcome is contained in exactly one of these sets. Then, if we take an arbitrary event A, we can divide A into the components that intersect with each partitioning set, namely $A \cap B_1$, $A \cap B_2$, and so forth.



The Law of Total Probability

Given a partition, $B_1, B_2, ...$, and an event A, the law of total probability states that

$$P(A) = \sum_i P(A, B_i) = \sum_i P(A|B_i)P(B_i).$$

Intuitively, since the whole sample space is divided into the different B_is, this rule breaks down the calculation of A happening into manageable chunks. Each term in the summation is "the probability that A happens, given B_i happening" weighted by how likely it is that B_i happens. Then by summing over all possible B_i, we know that we must be capturing all possible ways that A can occur since all parts of the sample space are contained in exactly one of the sets of our partition. The law of total probability is an indispensible tool for computing probabilities in practice.

Example 4.6 (Sadie's Possibly Late Package)**.** Sadie has still not received the package that Charles had ordered. While it is not late yet, Charles decides to figure out the probability

that the package will end up being late. Recall that company A is late 75% of the time while company B is late 15% of the time, and the store that Charles ordered from sends out 10% of packages with company A, and the rest with company B, which is more likely.

What is the probability that the package is late?

Solution

Note that A, B forms a partition of the sample space as every package is either sent with A or with B, and no package can be sent with both. Further, if L represents the event that the package is late, then we know that $P(L|A) = 0.75$ and $P(L|B) = 0.15$. Since $P(A) = 0.1$ and $P(B) = 0.9$, an application of the law of total probability gives

$$P(L) = P(L|A)P(A) + P(L|B)P(B) = (0.75)(0.1) + (0.15)(0.9) = 0.21.$$

As a result, knowing nothing else, the package has a probability of 0.21 of being late.

Example 4.7 (Charles' Many Urns)**.** While shopping at some garage sales one Sunday morning, Charles and Sadie stumble across a wonderful find! They see three urns which are **exactly** identical to the one that Charles had already purchased to store the different balls which turned out to not be balls at all. Realizing the opportunity they splurge and purchase them all, and then divide the various objects between the four urns, placing all the spheres in one container, all the cubes in another, all the pyramids in a third, and all the cones in a fourth.

Once done, they use a different selection mechanism. First, they pick an urn at random. Next, they random grab one of the items from within it. The distribution of colours and shapes is included in the following table.

Colour	**Sphere (Urn 1)**	**Cube (Urn 2)**	**Pyramid (Urn 3)**	**Cone (Urn 4)**
Red	2	3	2	0
Blue	1	0	0	6
Green	2	2	1	2
Yellow	0	4	2	1
Black	3	1	1	2

a. What is the probability that they select any of the five colours under this sampling scheme?
b. How does this change if the probability that each urn is selected is proportional to the number of items in it? (Thus, urn 1 is selected with probability 8/35, and so forth).

Solution

a. Let R, B, G, Y, and Bk be the events that a red, blue, green, yellow, or black ball are selected. Further, let U_1, U_2, U_3, and U_4 be the events that the first, second,

third, or fourth urn are selected. Then note that, according to the law of total probability,

$$P(R) = P(R, U_1) + P(R, U_2) + P(R, U_3) + P(R, U_4) = \sum_{j=1}^{4} P(R|U_j)P(U_j).$$

An equivalent argument holds for each of the other colours. Now,

$$P(R|U_j) = \frac{N_{R\cap U_j}}{N_{U_j}},$$

where $N_{R\cap U_j}$ is the number of red objects in urn j and N_{U_j} is the total number in urn j. Plugging this in and simplifying we get

$$P(R) = \sum_{j=1}^{4} P(R|U_j)\left(\frac{1}{4}\right) = \frac{1}{4}\left\{\frac{2}{8} + \frac{3}{10} + \frac{2}{6}\right\} = \frac{53}{240}.$$

We can apply analogous arguments to the other colours giving

$$\begin{aligned}
P(B) &= \frac{1}{4}\left\{\frac{1}{8} + \frac{6}{11}\right\} = \frac{59}{352} \\
P(G) &= \frac{1}{4}\left\{\frac{2}{8} + \frac{2}{10} + \frac{1}{6} + \frac{2}{11}\right\} = \frac{527}{2640} \\
P(Y) &= \frac{1}{4}\left\{\frac{4}{10} + \frac{2}{6} + \frac{1}{11}\right\} = \frac{34}{165} \\
P(Bk) &= \frac{1}{4}\left\{\frac{3}{8} + \frac{1}{10} + \frac{1}{6} + \frac{2}{11}\right\} = \frac{1087}{5280}.
\end{aligned}$$

Note that these probabilities sum to 1. In decimal these simplify to approximately 0.221, 0.168, 0.2, 0.206, 0.206.

b. Using the same setup as before, we have

$$P(R) = P(R, U_1) + P(R, U_2) + P(R, U_3) + P(R, U_4) = \sum_{j=1}^{4} P(R|U_j)P(U_j).$$

Now $P(U_1) = 8/35$, $P(U_2) = 10/35$, $P(U_3) = 6/35$, and $P(U_4) = 11/35$. Moreover, the conditional probabilities themselves do not change, and so instead we have

$$P(R) = \left\{\frac{2}{8}\cdot\frac{8}{35} + \frac{3}{10}\cdot\frac{10}{35} + \frac{2}{6}\frac{6}{35} + \frac{0}{11}\cdot\frac{11}{35}\right\} = \frac{N_R}{35}.$$

Note that when multiplying by the marginal probability of the urn, the denominator will always cancel. As a result, we end up with the total number of reds over 35, which leads to $P(R) = \frac{1}{5}$. The same will be true for the other colours, and as a result, if we choose the urn based on a weighted selection, this will result in equal probability once more.

4.3.3 Bayes' Theorem

We have seen the direct computation of marginal probabilities (while using an equally likely outcome model), the computation of conditional probabilities, the use of the multiplication rule for joint probabilities, and the use of the law of total probability to indirectly calculate marginal probabilities through conditioning arguments. Throughout these discussions we have been primarily concerned with keeping events A and B arbitrary. Everything that we have indicated for $P(A)$ holds for $P(B)$, as does $P(A|B)$ and $P(B|A)$. In reality, it will often be the case that conditioning on one of the events will be natural, while conditioning on the other will be more tricky. In these events, it can be useful to be able to transform statements regarding $P(A|B)$ into statements regarding $P(B|A)$, and vice versa.

Note that because the definitions are symmetric,

$$P(A|B)P(B) = P(A, B) = P(B|A)P(A).$$

This is an application of the multiplication rule in two different orientations. If we divide both sides of the equality by $P(B)$, assuming that it is not 0, then we get

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Now, if we form a partition, say $A, A_2, A_3, ...$, then we can rewrite $P(B)$ using the law of total probability as

$$P(B) = P(B|A)P(A) + P(B|A_2)P(A_2) + \cdots = P(B|A)P(A) + \sum_{i=2} P(B|A_i)P(A_i),$$

which can replace $P(B)$. Taken together this gives a result known as Bayes' Theorem.[4]

Bayes' Theorem

Suppose that there are two events, A and B, with $P(B) > 0$. Moreover, suppose that A taken with $A_2, A_3, ...$ forms a partition. Then Bayes' Theorem states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + \sum_{i=2} P(B|A_i)P(A_i)}.$$

Bayes' Theorem allows us to convert statements regarding $P(B|A)$ into statements regarding $P(A|B)$. Note that, as we derived above, Bayes' Theorem is an application of the multiplication

[4] Bayes' Theorem is named in the same way that the Bayesian interpretation of probability is, and that Bayesian statistics is more broadly. The connections are more than merely surface: Bayes' Theorem can be viewed as the primary technique with which we can update subjective beliefs about the world. Importantly, however, even those who use a Frequentist view of statistics accept the math of Bayes' Theorem, and use it frequently.

rule and an application of the law of total probability.[5] Sometimes we may have $P(B)$ directly, rendering the law of total probability in the denominator unnecessary.

Often, the natural partition to select when we do need the law of total probability is to take A and A^C. Note that any set with its complement forms a partition, since by definition they occupy the entire space and are non-overlapping. When this is done we get the slightly more compact relationship of

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}.$$

Bayes' Theorem differs from our previous relationships as it allows us to translate one set of conditional probabilities into another. Every other relationship we have looked at has moved between types of probabilities, whereas Bayes' Theorem deals directly with conditional relationships.

The most commonly cited example application is in medical testing. Suppose that we know the performance characteristics of a particular medical test: it is 99% accurate for positive cases, and 95% accurate for negative cases. That is, with probability 0.99 it correctly returns positive when an individual is infected, and with probability 0.95 it returns negative when an individual is not infected.These are both statements of conditional probability. If we take A to be the event that the test returns positive, and B to be the event that the patient is infected,[6] then we are saying that $P(A|B) = 0.99$ and $P(A^C|B^C) = 0.95$ which means that $P(A|B^C) = 0.05$. Suppose that we know that, across the entire population, one in a thousand individuals is likely to be infected. This means that $P(B) = 0.001$.

Now if a random individual goes into a doctor's office and tests positive for the disease, how likely are they to actually be infected? In this case we want to know the probability of them being infected given that they have tested positive. In notation, this is $P(B|A)$. We do not know this quantity directly, but given an application of Bayes' Theorem, we can find it. Using the natural partition of B and B^C, we get

$$P(B|A) = \frac{P(A|B)P(A)}{P(A|B)P(B) + P(A|B^C)P(B^C)} = \frac{(0.99)(0.001)}{(0.99)(0.001) + (0.05)(0.999)} \approx 0.019.$$

[5] In fact, I would go as far as to suggest that learning Bayes' Theorem by itself is less important than fully grasping the definition of conditional probability alongside the multiplication rule and the law of total probability. I myself do not remember Bayes' Theorem directly, but can write it down directly from these definitions without any thought.

[6] It is worth drawing attention to the language that we have started to use at this point in the notes regarding "events". In this case our sample space would actually be formed using pairs of information. In particular, we might have

$$\mathcal{S} = \{(\text{Pos. Test}, \text{Illness}), (\text{Neg. Test}, \text{Illness}), (\text{Pos. Test}, \text{No Illness}), (\text{Neg. Test}, \text{No Illness})\}.$$

Then the event A here is actually $A = \{(\text{Pos. Test}, \text{Illness}), (\text{Pos. Test}, \text{No Illness})\}$ and B is $\{(\text{Pos. Test}, \text{Illness}), (\text{Neg. Test}, \text{Illness})\}$. It is far more clunky to make explicit these events, and so we move towards using more natural language. Until you feel confident that you can identify the specific outcomes associated with the events of interest, it is worth writing these out in full.

That is, despite the fact that this test is exceptionally effective at detecting this disease, a positive test still means that an individual has a probability of only 0.019 of actually having the illness.[7]

Example 4.8 (Sadie's Late Package)**.** Sadie eventually received the package that Charles had sent, but it arrived very late. Sadie was not home when the package was delivered and there no obvious markings on the box to indicate which of the two delivery companies had sent it.

Given this information, and knowing that company A is late 75% of the time while company B is late 15% of the time, and the store that Charles ordered from sends out 10% of packages with company A, and the rest with company B, what is the probability that the package was delivered by each of the two companies?

Solution

We want to determine $P(A|L)$ and $P(B|L)$. We know that $P(L|A) = 0.75$ and $P(L|B) = 0.15$. Moreover, we know that $P(A) = 0.1$ and $P(B) = 0.9$. Applying Bayes' Theorem directly we get

$$P(A|L) = \frac{P(L|A)P(A)}{P(L|A)P(A) + P(L|B)P(B)} = \frac{(0.75)(0.10)}{(0.75)(0.10) + (.10)(0.90)} = \frac{5}{11}.$$

Similarly, we get

$$P(B|L) = \frac{P(B|A)P(B)}{P(L|A)P(A) + P(L|B)P(B)} = \frac{(0.15)(0.90)}{(0.75)(0.10) + (.10)(0.90)} = \frac{6}{11}.$$

Note, we could have also used the fact that $P(A|L) + P(B|L) = 1$ to determine this. As a result, it is more likely that B delivered the package than A.[8]

Bayes' Theorem highlights a key lesson when considering conditional probabilities, and it's a common mistake to make which should be avoided at all costs. Namely, we cannot interchange $P(A|B)$ with $P(B|A)$. These probabilities are not necessarily highly correlated with

[7]This counter intuitive fact was an intensely frustrating reality for statisticians everywhere during the height of the COVID-19 pandemic, when politicians and the population at large turned away from testing owing to its perceived "ineffectiveness". The quantity of interest for knowing how good a test is is $P(A|B)$ and $P(A^C|B^C)$. However, if a disease is sufficiently rare, with $P(B)$ sufficiently small, then no matter how effective the tests are you will likely have $P(B|A)$ to be low. Note that $P(B|A) >> P(B)$ in the example, and this will also be true in general. A single test cannot say with certainty, however, they are an incredibly effective tool at reducing our uncertainty.

[8]Note that this is another example of a counterintuitive result. Here, A is far more likely to be late, but it is more likely that B delivered a late package to us than A because of the **base rates**. That is, $P(B)$ is much higher to begin than $P(A)$, and so that is hard to overcome. It is worth noting, however, that originally $P(B)$ is 9 times more likely than $P(A)$, and after knowing that package is late it is $\frac{6}{5} = 1.2$ times more likely. This major reduction in the relative likelihood is owed to how much more likely A is to deliver late packages than B.

one another, and it is important to distinguish clearly which is the probability of interest.[9] The stakes of these types of confusion can be quite high, and it is tremendously important to ensure that you are conditioning on the correct events. Fortunately, Bayes' Theorem allows us to translate between events for conditioning, giving a mechanism from translating between the two.[10]

4.4 Independence

We have seen that, in most cases, conditioning on an event changes the probability of that event. For instance, if we want to know the probability it is raining, if we condition on knowing that it is a day full of gray skies, the conditional probability is likely higher than the marginal probability. By considering how the marginal and conditional probabilities differ, we are in effect indicating a dependence of the events on one another. In terms of probability, this dependence is captured by an influence on the degree of uncertainty present depending on what we know.

It is totally possible that two events do not influence one another at all. The weather outside today is likely not influenced by your favourite sports team's performance last night.[11] In this case, we would have $P(A|B) = P(A)$.

We saw an example of this previously when we wanted to know the probability of a randomly selected card being a heart (A) given that it was an ace (B). We found that this was $\frac{1}{4}$, exactly the same as the probability if we did not know that it was an ace. Thus here we have $P(A|B) = P(A)$. We could have also said that $P(B|A) = P(B) = \frac{1}{13}$. The symmetry of these events makes it somewhat more convenient to express this relationship differently.

Instead of writing $P(A|B) = P(A)$ and $P(B|A) = P(B)$, we can multiply the first relationship by $P(B)$ on both sides, or the second by $P(A)$ on both sides. The multiplication rule gives

[9] Mixing up $P(B|A)$ and $P(A|B)$ is often called *confusion of the inverse*, and it can lead to very faulty conclusions when ignored. In the medical testing example above, it is important to not confuse "the probability that the test returns positive, assuming you have the illness" with "the probability that you have the illness, given that the test returns positive." This type of faulty logic has long been used to justify discriminatory behaviour in medicine, the law, and so on. A thorough understanding of statistics and probability helps to ensure that these types of errors are not made, and gives you the tools to push-back on the spots where people are making these arguments incorrectly, particularly when the stakes are high or harm is being done.

[10] When discussing Bayes' Theorem we introduced two frequently occurring sources of error when reasoning about probabilities, and showed how to remedy them. *Confusion of the inverse* occurs when you mix up $P(A|B)$ with $P(B|A)$, and can lead to disastrous consequences. The *base rate fallacy* occurs when you fail to take into account the rareness of the marginal events and instead only consider the conditionals (such as seeing that delivery company A was more likely to be late than B, without considering that B was more likely to be used than A). These are both examples of the challenges at the heart of probability and statistics: namely, the subjects are fairly unintuitive once moving beyond the basics. As a result, we need to rely on building up our intuitions over time by making use of the formal rules that we are able to derive.

[11] Though, perhaps the world will freeze over if *(insert-your-least-favourite-team-here)* wins the *(insert-the-name-of-the-championship-for-the-league-you-care-about-here)*?

$P(A|B)P(B) = P(A, B)$, and so the first relationship becomes $P(A, B) = P(A)P(B)$. The second follows exactly the same. Any two events that satisfy this relationship are said to be independent.

Definition 4.5 (Independence)**.** Any two events, A and B, which satisfy $P(A, B) = P(A)P(B)$ are said to be independent. If A and B are independent we write $A \perp B$, and read "A is independent of B". Any events which are not independent are said to be dependent, and we write $A \not\perp B$.

Note that independence is always a symmetric property: if A is independent of B, then B is independent of A. To check whether two events are independent, we check whether their joint probability is equal to the product of their marginal probabilities.

Properties of Independence

Note that if $A \perp B$, then $A^C \perp B$, $A^C \perp B^C$, and vice versa. To see this note

$$P(A^C, B) + P(A, B) = P(B),$$

by the Law of Total Probability. Then, by independence of A and B this gives

$$\begin{aligned} P(A^C, B) + P(A)P(B) &= P(B) \\ \implies P(A^C, B) &= P(B) - P(A)P(B) \\ &= P(B)(1 - P(A)) \\ &= P(B)P(A^C), \end{aligned}$$

as is required. The other combinations follow in the same manner.

If $P(A) \neq 0$, then we can divide both sides by $P(A)$ to gives $P(B|A) = P(B)$. Similarly, if $P(B) \neq 0$, then we can divide both sides by $P(B)$ to give $P(A|B) = P(A)$. This expression in terms of conditional probabilities is the more intuitive expression of independence. It directly captures the idea that "knowing B does not change our belief about A". However, we must be careful. This conditional argument is only valid when the event that is being conditioned on is not probability 0, where the defining relationship, $P(A, B) = P(A)P(B)$, will hold for all events. Recall that, in general, $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. It is only when assuming independence that this simplifies further.

Example 4.9 (Independence and Board Games)**.** Charles and Sadie are invited over to a friends house, Garth, to play some board games. Both of them are quite excited by this prospect as board games feel like the next logical step for the two of them, being as into probability and games as they are! Garth has a large board game collection, and begins to explain the games to Charles and Sadie across a variety of axes:

- Whether the games are competitive or cooperative.
- How many players the games play best at ($\{1, 2, 3, 4+\}$).
- How "heavy" the games tend to be (friendly for everyone, moderately involved, or very heavy).

While Garth continues on about several other topics including the themes, the mechanics, or the rating on BoardGameGeek, Charles drifts off wondering whether the traits listed are independent in Garth's collection. Suppose that Garth has 100 games, of which:

- 25 are cooperative.
- 5 are best played at one player, 20 at two, and 55 at three.
- 10 are friendly for everyone and 45 are moderately involved.

a. If 5 games are best played at two players and are cooperative, are these events independent?
b. If 20 games are heavy and best played with 4+ players, are these events independent?
c. If 0 games are both moderately involved and cooperative, are these events independent?
d. Charles figures that competitive games and two player games are independent. How many competitive two player games are there?
e. Is it possible that competitive games are independent of heavy games?

Solution

a. We are suggesting that $P(A, B) = 0.05$, where A is "two player" and B is "cooperative". We know that $P(B) = 0.25$ and $P(A) = 0.20$. Calculating $P(A)P(B) = (0.20)(0.25) = 0.05 = P(A, B)$, and so these traits **are** independent.

b. We have $P(A, B) = 0.2$ where A is "heavy" and B is "4+ players. We know that $P(A) = 0.45$ and $P(B) = 0.20$, and so $P(A)P(B) = (0.45)(0.20) = 0.09 \neq P(A, B)$. As a result, these traits are **not** independent.

c. We know that $P(A) > 0$ and $P(B) > 0$ for A being moderately involved and B being cooperative. As a result, $P(A)P(B) > 0$, and so if $P(A, B) = 0$, they must **not** be independent.

d. Taking A to be "competitive" we have $P(A) = 0.75$. Taking B to be "two player" we have $P(B) = 0.20$. As a result, if $A \perp B$ then $P(A, B) = P(A)P(B) = (0.75)(0.20) = 0.15$. Thus, there must be 15 competitive, two player games.

e. Taking A to be "competitive" we have $P(A) = 0.75$. Taking B to be "heavy" weh ave $P(B) = 0.45$. Thus, if they were independent, we would have $P(A, B) = P(A)P(B) = (0.75)(0.45) = 0.3375$. This would require 33.75 games to be heavy, competitive games and so we must conclude that they are **not** independent.

4.4.0.1 Mutually Exclusive Events

Importantly, if $A \perp B$, then $A \cap B \neq \emptyset$ unless either $A = \emptyset$, $B = \emptyset$, or both. To see this recall that $P(\emptyset) = 0$, and so if $A \cap B = \emptyset$ then $P(A \cap B) = P(A)P(B) = P(\emptyset) = 0$. This only holds if either $P(A) = 0$ or $P(B) = 0$. This may seem to be a rather technical point, however, it is the source of much confusion regarding independence. In particular, it is common to mistake independent events for mutually exclusive events.

Definition 4.6 (Mutually Exclusive Events)**.** Two events, A and B are said to be mutually exclusive if they are disjoint. In particular, if $P(A, B) = 0$ then A and B are mutually exclusive. If A and B are mutually exclusive events, with $P(A) > 0$ and $P(B) > 0$, then $A \not\perp B$.

Whenever only one event from a set of events can happen, we refer to the events as being mutually exclusive. If one happens, we know that the others did not. Mutually exclusive events are always dependent since knowing that A occurs dramatically shifts our belief about B,C, and D.[12]

The primary concern with mutually exclusive and independent events is a linguistic one. We often use words like independent to mean unrelated, and in a sense, mutually exclusive events are unrelated in that one has nothing to do with another. However, in statistics and probability, when we discuss independence, it is not an independence of the events themselves but rather an independence relating to our beliefs regarding the uncertainty associated with the events. In this sense, mutually exclusive events are very informative regarding the uncertainty associated with them.

Example 4.10 (Charles and Sadie Cooking Dinner)**.** Suppose that Charles and Sadie always eat dinner together. It will either be the case that Charles cooks at home, that Sadie cooks at home, that the two of them order in, or that they go out to eat. If we take these to be four events, A, B, C, and D, then are any of these events independent? Mutually exclusive? Explain.

Solution

Assuming, as it seems reasonable given the information in the question, that only one of the possible tasks occurs for dinner in a night, then we *know* that

$$P(A, B) = P(A, C) = P(A, D) = P(B, C) = P(B, D) = P(C, D) = 0.$$

As a result, all of the events described are mutually exclusive, and correspondingly, are *not* independent.

[12] Namely, we know that all of the others are then impossible.

Table 4.6: Frequency and corresponding proportions of enrolment by faculty in a university.

Faculty	Enrolment	Proportion
Arts	1266	0.2182382
Computer Science	749	0.1291157
Education	786	0.1354939
Engineering	1315	0.2266851
Law	266	0.0458542
Nursing	543	0.0936046
Science	876	0.1510084
Total	**5801**	**1.0000000**

4.5 Contingency Tables

Through to this point we have discussed probabilities in the abstract, either through an enumeration of equally likely outcomes, or else by directly specifying the likelihood of various events. While these are useful in many regards, we are often looking for more concise manners of summarizing information of interest. One tool for accomplishing this is a contingency table.

Definition 4.7 (Contingency Table)**.** A contingency table is a tabular summary of information which summarizes the joint probabilities of two or more variables. Typically a contingency table will take one factor for the columns and a secondary factor for the rows, where each cell then represents the frequency with which observations occur in the two corresponding categories simultaneously.

To begin, you could imagine constructing a frequency table relaying the frequency with which undergraduate students are enrolled in various faculties at a particular university. This tells you, of the whole population of students at the university, what is the faculty breakdown. By dividing the number in each faculty by the total number of students, you convert the frequencies to proportions, and these proportions can be viewed as probabilities. (See Table 4.6.[13]) The interpretation of proportions as probabilities implies a very specific statistical experiment. In particular, the proportion represents the probability that an individual selected at random from the entire population has the given trait. This is frequently a probability of interest, which makes these summary tables a useful tool.[14]

When a single trait is displayed we refer to these tabular summaries as frequency tables or frequency distributions. A contingency table instead plots two or more traits on the same

[13]Note, the values in this table are *roughly* inspired by a subset of the Fall 2022 University of New Brunswick Enrolment Numbers.

[14]This provides further emphasis for the utility of urn models. If you imagine an urn that has balls with two or more traits (say like those in Example 4.3), then randomly selecting a single ball from this population has an analogous probability distribution.

Table 4.7: Frequency of enrolment by faculty and year of study in a university.

	Year 1	Year 2	Year 3	Year 4	Year 5+	
Arts	222	276	273	225	270	**1266**
Comp. Sci.	164	87	164	90	244	**749**
Education	100	136	184	128	238	**786**
Engineering	189	290	354	298	184	**1315**
Law	54	50	58	37	67	**266**
Nursing	55	90	154	112	132	**543**
Science	242	114	206	183	131	**876**
	1026	**1043**	**1393**	**1073**	**1266**	**5801**

table, with each cell representing the frequency of both traits occurring simultaneously in the population. Extending the university example, we may further include the student's current year of study to see the breakdown of both faculty and year of study, in one table.

By including two (or more) factors in the table we are able to capture not only the marginal probabilities for the population, but also the joint probabilities for the population, and in turn, the conditional probabilities for the population. Being able to concisely summarize all of these concepts regarding traits in a population of interest renders contingency tables immensely useful in the study of uncertainty broadly.

Consider the two-way contingency table, Table 4.7. Each cell consists of frequency with which a combination of the two traits occurs in the population.[15] If we take events corresponding to each of the levels of the two variables of interest, then these central cells represent the frequency of joint events. That is, each interior cell gives the total number of observations with a set level for variable one[16] AND a set level for variable two[17]. For instance, there are 50 individuals who are in Law and studying in Year 2.

Each row is then summed, with the total number following into the corresponding row recorded in the right hand margin. each column is summed, with the total number corresponding to the given column recorded in the bottom margin. For instance there are a total of 786 students in Education and a total of 1393 in Year 3. Then the margin totals are summed and the total is recorded in the lower right margin space (in this case, 5801). Whether the rows or columns are summed, they should sum to the same total, which is the total of the population under consideration. This is the same as simply adding all of the observed interior frequencies. To turn a frequency into a probability, you need only divide the correct frequency by the correct total.

For the standard joint probabilities, you take the interior cell count and divide by the population total. Here we are saying that some fixed number, m, of the N total individuals have

[15]That is, the number of students who are enrolled in that faculty, in that year of study.
[16]Faculty.
[17]Year of study.

Table 4.8: Proportions of enrolment by faculty and year of study in a university.

	Year 1	Year 2	Year 3	Year 4	Year 5+	
Arts	0.0382693	0.0475780	0.0470609	0.0387864	0.0465437	**0.2182382**
Comp. Sci.	0.0282710	0.0149974	0.0282710	0.0155146	0.0420617	**0.1291157**
Education	0.0172384	0.0234442	0.0317187	0.0220652	0.0410274	**0.1354939**
Engineering	0.0325806	0.0499914	0.0610240	0.0513705	0.0317187	**0.2266851**
Law	0.0093087	0.0086192	0.0099983	0.0063782	0.0115497	**0.0458542**
Nursing	0.0094811	0.0155146	0.0265471	0.0193070	0.0227547	**0.0936046**
Science	0.0417169	0.0196518	0.0355111	0.0315463	0.0225823	**0.1510084**
	0.1768661	**0.1797966**	**0.2401310**	**0.1849681**	**0.2182382**	**1.0000000**

both traits under consideration. For instance, the joint probability that a student is in Law and studying in Year 2 is 0.0086192.[18] If instead you wish to find a marginal probability, you have to consider the value in the corresponding margin: this is the total number of individuals with the given trait, ignoring the level of the other variable. These marginal values are also divided by the total population size. For instance, there is a 0.1354939 probability of observing a student in Education and a probability 0.240131 of observing a student in Year 3.

Outside of joint and marginal probabilities, we can also find conditional probabilities. To do so, we restrict our focus to either only one row, or one column. Then, we can take the joint cell and divide by the value in the margin, which gives the conditional probability of interest. Note, this works with *either* the contingency table directly **or** with the propensity table. The reasoning is that the propensity table divided by the same totals in the numerator and the denominator. Suppose we take some cell, $A \cap B$, which has $N_{A,B}$ in the contingency table. Then,

$$P(A \cap B) = \frac{N_{A,B}}{N} \qquad \text{and} \qquad P(B) = \frac{N_B}{N}.$$

If we consider then forming $P(A|B)$ we get

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{N_{A \cap B}/\cancel{N}}{N_B/\cancel{N}} = \frac{N_{A,B}}{N_B}.$$

Thus, given that we know a student is in Education, then the probability that they are in Year 3 is approximately 0.2341.

Note that these procedures are exactly in line with what we had seen before. The conditional probability is defined as the joint probability divided by the marginal probability. The process of computing the marginal can be seen as an application of the law of total probability. As

[18] It is worth reemphasizing what this probability actually means. If we were to randomly sample, with equal probability, individuals from this population then the probability that an individual selected has both the faculty and year specified is given by the joint probability. That is to say, if we did this over and over and over again (with replacement) in the long-term, these probabilities represent proportion of time those combinations would be observed.

a result, contingency tables can be a useful, tangible tool for investigating the techniques we have been discussing: they are not a substitute for direct manipulation of the mathematical objects, but they can present insight into the underlying processes where it may be hard to derive that insight otherwise.

The Law of Total Probability: Contingency Table's Version

Recall that the law of total probability states that, if $B_1, \dots, B_n$ forms a partition of the sample space, then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) = \sum_{i=1}^{n} P(A, B_i)$$

. In a contingency table it is easier to see how either of the two factors at play (either those in the rows or the columns) forms a partition of the space. Every observation has to fall in exactly one row and exactly one column. Thus, if we want to know the marginal frequency of a single trait (say represented by some row of the table), then one way to find this total is to sum up every observation in each of the corresponding columns. This is *precisely* the same process as the law of total probability. Note that, denoting each column as B_i, and supposing there are k total columns, then the total number in a row is taken to be

$$N_A = \sum_{i=1}^{k} N_{A,B_i},$$

simply as the summation of the corresponding row. By definition, $P(A) = \frac{N_A}{N}$, and so dividing both sides by N gives

$$P(A) = \frac{N_A}{N} = \frac{1}{N}\sum_{i=1}^{k} N_{A,B_i} = \sum_{i=1}^{k} \frac{N_{A,B_i}}{N} = \sum_{i=1}^{n} P(A \cap B_i).$$

It is important to note that there is redundant information within a contingency table. For instance, the margins need not be listed explicitly, as they can be directly calculated from the interior points. Same goes for interior points, given the margins of the table (assuming *some* interior points are also presented). This can be useful for a compact representation of the information, and manipulating these tables – being able to find the required information in many places – should become familiar to you as you continue to work with them more and more.

Example 4.11 (The Evolving Contents of Charle's Urns)**.** After learning of contingency tables, Sadie points out to Charles that with the whole urn debacle they went through, the two of them actually ended up using a contingency table to summarize the information. How neat! In the interim, there has been some development in the contents of Charles's urns, and armed with the new knowledge of contingency tables, the following summary is produced.

	Cone	Cube	Pyramid	Sphere	
Black	2	A	1	3	**10**
Blue	1	2	1	1	**5**
Green	3	3	B	5	**15**
Red	2	C	3	2	**D**
Yellow	5	3	4	1	**13**
	13	**E**	**13**	**12**	**50**

Sadly, Charles does not have the best writing and Sadie cannot make out what values were written in for the cells marked A, B, C, D, and E.

a. What are the values for the missing values?
b. What is the probability that a black cube is drawn?
c. What is the probability that a red object is drawn?
d. What is the probability that a cube is drawn?
e. Given that the drawn object was a pyramid, what is the probability that it is green?
f. Given that the object is green, what is the probability that it is a pyramid.

Solution

a. We start by filling in the missing cells, in order.

   i. For A we can note that $2+A+1+3=10$, which gives that $A=4$. We could have also tried a column sum, however, this would involve 3 unknowns and so would not have given a numeric result.

   ii. For B either a row sum or column sum would produce the correct answer. We either take $3+3+B+5=15$ giving $B=4$ or $1+1+B+3+4=13$ giving $B=4$.

   iii. C is more challenging as we either have that $2+C+3+2=D$ or $4+2+3+C+3=E$[19] As a result, for this technique we will need to find either D or E first. There is an alternative technique to use, which would be to note that we have *all* the internal points specified at this point, and we know that these sum to 50. As a result, we could note that $C=50-10-5-15-13-2-3-2=0$ or that $C=50-13-13-12-4-2-3-3=0$.

   iv. We can find D either by subtracting from the total, $D=50-10-5-15-13=7$, or by adding along the row, $2+0+3+2=7$.[20] Note that with $D=7$, had we not found $C=0$ above, we could now take $2+C+3+2=7$ and find $C=0$.

   v. Like D, we have two options for working this out. Either $E=50-13-13-12=12$ or $E=4+2+3+0+3=12$.[21] Like with D, had we solved for E first, then we could find C via $4+2+3+C+3=12$.

b. Here we want $P(\text{Black}, \text{Cube})$. This is given by the frequency of black cubes divided

by 50. Thus,

$$P(\text{Black}, \text{Cube}) = \frac{A}{50} = \frac{4}{50} = 0.08.$$

c. Here we want $P(\text{Red})$. This is given by the marginal frequency of red objects divided by 50. Thus,

$$P(\text{Red}) = \frac{D}{50} = \frac{7}{50} = 0.14.$$

d. Here we want $P(\text{Cube})$. This is given by the marginal frequency of cubes divided by 50. Thus,

$$P(\text{Cube}) = \frac{E}{50} = \frac{12}{50} = 0.24.$$

e. Here we want $P(\text{Green}|\text{Pyramid})$. This is given by the frequency of green pyramids divided by the marginal frequency of pyramids. Thus

$$P(\text{Green}|\text{Pyramid}) = \frac{B}{13} = \frac{4}{13} \approx 0.308.$$

f. Here we want $P(\text{Pyramid}|\text{Green})$. This is given by the frequency of green pyramids divided by the marginal frequency of green objects.[22] Thus

$$P(\text{Pyramid}|\text{Green}) = \frac{B}{15} = \frac{4}{15} \approx 0.267.$$

It is also important to recognize that independence and mutually exclusive events can be codified via the table as well. Zeros on the interior points indicate events which are mutually exclusive: if we know that one of them occurred, we also know that the other one did not. For independence, it requires a degree of solving proportions. We can either check that the joint probability $(N_{A,B}/N)$ is equal to the product of the two marginal probabilities, $(N_A N_B/N^2)$, or else (assuming that the events are all non-zero), that the conditional probability $(N_{A,B}/N_A)$ equals the marginal probability N_B/N. Either way this is represented by $N \times N_{A,B} = N_A N_B$, and when this holds, we can conclude that the events are independent.

Example 4.12 (Independent or Mutually Exclusive Urn Shapes)**.** After helping Charles *neatly*

[22]Note here, $A = 4$ has been filled in.
[22]We have filled in $C = 0$ here.
[22]This uses $A = 4$ and $C = 0$.
[22]Note, we could also apply Bayes' Theorem to find the result here. We know that $P(\text{Green}|\text{Pyramid}) = 4/13$, that $P(\text{Green}) = 15/50$ and that $P(\text{Pyramid}) = 13/50$. Thus, Bayes' Theorem gives

$$P(\text{Pyramid}|\text{Green}) = \frac{(4/13)(13/50)}{15/50} = \frac{4}{15},$$

exactly as in the direct derivation.

fill in the contingency table, Sadie begins to wonder about whether there are any shape-colour combinations which are independent, or if any are mutually exclusive.

	Cone	Cube	Pyramid	Sphere	
Black	2	4	1	3	**10**
Blue	1	2	1	1	**5**
Green	3	3	4	5	**15**
Red	2	0	3	2	**7**
Yellow	5	3	4	1	**13**
	13	**12**	**13**	**12**	**50**

a. Are there any events represented in the contingency table which are mutually exclusive?
b. Are there any events represented in the contingency table which are independent?

Solution

a. Mutually exclusive events are codified via a 0 frequency. We can see then that Cube and Red are mutually exclusive. That is, if we know that an object is red, we also know that it is not a cube. And if we know that an object is a cube, then we know that it is not red.

b. Independence is more cumbersome to check. We require the product of marginal counts to be equal to the total times by the joint counts. As a result, the easiest way to check is to consider the product of each marginal value in the columns by each marginal value in the row, divided by 50. If this value corresponds to the value in that row/column pairing, then we know that those features are independent. Note that immediately we can rule out any products which are not divisible by 50, since we know that $N_{A,B}$ is an integer for all A, B and $N_{A,B} = N_A N_B/50$ if there is independence. To this end, we only need to check the product of each of the row totals with each of 12 and 13 (as those are the only two unique column totals). This gives, in order 120, 130, 60, 65, 180, 195, 84, 91, 156, and 169. None of these values are divisible by 50 and as a result we know that there is **no independence** codified in this table.

4.5.0.1 Contingency Tables and Data Frames in R

In R we can use the `table` and `prop.table` to calculate the (interior) of a contingency table. This will return a `table` type object in R, which can be thought of as a matrix of sorts. On it, we can use the functions `rowSums` and `colSums` to get the summation of the rows and columns, respectively. Typically the object we pass to `table` will be a **data frame**. A data frame is another R object type that we have not yet seen. The idea with a data frame is that we have multiple columns with different variables (of possibly different types) represented. Each row corresponds to a single observation, and then each column is read as a feature of those

observations. Data frames are essentially large spreadsheets (or data tables) which indicate the various observations that we have. In practice, data frames are the most commonly used object in an R analysis, and we will see them plenty going forward.

```
# As always, when randomization is to be used, we call
# the set.seed function to ensure that the analysis can be
# replicated.
set.seed(314)

# Begin by Defining the possible shapes and colours for
# the objects in Charles's urns
shapes <- c("Sphere", "Cube", "Pyramid", "Cone")
colours <- c("Red", "Blue", "Green", "Yellow", "Black")

# Use sample to draw 50 different shapes, and 50 different
# colours. We are drawing *with* replacement.
all_shapes <- sample(x = shapes, size = 50, replace = TRUE)
all_colours <- sample(x = colours, size = 50, replace = TRUE)

# Now we form the data frame. This will consist of two columns,
# one called "Colours" and one called "Shapes". The values to
# place here will correspond to the random colours and shapes we
# sampled above.
charles_data <- data.frame("Colours" = all_colours, "Shapes" = all_shapes)

# To get a sense of the data frame we can use the head function
# which will return the first few rows of our data frame.
head(charles_data)
##   Colours  Shapes
## 1   Green    Cube
## 2    Blue    Cone
## 3     Red    Cone
## 4     Red Pyramid
## 5  Yellow    Cone
## 6     Red Pyramid
```

With a data frame specified, we can then use the `table` function called on it to form a contingency table.

```
charles_c_table <- table(charles_data)

charles_c_table
```

```
##          Shapes
## Colours  Cone Cube Pyramid Sphere
##   Black     2    4       1      3
##   Blue      1    2       1      1
##   Green     3    3       4      5
##   Red       2    0       3      2
##   Yellow    5    3       4      1
```

Then, to get the totals, we can either use `rowSums` or `colSums` to return a vector with the corresponding row or column sums.

```
marginal_shapes <- colSums(charles_c_table)
marginal_colours <- rowSums(charles_c_table)

marginal_shapes
marginal_colours
##    Cone    Cube Pyramid  Sphere
##      13      12      13      12
##  Black   Blue  Green    Red Yellow
##     10      5     15      7     13
```

Finally, we can take our formed contingency table, and use `prop.table` on it, in order to return a table with the proportions (rather than the frequencies).

```
charles_p_table <- prop.table(charles_c_table)

charles_p_table
##          Shapes
## Colours  Cone Cube Pyramid Sphere
##   Black  0.04 0.08    0.02   0.06
##   Blue   0.02 0.04    0.02   0.02
##   Green  0.06 0.06    0.08   0.10
##   Red    0.04 0.00    0.06   0.04
##   Yellow 0.10 0.06    0.08   0.02
```

Exercises

Exercise 4.1. Four cards are dealt, in order, from a standard pack of 52 cards.

a. What is the probability that all four are spades?

b. What is the probability that two or fewer are spades?
c. What is the probability that all four are spades, given that the first two are spades?
d. What is the probability that spades and hearts alternate?

Exercise 4.2. Two cards are dealt from a deck of 52 cards. Find the probability that the second card dealt is a heart.

Exercise 4.3. A manufacturer of computer chips bins their manufactured chips based on the quality. Their production line ends up producing 1 high quality chip for every 2 medium quality chips and 2 low quality chips. That is, the proportions of high-to-medium-to-low remains $1 : 2 : 2$. For each quality level of chip, the probability that it is of unacceptable standard is 0, 0.1, and 0.2 respectively. Suppose a bin of chips is selected at random, two chips are tested, and found to be satisfactory.

a. What is the probability it is a high quality bin?
b. What is the probability it is a medium quality bin?
c. What is the probability it is a low quality bin?

Exercise 4.4. Consider the following argument from Lewis Carroll, which proves mathematically that no urn can have two balls of the same colour in it.

> Suppose that there are balls which are either black, B, or white, W, and two of unknown colours are contained in an urn with equal likelihood. That is,
>
> $$P(BB) = P(BW) = P(WB) = P(WW) = \frac{1}{4}.$$
>
> Now, consider adding a black ball to the urn so that
>
> $$P(BBB) = P(BWB) = P(WBB) = P(WWB) = \frac{1}{4}.$$
>
> Consider then selecting a ball at random. The chance that this ball is black is
>
> $$P(B) = \frac{2}{3}.$$
>
> Thus, there is a 2/3 chance that the ball we select is black, and as a result, there must be 2 black balls and 1 white ball, and so we could not have had two black balls to start!

a. Show that the probability of drawing a black ball in the described setup is indeed 2/3.
b. Is the logic sound? Why?

Exercise 4.5. Of people in a certain city who bought a new vehicle in the past year, 12% of them bought an electric vehicle and 5% of them bought an electric truck. Given that a person bought an electric vehicle, what is the probability that it was a truck?

Exercise 4.6. Mo and Fran each roll a die. The person who rolls the highest number wins; if they roll the same number, they both lose.

a. What is the probability that Fran wins?
b. If Mo rolls a 3, what is the probability that Mo wins?
c. If Mo rolls a 3, what is the probability that Fran wins?
d. If Mo wins, what is the probability that Fran rolled a 3?
e. If Mo wins, what is the probability that Mo rolled a 3?

Exercise 4.7. A geneticist is studying two genes. Each gene can be either dominant or recessive. A sample of 100 individuals has 56 individuals with both genes dominant, 6 individuals with both genes recessive, 24 individuals with only gene one dominant, and the remaining 14 with only gene two dominant.

a. What is the probability that a randomly sampled individual has dominant gene one?
b. What is the probability that a randomly sampled individual has dominant gene two?
c. Given that gene one is dominant, what is the probability that gene two is dominant?
d. These genes are said to be in linkage equilibrium if the event that gene one is dominant is independent of the event that gene two is dominant. Are these genes in linkage equilibrium?

Exercise 4.8. A lot of 10 components contains 3 that are defective. Two components are drawn at random and tested. Let A be the event that the first component drawn is defective, and let B be the event that the second component drawn is defective.

a. What is $P(A)$?
b. What is $P(B|A)$?
c. What is $P(A \cap B)$?
d. What is $P(A^C \cap B)$?
e. What is $P(B)$?
f. Are A and B independent?

Exercise 4.9. A lot containing 1000 components contains 300 that are defective. Two components are drawn at random and tested. Let A be the event that the first component drawn is defective, and let B be the event that the second component drawn is defective.

a. What is $P(A)$?
b. What is $P(B|A)$?
c. What is $P(A \cap B)$?
d. What is $P(A^C \cap B)$?
e. What is $P(B)$?
f. Is it reasonable to treat A and B as though they are independent?

Exercise 4.10. A certain delivery service offers both express and standard delivery. Seventy-five percent of parcels are sent by standard delivery, and the rest are sent by express. Of those sent standard, 80% arrive the next day, and of those sent express, 95% arrive the next day. A record of a parcel delivery is chosen at random from the company's files.

a. What is the probability that the parcel was shipped express and arrived the next day?
b. What is the probability that it arrived the next day?
c. Given the package arrived the next day, what is the probability that it was sent express?

Exercise 4.11. A quality control program at a food production facility involves inspecting finished product for safety. The proportion of items that actually are unsafe is 0.0002. If an item is found to be unsafe, the probability is 0.995 that it will fail the inspection. If an item is safe, the probability is 0.99 that it will pass the inspection.

a. If an item fails the inspection, what is the probability that it is unsafe?
b. Which of the following is more correct interpretation to the previous answer:

    i. Most items that fail inspection are safe.
    ii. Most items that pass inspection are unsafe.

c. If an item passes inspection, what is the probability that it is safe?
d. Which of the following is the more correct interpretation to the previous answer:

    i. Most items that fail inspection are unsafe.
    ii. Most items that pass inspection are safe.

e. Explain why a small probability in part (a) is not a problem, so long as the probability in part (c) is large.

Exercise 4.12. A patient goes to see a doctor. The doctor performs a test with 99 percent reliability–that is, 99 percent of people who are sick test positive and 99 percent of the healthy people test negative. The doctor knows that only 1 in 10000 people in the country are sick. If the patient tests positive, what are the chances the patient is sick?

Exercise 4.13. Let A and B be events with $P(A) = 0.8$ and $P(A \cap B) = 0.2$. For what value of $P(B)$ will A and B be independent?

Exercise 4.14. Let A and B be events with $P(A) = 0.5$ and $P(A \cap B^C) = 0.4$. For what value of $P(B)$ will A and B be independent?

Exercise 4.15. Prove that if $A \perp B$ then $A \perp B^C$, $A^C \perp B$ and $A^C \perp B^C$.

Exercise 4.16. Show that if $A \perp B$ then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

5 Summarizing Statistical Experiments with Random Variables

5.1 The Need for Random Variables

When introducing probability originally we worked from a sample space and then corresponding events. This is a very general framework which allows us to effectively analyze any statistical experiment. Sample spaces are not restricted to be numeric, for instance, and events are simply subsets of the sample space. As a result, this framework provides the tools for capturing uncertainty quantification in almost any setting. Still, the need to enumerate sample spaces and events over complex sets of arbitrary items is cumbersome and may prevent succinct representations of the underlying phenomena. Often, rather than caring about the entire space of outcomes from an experiment of interest, we are primarily concerned we a summary of the experiment. When we can summarize the experiment using a numerical quantity, we are able to define a random variable.

Definition 5.1 (Random Variable)**.** A random variable is a numeric quantity whose specific value depends on chance through the outcome of a statistical experiment. Formally, a random variable is a mapping from the result of an experiment to a set of numbers.

By reporting the numeric value of the random variable, we are able to summarize the key part of the experiment, succinctly. For instance, suppose that we are repeatedly tossing a coin. If we toss the coin 100 times, then the sample space is going to consist of 2^{100} total possible outcomes, each of which is a sequence of 100 heads and tails.[1] Instead, it may be more convenient to assign a random variable to be the number of heads that show up on the 100 tosses of the coin. In this case, the random variable takes on a non-negative integer between 0 and 100. In many situations, such a summary may be all the is relevant from the experiment.

It is important to recognize that information *is* lost when we define this random variable. If X summarizes the number of heads in 100 tosses of a coin, then provided X you are not able to answer the question "what was the 23rd toss of the coin?" As a result, random variables smooth over the unnecessary information, summarizing the parts of the statistical experiment that we care for. For any statistical experiment, however, we need to carefully define the

[1] Recall our discussions of combinatorics.

random variables which are truly of interest to us. For instance, if the 23rd toss of the coin was integral to our decision-making, then perhaps we define a random variable Y which counts the number of heads that show up on the 100 tosses of the coin, but does so with negative numbers if the 23rd toss was a tail, and with positive numbers if it was a head. Then, provided Y you can answer "how many heads showed up in the tosses of the coin? by calculating $|Y|$, and you can answer"what was the 23rd toss of the coin?" by looking at $\text{sign}(Y)$.

Example 5.1 (Random Variables at the Coffee Shop)**.** Back at their favourite coffee shop, Charles and Sadie are sitting in their favourite chairs, discussing random variables and watching the cash register for the inevitable sequence of customers who will arrive there. Charles suggests playing a game together, which is creatively called "how many random variables can we name that have to do with the statistical experiment of watching for customers at a cash register?" Seeing as how catchy the name is, Sadie is excited to play along, and so they start.

Name several distinct random variables which could be observed via the described statistical experiment.

Solution

There are essentially countless different random variables that could be named here, depending on what is of interest. The important concept is that each random variable needs to be a numeric quantity which is calculable from the statistical experiment. For instance:

- The number of people who arrive at the cashier in the next hour.
- The length of time until the next customer arrives at the cashier.
- The amount of money that the next customer spends when paying at the cashier.
- A value of 1 if the next customer is wearing a hat, and a value of 0 otherwise.
- The number of different items that are ordered by all customers over the next hour.
- The number of words that the cashier says to the next customer, prior to payment.
- A count of the number of red items of clothing that can be seen being worn by the next 15 customers.

Because of their ability to summarize effectively and flexibly the pertinent components of a statistical experiment, random variables are the default paradigm for discussing randomness. When discussing a random variable we will typically use capital letters, such as X, to represent the random quantity with an unknown value. In the event that an experiment is actually performed, and a value is realized for the random variable, we will record this value as a lower case letter, such as x. For instance, the number of heads showing in 100 flips of a coin is an unknown quantity depending on chance which we call X. Once we have flipped the coin 100 times and observed 57 heads, we denote this as $x = 57$.

The importance of this notation is merely to emphasize what values are unknown and random, and what values are known numeric quantities. Because x is a known value taking on some

set number we will not often speak of probabilities involving x.[2] Instead, we wish to translate the language of probability that we have built to statements regarding the random variable X.

The random variable X has a corresponding "sample space" of possible values that it can take on. We will typically refer to this as the *support* of the random variable, though borrowing other terms from math courses (such as *range*) will work also. This support, which we can think of as directly analogous to the sample space of arbitrary elements from before, will depend on the possible realizations from the underlying experiment. We typically will denote the support of a random variable X as $\mathcal{X}$. After the experiment has been performed the random variable will take on a single value from this set.[3] We can sometimes compactly describe the set of possible values for a random variable, for instance, by stating all of the integers, or integers between 5 and 10, or even values less than 1000. This allows for compact descriptions of $\mathcal{X}$ even when the set $\mathcal{S}$ is challenging to describe. The probability of realizing any outcome in $\mathcal{X}$ is dictated by the underlying probability model.

When introducing the concepts of probability we indicated that probability was assigned to events. With random variables, this remains true. As a result we need to define events in terms of random variables. When we have a random variable, X, an event is defined as any set of values that it can take on. For instance, we may have the event $X = 4$, or the event $X \geq 18$, or the event $2 \leq X \leq 93$, or the event $X \in \{2, 4, 6, 8, 10\}$. In each case these are subsets of the possible values that the random variable can take on, based on the outcome of the experiment.[4]

Just as before these events can be simple events (comprised of a single outcome) or compound events (comprising of multiple outcomes). The event $\{X = 5\}$ is a simple event, whereas the event $\{X \geq 25\}$ is a compound event. With events defined in this way, we can translate the other concepts from the explicit event-based probability. Specifically, we think of the experiment producing a numerical outcome, and then use the same sets of tools as before applied to numeric events.

Example 5.2 (Expanding on Random Variables at the Coffee Shop)**.** After tiring of their game of "how many random variables can we name that have to do with the statistical experiment of watching for customers at a cash register?", Charles and Sadie fall into a deeper discussion of some of the some of their favourite random variables named during the play through. They

[2]We can actually discuss probabilities revolving around x, but they are all deeply uninteresting. Every probability will either be 1, or 0. If I tell you that $x = 5$, then $P(x = 5) = 1$ and $P(x = 2) = 0$. While this is not a particularly *interesting* probability statement, it is true, and somewhat surprisingly, can arise in meaningful ways.

[3]The support of a random variable is directly analogous to the sample space from the statistical experiment. That is, it is the set of possible observations which can be made of the random variable. As a result, it is *sometimes* permissible to call the support the "sample space of the random variable", in an informal manner. The informality here is important. Formally, a random variable X is a function which maps from the sample space to its support, which is to say $X : \mathcal{S} \to \mathcal{X}$.

[4]That is to say, events in the case of random variables are subsets of $\mathcal{X}$.

take turns describing the support of a chosen random variable, as well as listing examples of simple and compound events that can be translated into the language of random variables.

Choose a random variable identified in Example 5.1 and indicate its support, as well as possible simple and compound events that could be observed in terms of it.

Solution

Suppose that we take X to be a random variable representing "The number of people who arrive at the cashier in the next hour." The support of this random variable is likely best select as the non-negative integers. That is, we take

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

There may be a reasonable maximal value for the random quantity (such as knowing the number of people who are within an hour radius of the coffee shop, or else knowing the number of people who are presently alive), however, the simpler "all non-negative integers" will likely suffice.
For simple events we may consider $\{X = 0\}$ or $\{X = 10\}$ or $\{X = 31415\}$. In every case, the simple event takes the form $\{X = x\}$.
For compound events we could consider $\{X > 0\}$, representing the event that at least one customer arises, we could consider $\{X \leq 10\}$, the event where no more than 10 customers show up, or we could consider something a little more abstract, like $X \in \{0, 3, 5, 9, 19, 25\}$ or X is an even number. In all these cases the key point is that compound events have more than one way of being satisfied (in that they correspond to more than one event).

When considering random variables there is a key distinction between two types of random variables, discrete and continuous.

Definition 5.2 (Discrete Random Variable)**.** A discrete random variable is any random variable whose support is either finite or else countably infinite.

Definition 5.3 (Continuous Random Variable)**.** A continuous random variable is any random variable whose support is uncountably infinite.

Countable and Uncountable Sets

We say that a set is **countable** whenever we can enumerate the elements of the set using the positive integers. If we have a set with a finite number of elements in it, then we say that it is countable because each element in the set can get assigned an integer value (just using 1 through to the cardinality of the set).
If we take a set like the positive integers, $\{1, 2, 3, \dots\}$, then this is an infinitely large set. However, it is still countable because we can assign each element of the set an integer

value (just use the same integer it is corresponding to!). What if we have the set of positive, even integers ({2,4,6,8,...})? Each of these is just 2 times one of the counting numbers, in order, and so these too are countable.
What if we took the set of real numbers[5] in the interval $[0, 1]$? In this case there is **no way** to map each of these values to the values $\{1, 2, 3, 4, \dots\}$ and as a result the interval $[0, 1]$ is **not countable**. It is infinitely large, but it remains infinitely large even after we have enumerated an infinite set of the items in it.
A key difference between countable and uncountable sets is that, with a countable set, we can[6] sum over the elements of the set. We cannot sum over the elements of an uncountable set, as there would be no way to actually order and enumerate the items. In this course the detailed mathematics of countable versus non-countable sets is tangential to the main point which is in distinguishing between discrete and continuous random variables.

Typically, discrete random variables will take on some collection of the integers, where continuous random variables will be defined on some interval (or set of intervals). That is, we may take discrete random variables to be defined on $\{0, 1, 2, 3, \dots, 100\}$ or $\{0, 1, 2, 3, \dots\}$ or $\mathbb{Z}$[7]. For continuous random variables we may take $X \in [0, 1]$ or $X \in (0, \infty)$ or $X \in (-\infty, 129]$.

Example 5.3 (The Countability of Coffee Shop Random Variables)**.** Charles and Sadie have been discussing the features of all of the random variables they identified at the coffee shop for a long time when Sadie points out that they have not been differentiating between discrete and continuous random variables. Charles thinks "how could we have overlooked this?" and seeks to remedy the situation, immediately!

List at least one discrete and one continuous random variable that could arise via the statistical experiment of watching the cashier at a coffee shop over time (as described in Example 5.1).

Solution

For this we categorize the random variables originally defined in the solution to Example 5.1.

- "The number of people who arrive at the cashier in the next hour."

  "The number of different items that are ordered by all customers over the next hour."

  "The number of words that the cashier says to the next customer, prior to payment."

  "A count of the number of red items of clothing that can be seen being worn by the next 15 customers."

[6]Recall that "real numbers" are just all of the standard numbers that we think of (decimals, whole numbers, fractions, negative values, etc.).
[6](in principle)
[7]This notation refers to the set of all integers

These are all **discrete** random variables, as the only values they can take on are the counting integer values.

- "The length of time until the next customer arrives at the cashier." This is a **continuous** random variable, as it can take on any value over an interval (say, the interval $(0, 28800)$ if the store is open for 8 hours and the customer were there the whole time).
- "The amount of money that the next customer spends when paying at the cashier." This is a **discrete** random variable, as it must take a finite set of values. One way to make this clear is to count the value in cents, and then it will be only integer values.
- "A value of 1 if the next customer is wearing a hat, and a value of 0 otherwise." This is a **discrete** random variable as it can only be from $\{0, 1\}$.

5.2 Probability Distributions and Probability Mass Functions

We will discuss continuous random variables later. For now, we turn our focus to discrete random variables. One of the major utilities of random variables is that they provide a shorthand for summarizing the results of a statistical experiment. To this end, there are several tools which have been developed centered on random variables which help to expedite the analysis of the corresponding probabilities. Chief among these tools is the concept of a probability distribution.

Definition 5.4 (Probability Distribution)**.** A probability distribution is a mathematical statement describing the probabilistic behaviour of a random variable.

Distributions capture the underlying random behaviour of the random variables of interest, and in so doing, summarize information regarding the experiment or process that is being considered. When concerned with discrete random variables, we typically summarize probability distributions through the use of a probability mass function.

Definition 5.5 (Probability Mass Function)**.** A probability mass function is a function, $p(x)$, which maps possible values for a discrete random variable (the set $\mathcal{X}$) to the probabilities corresponding to those events.[8]

If a random variable X has support $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$, then a probability mass function is a function $p(x)$ such that $p(x_1) = P(X = x_1)$, $p(x_2) = P(X = x_2)$, and so on through

[8] That is, $p(x)$ is a function, $p \colon \mathcal{X} \to [0, 1]$.

to $p(x_k) = P(X = x_k)$. Once a probability mass function is known, all of the probabilistic behaviour of the random variable is fully described.

Example 5.4 (The Coin Game for Three)**.** Sometimes Charlie and Sadie are joined by their friend Garth on their trips to the coffee shop. Also a probability aficionado, Garth contently joins in the game with Charles and Sadie where three coins are flipped, and depending on the results, one friend pays for the group. Garth's order is typically less than Charles and Sadie and so Sadie proposes the following modified game.

> A fair coin is tossed three times. If all tosses of the coin show the same symbol, Garth pays. Otherwise, if there are more heads than tails, Charles pays. Finally, if there are more tails than heads, Sadie pays.

Help to simplify this statistical experiment by first defining a random variable, X which can be used to encode this game and then record the probability mass function of X.

Solution

One choice of X is to allow X to be the number of heads that show on three tosses of the coin. In this case we have $\mathcal{X} = \{0, 1, 2, 3\}$. If $x = 0$ or $x = 3$ is observed then Garth pays. If $x = 2$ is observed then Charles pays. Finally, if $x = 1$ is observed, Sadie pays. Thus writing down the probability mass function for X also provides an easy way for computing the probability of each player having to pay.
In order for $X = 0$, we must have all tails come up. There are $2^3 = 8$ total possible sequences of 3 coin tosses, and only 1 of these results in $X = 0$, therefore $P(X = 0) = p(0) = \frac{1}{8}$. The same logic applies to $X = 3$ where we need to toss all heads. There are several techniques for finding $p(1)$ and $p(2)$.[9] The most direct way is to recognize that there are exactly 3 ways of observing 1 head (it can be in the first, second, or third toss). Thus, $p(1) = \frac{3}{8}$. For $p(2)$, the same logic applies except we ask "where is the one tail?" to give $p(2) = \frac{3}{8}$. Then, taken together, this results in

$$p(x) = \begin{cases} \frac{1}{8} & x \in \{0, 3\} \\ \frac{3}{8} & x \in \{1, 2\} \\ 0 & \text{otherwise.} \end{cases}$$

We can also read off from this result that Garth pays $\frac{1}{4}$ of the time in the long run, while Sadie and Charles each pay $\frac{3}{8}$ of the time.[10]

[10] One particularly elegant technique is to realize that $p(1)$ and $p(2)$ must be the same by interchanging the roles of heads and tails. Thus, we have $p(1) + p(2) + 0.25 = 1$ by the unitary property of probability, and rearranging gives $p(1) = p(2) = 0.375$.

[10] We can actually write down this probability mass function more succinctly as $p(x) = \binom{3}{x}\frac{1}{8}$. You can check this holds for yourself, and we will understand why later on!

The previously outlined conditions for probabilities must still hold when using probability mass functions. As a result, we know that probabilities are all between 0 and 1, and so we must have $0 \leq p(x) \leq 1$, for all $x \in \mathcal{X}$. Moreover, we saw previously that summing the probabilities over the full sample space gave a value of 1. Correspondingly, we must have that

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

These two properties are often used to define a *valid* probability mass function, and we can use these properties to both check whether a given mass function is valid and to turn a given function into a valid probability mass function.

Example 5.5 (Charles's Messy Writing Strikes Again)**.** Charles is reading through some notes regarding a new game under development which, like most good games, relies at least a little bit on chance. In the notes there is a probability mass function written down, which, the best Charles can make out, reads

$$p(x) = \begin{cases} 3c & x = 0 \\ 0.6 & x = 1 \\ 7c & x = 2. \end{cases}$$

Frustrated with the illegibility, Charles brings the problem to Sadie who points out, if it really is $p(0) = 3c$ and $p(2) = 7c$, they actually have all the information required to solve the problem.

What is the probability mass function, assuming Charles is reading it correctly.

Solution

We know that two properties must be true of all probability mass functions. First $0 \leq p(x) \leq 1$. Second, $\sum_{x \in \mathcal{X}} p(x) = 1$. The first property tells us that $0 < 3c < 1$ and that $0 < 7c < 1$. This means that $c > 0$, and that $c < 1/3$ and $c < 1/7$. Using the second property we get that

$$1 = 3c + 0.6 + 7c \implies 0.4 = 10c.$$

This tells us that $c = 0.04$, which satisfies all of the previous properties. Taking $c = 0.04$ gives a probability mass function of

$$p(x) = \begin{cases} 0.12 & x = 0 \\ 0.6 & x = 1 \\ 0.28 & x = 2. \end{cases}$$

When solving questions related to probabilities using a probability mass function, the same secondary properties apply. Notably, if we want to know $P(X \in A)$, for some set of possible values A, then we can write

$$P(X \in A) = \sum_{x \in A} p(x).$$

This can be particularly helpful, for instance, if we want to know $P(X \leq c)$ for some constant value c. In this case we know that the possible values range from teh smallest value X can take on, through to c, giving for instance,

$$P(X \leq c) = \sum_{x=0}^{c} p(x),$$

if $X \geq 0$. Rules regarding the complements of events continue to hold as well, where for instance, $P(X > c) = 1 - P(X \leq c)$, giving a useful avenue for simplifying probability calculations.

Example 5.6 (Charles and Sadie: Amateur Ornithologists)**.** Charles and Sadie watched a documentary about birds together, which had Sadie become quite interested in bird watching. Charles is skeptical[11] but agrees to go along with Sadie, supposing that there is a high enough probability of something interesting happening. Sadie scours the scholarly literature and determines that, in the area, the probability of seeing x rare birds over a five hour birding session has a probability mass function given by

$$p(x) = \frac{e^{-3}3^x}{x!},$$

where $x \geq 0$.[12]

What is the probability that during their bird watching adventure, Charles and Sadie see at least 1 rare bird?

Solution

We are interesting in the probability $P(X \geq 1)$, which we can write down explicitly as

$$P(X \geq 1) = \sum_{x=1}^{\infty} \frac{e^{-3}3^x}{x!}.$$

While, strictly speaking, it is possible to solve this infinite summation, it is an infinite summation and we would rather not. Instead we can realize that, if we take the event $A = \{X \geq 1\}$ then A^C is the event $\{X < 1\} = \{X = 0\}$. As a result, using our elementary rules of probability we get that

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-3}3^0}{0!} = 1 - e^{-3} \approx 0.95.$$

As a result, there is an approximately 0.95 probability that they see a rare bird while bird watching for five hours.

[11] Charles is still not entirely sure if birds are even real, let alone worth watching.
[12] How Sadie found such a concrete answer, I will never know.

With events defined in terms of random variables, we can talk about events as being independent of each other or mutually exclusive using the familiar definitions. On a related note, we can talk of joint and conditional probabilities, relating to multiple events. With joint probabilities, it is often easiest to combine the event into a single, compound event, and find the marginal probability of that event. For instance, if you have the events X is even and $X \leq 15$, then the intersection of these events is $X \in \{2, 4, 6, 8, 10, 12, 14\}$ (supposing $X > 0$).

Example 5.7 (Charles: The Ornithological Photographer)**.** After a single time bird watching Charles is hooked! The next time that they go out, Charles brings a camera to snag some beautiful memories of the majesty they are witnessing. The only trouble is that Charles is not particularly good at taking photographs of stills, let alone of creatures that can fly. Charles is not yet certain, but suspects that every time a photograph is snapped, there is a 0.1 probability that it turns out good. Because the camera is a film camera, the number of photos taken is a key metric, and Charles works out that, if X is the number of bad photographs taken for every good photograph, the probability mass function for X is given by

$$p(x) = (0.1) \times (0.9)^x,$$

where $x \geq 0$ is an integer.

a. What is the probability that Charles takes exactly three bad photos before the first good one?
b. If A is the event that Charles takes three bad photos before the first good one, and B is the event that Charles takes more than four photos before the first good one, what is $P(A, B)$? What does this means about A and B?
c. Suppose that Charles has taken 2 photos, both of which are bad. What is the probability that at least two more bad photos are taken before a good one?
d. What is the probability that at least 2 bad photos are taken?
e. Do the results of (c) and (d) suggest an independence?
f. List two events, not previously described, which are independent of one another.

Solution

a. Here, we want $p(3) = (0.1) \times (0.9)^3 = 0.0729$.

b. Considering $A \cap B$ we have $\{X = 3\} \cap \{X \geq 4\} = \emptyset$. As a result, $P(A, B) = 0$ since there is no overlap between A and B. As a result, these events are mutually exclusive.

c. Here, we can frame this probability as

$$P(X \geq 4 | X \geq 2)$$

This can be directly computed using the definition of conditional probability,

$$\begin{aligned} P(X \geq 4 | X \geq 2) &= \frac{P(X \geq 4, X \geq 2)}{P(X \geq 2)} \\ &= \frac{P(X \geq 4)}{1 - P(X < 2)} \\ &= \frac{1 - P(X < 4)}{1 - (0.1 + 0.1 \times 0.9)} \\ &= \frac{1 - (0.1 + 0.1 \times 0.9 + 0.1 \times 0.9^2 + 0.1 \times 0.9^3)}{1 - (0.1 + 0.1 \times 0.9)} \\ &= 0.81. \end{aligned}$$

d. Through direct calculation we get

$$P(X \geq 2) = 1 - P(X < 2) = 1 - (0.1 + 0.1 \times 0.9) = 0.81.$$

e. Combining (c) and (d) we find that the probability that an additional two photos are required, given that two have already been taken, is the same as the probability that two photos are required, without knowing that any have been taken. While this feels like an independence type of property, it is not precisely independence. Note that the event $\{X \geq 4\}$ immediately gives the event $\{X \geq 2\}$ and so they cannot be independent of one another.[13] For this to be independent we would require $P(A, B) = P(A)P(B)$ which is not true of any of the events discussed.

f. One option is to consider the event $X \geq 0$ with any other event. Because $\{X \geq 0\} = \mathcal{X}$, no information can be gained conditioning on this event.

5.3 Multiple Random Variables and Joint Probability Mass Functions

While we can discuss independence, joint probabilities, and conditional probabilities relating to events on the same random variable, it is often of interest to combine multiple random variables. Sometimes these random variables will be multiple versions coming from the same distribution, and other times they will be coming from multiple different distributions. In either event, frequently our main concern is in summarizing the probabilistic behaviour of two or more random quantities.

[13] Instead of independence, we may describe this distribution as being "memoryless". If you know that you have been waiting a certain amount of time for something to happen (a good picture to get taken) that does not change your belief about how much longer you have to wait; the process "forgets" that it has been ongoing for sometime.

A Discussion of Distributional Language

Note that when we talk of a random variable "following" a particular distribution, we are saying that the probability mass function of the random variable is described by that distribution's mass function. Thus, if two random variables "share a distribution", we just mean that their probabilistic behaviour is described by the same underlying mass function.
For instance, if I flip a coin 10 times, and you flip a different coin 10 times, and we each count the number of heads that show up, we can say that the random quantity for the number of heads I observe will have the **same distribution** as the random quantity for the number of heads that you observe.
These two quantities are not equal, in general, but they are described by the same random processes.

When we describe the distribution of a particular random variable, we are implicitly describing the marginal probabilities associated with that quantity. Just as before, the marginal probabilities describe the behaviour of the random variable alone. What happens when we want to be able to describe multiple components of an experiment, together? For this we require extending the idea of a joint probability beyond the concept of events.

Suppose that we roll two six-sided fair dice. Let X denote the sum of the two dice, and let Y denote the maximum value showing on the two dice. X is a discrete random variable taking on values between 2 and 12, while Y is a discrete random variable taking on values between 1 and 6. The supports for the two random variables are different from one another and so immediately we know that their probabilistic behaviour must be different, despite the fact that both random variables summarize the same statistical experiment.

We can also immediately see that the two random variables, while not equal to each other, are dependent. For instance, if you know that $Y = 1$ then you know that $X = 2$, and if you know that $Y = 3$, then you know that $X \leq 6$. To begin to capture the joint behaviour of X and Y we introduce the joint distribution and joint probability mass function.

Definition 5.6 (Joint Distribution)**.** A joint probability distribution describes the joint probabilistic behaviour of two or more random variables, simultaneously.

Definition 5.7 (Joint Probability Mass Function)**.** A joint probability mass function describes the behaviour of the joint distribution of multiple random variables. For a set of random variables, $X_1, \ldots, X_n$, the joint probability mass function assigns a probability value for every *tuple* of values that $(X_1, \ldots, X_n)$ can take on.

The joint probability mass function is analogous to the marginal probability mass function, only it considers joint events rather than marginal ones. Suppose that you have two random variables, X and Y. The joint probability mass function assigns a probability value for every

pair of values that (X, Y) can take on. That is, $p_{X,Y}(x, y) = P(X = x, Y = y)$. Then, once you know the joint behaviour of X and Y, you can fully summarize the combined behaviour of the underlying experiment.

Example 5.8 (Charles and Sadie's Bird Outings)**.** Charles and Sadie have both gotten deeply into their ornithological adventures. They go on trips together, Sadie is responsible for spotting the rare birds, and then Charles for snapping the photos. Charles has settled on a camera setup that allows for 10 pictures to be taken before changing the film. The strategy that they follow is to go out and look for a rare bird. When one is spotted, Charles takes 10 photos, trying to get as many good photos as possible. Then, the film is replaced and they repeat the process.

If X is a random variable representing the number of good photos that are taken, and Y is the number of birds that are seen on the trip, then the joint probability mass function of X and Y is

$$p_{X,Y}(x, y) = \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!},$$

with $0 \leq x \leq 10y$, and $y \geq 0$.

a. What is the probability that one bird is seen and there are no good photos taken?
b. What is the probability that there is one or fewer good photos taken and two birds seen?

Solution

a. Here we want $Y = 1$ and $X = 0$. Thus, we compute

$$p_{X,Y}(0, 1) = \binom{10(1)}{0} \times (0.25)^0 \times (0.75)^{10(1)-0} \times \frac{e^{-3}3^1}{1!} \approx 0.008411.$$

b. Here we want $\{X \leq 1\}$ and $Y = 2$. This is the same as $X = 0$ with $Y = 2$ or $X = 1$ with $Y = 2$. Thus,

$$\begin{aligned} p_{X,Y}(0, 2) + p_{X,Y}(1, 2) &= \binom{10(2)}{0} \times (0.25)^0 \times (0.75)^{10(2)-0} \times \frac{e^{-3}3^2}{2!} \\ &\quad + \binom{10(2)}{1} \times (0.25)^1 \times (0.75)^{10(2)-1} \times \frac{e^{-3}3^2}{2!} \\ &\approx 0.00547. \end{aligned}$$

5.3.1 Joint Probability Distributions as Contingency Tables

In practice, joint probability mass functions can be thought of as analogous to the contingency tables we previously saw. If the first variable represents the first random variable being con-

sidered, and the second variable represents the second random variable, then each cell of the contingency table assigns a probability to one of the joint events that could be observed in the experiment. Joint distributions are a useful generalization of contingency tables as they allow us to compactly represent not only two different random variables, but sometimes many more. All of the definitions used for the case of two random variables extend naturally to three, four, and beyond.

Example 5.9 (Charles and Sadie: Rock-Paper-Scissors Experimentation)**.** Charles and Sadie were once asked if rock-paper-scissors may be an easier way to solve their "who pays for coffee" dilemmas. Though both of them rejected the description of their coffee games as a "dilemma", they had never really given it much thought. One day while bird watching, during a particularly long break with no birds, they begin to think through whether this could work or not. To this end, the friends play 1000 games, recording the results of each game into a contingency table. They suspect that this provides the true long run proportion of occurrences. To encode these games numerically, they take $X = -1$, $X = 0$, and $X = 1$ to represent Charles playing rock, paper, and scissors respectively, and they take $Y = -1$, $Y = 0$, and $Y = 1$ to represent the same events from Sadie.

	$Y = -1$	$Y = 0$	$Y = 1$
$X = -1$	400	100	50
$X = 0$	10	200	40
$X = 1$	50	50	100

What is the joint probability mass function that corresponds to this contingency table?

Solution

We can explicitly enumerate the possibilities to define the joint probability mass function. That is

$$p_{X,Y}(x,y) = \begin{cases} 0.4 & x = -1, y = -1 \\ 0.1 & x = -1, y = 0 \\ 0.05 & x = -1, y = 1 \\ 0.01 & x = 0, y = -1 \\ 0.2 & x = 0, y = 0 \\ 0.04 & x = 0, y = 1 \\ 0.05 & x = 1, y = -1 \\ 0.05 & x = 1, y = 0 \\ 0.1 & x = 1, y = 1. \end{cases}$$

While it may be possible to express this in a more compact way, this fits the criteria for a joint probability mass function.

5.4 Independence of Random Variables

If we continue to consider the case of a bivariate (two variable) joint distribution, we can use this setting to introduce the independence of random variables. Recall that the joint probability mass function of X and Y is a function, $p_{X,Y}(x, y) = P(X = x, Y = y)$. We have also introduced the marginal mass functions, $p_X(x) = P(X = x)$ and $p_Y(y) = P(Y = y)$. Further, we have said that two events, A and B, are independent if their joint probability is equal to the product of their marginal probabilities, that is $P(A, B) = P(A)P(B)$.

If we take $A = \{X = x'\}$ and $B = \{Y = y'\}$, then if $A \perp B$ we can write $p_{X,Y}(x', y') = p_X(x')p_Y(y')$. If this holds for every possible x' and every possible y', then we say that X and Y are independent random variables, and we write $X \perp Y$.

Definition 5.8 (Independent Random Variables)**.** Two random variables, X and Y are said to be independent random variables if their joint probability mass function is given by the product of their marginal probability mass functions,

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

We write $X \perp Y$, and read: X is independent of Y.

In words, two random variables are independent whenever all possible combinations of events between them are independent.

Example 5.10 (Charles the Sports Fan)**.** Charles is a major sports fan and has been particularly fond of *hurling* ever since a trip to Ireland. Each week, Charles watches with deep investment, becoming very attached to the outcome. Because of this attachment, Charles is willing to do just about anything to help out from behind the television set, which mostly consists of wearing the right coloured clothing. Charles's theory is that by increasing the number of articles of green clothing that are worn, the number of scores in the game will also increase.

Let X represent the number of articles of green clothing that Charles is wearing, with $X \in \{0, 1, 2, 3, 4\}$, and Y is the number of scores in the game. Suppose that

$$p_X(x) = \frac{(x-2)^2}{10},$$

and that

$$p_Y(y) = \frac{e^{-45} \times 45^y}{y!}.$$

a. What is the joint probability mass function for X and Y, assuming that they are independent?
b. If it is determined that, assuming Charles wears no green clothing, the probability of 50 scores is 0.01, are these random variables independent?

Solution

a. If $X \perp Y$, then the joint probability mass function is the product of the two, which is to say

$$p_{X,Y}(x,y) = \frac{(x-2)^2}{10} \cdot \frac{e^{-45} \times 45^y}{y!} = \frac{e^{-45}}{10} \times (x-2)^2 \times \frac{45^y}{y!}.$$

b. Using the joint probability mass function found in (a),

$$p_{X,Y}(0,15) = \frac{e^{-45}}{10} \times (0-2)^2 \times \frac{45^{50}}{(50)!} \approx 0.017.$$

This is not equal to 0.01, and so it must be the case that $X \not\perp Y$.

Example 5.11 (The Independence Argument)**.** Charles and Sadie are having a disagreement about the nature of the probabilities on their bird outings. Charles claims that, without knowing the marginal probability mass functions of the number of good pictures taken, and the number of rare birds that are seen, there is no way to tell whether the the two random quantities are independent or not. Sadie cannot say exactly why this argument feels wrong, but insists that the two quantities must not be independent.

Who is correct, and why? Recall that

$$p_{X,Y}(x,y) = \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!},$$

with $0 \leq x \leq 10y$, and $y \geq 0$

Solution

In this case Sadie is correct. The most clear way of seeing this is by noting that the support set for X depends on Y. The question to ask yourself is: how could two functions which are independent of one another multiply together so that one's support depends on the other's? It cannot happen. Put differently: we know that these cannot be independent as, if $Y = 0$ then $P(X = 0) = 1$ as Charles will not take any pictures.

Equivalent Definition of Independence

There is an equivalence between the described definition, and a slightly more intuitive definition for independence.
Whenever $X \perp Y$ any two events corresponding to X and Y, say $X \in A$ and $Y \in B$ are independent. The subtle distinction is that in our previous definition, we were only

concerned with simple events of the form $X = x'$ or $Y = y'$. Here we allow any two arbitrary events.
Note that if we have the above definition holding for simple events, then

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x \in A} P(X = x, Y \in B) \\ &= \sum_{x \in A} \sum_{y \in B} P(X = x, Y = y) \\ &= \sum_{x \in A} \sum_{y \in B} p_X(x) p_Y(y) \\ &= \left(\sum_{x \in A} p_X(x) \right) \left(\sum_{y \in B} p_Y(y) \right) \\ &= P(X \in A) P(Y \in B). \end{aligned}$$

Even by only making the assumption for simple events, the conclusion regarding compound events follows naturally. Whenever any two random variables are known to be independent we know that any two events corresponding to these random variables will be independent. Moreover, we can directly write down the joint probability mass function by taking the product of the two marginals.

5.5 Conditional Probability Distributions

When introducing events, we discussed how the concepts of independence and dependence could be understood more intuitively through the use of conditional probabilities. The same is true for random variables.

Definition 5.9 (Conditional Distributions)**.** A conditional distribution of a random variable captures the probabilistic behaviour of a random variable, given information regarding another (or several other) random variables.

Definition 5.10 (Conditional Probability Mass Function)**.** A conditional probability mass function assigns probability values associated with any conditional event between multiple random variables. For instance, if there are two random variables X and Y, the conditional probability mass function of X given Y characterizes events of the form X given $Y = y$. Mathematically, the conditional mass function is

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

This definition is analogous to the formula for conditional probabilities more generally, give by the joint distribution over the marginal distribution. To determine the probability of any event (for X) given some information about Y, you plug-in $X = x$ and $Y = y$ into the conditional probability mass function. If you want to condition on more than one random variable, the quantities extend in exactly the same way, where for instance

$$P(X = x|Y = y, Z = z) = \frac{P(X = x, Y = y, z = z)}{P(Y = y, Z = z)}.$$

Example 5.12 (Charles Accepts the Argument)**.** After some convincing, Charles accepts the argument that the number of good pictures taken and the number of rare birds seen cannot be independent events. What really makes this clear, however, is when Sadie points out that the marginal probability mass function for the number of rare birds seen is given by

$$p_Y = \frac{e^{-3}3^y}{y!}.$$

Taken together with the fact that

$$p_{X,Y}(x, y) = \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!},$$

with $0 \leq x \leq 10y$, and $y \geq 0$, Charles realizes that the conditional distribution of $X|Y$ can be computed.

a. What is the conditional probability mass function of X given Y?
b. Given that 2 birds are seen, what is the probability that any good photos are taken?

Solution

a. For the conditional probability mass function we take $p_{X,Y}(x, y)/p_Y(y)$\$. This gives

$$p_{X|Y}(x|y) = \frac{\binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!}}{\frac{e^{-3}3^y}{y!}} = \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x}.$$

b. To represent this, we want $P(X \geq 1|Y = 2)$. From properties of probability, we know that $P(X \geq 1|Y = 2) = 1 - P(X < 1|Y = 2) = P(X = 0|Y = 2)$, and so using the conditional probability mass function found previously, this gives

$$P(X = 0|Y = 2) = \binom{20}{0}(0.25)^0 \times (0.75)^{20} = 0.75^{20}.$$

As a result, the probability of interest is $1-0.75^{20} \approx 0.9968$, which is the probability that Charles takes at least one good picture if two birds are seen.

As was discussed, the joint probability mass function of two independent random variables is given by the product of the marginals. If $X \perp Y$, then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$. Combined with the expression for the conditional probability mass function this results in $p_{X|Y}(x|y) = p_X(x)$. That is, whenever two variables are independent, the conditional probability mass function is exactly equal to the marginal probability mass function.

We saw that this was true for events, and the same reasoning applies here. This result gives an intuitive method for interpreting the independence of random variables. Two random variables are independent whenever any information about the one does not provide information about the other; when they are completely uninformative for one another. With this intuitive description it is easier to infer when independence of random variables is reasonable. Doing so is a useful skill for manipulating probability expressions.

Example 5.13 (Charles Exploration of Independence)**.** A little shaken after the previous independence mishap, Charles is committed to better understanding independence at an intuitive level. As a result, every time a pair of random quantities are seen together Charles has taken to deciding whether or not the underlying random variables would be independent.

For each of the following, indicate whether the pairs of random quantities are likely to be independent (and why).

a. Charles is buying a butternut squash, and is considering the length of the squash (X) and the weight of the squash (Y) as the most relevant measurements.
b. Charles is sitting at an intersection which has a separated bike path in front of it, and a lane of vehicular traffic. Take X to be the number of bikes passing by over a length of time, and Y to be the number of cars passing by over a certain time.
c. As the weekend comes around Charles prepares for the hurling match, once again considering whether the number of green items of clothing worn (X) has an impact on the number of scores (Y).
d. Charles and Sadie are playing battle dice, where Charles rolls a die and compares the result (X) to the result of a separate die rolled by Sadie (Y).

Solution

a. The length of a butternut squash and its weight are likely dependent. The longer a squash is, all else equal, the more likely it will be heavier, and vice versa. Both of these measure the size of the squash, and as a result, will be connected.

b. This is an interesting example which likely could be argued either way. On one hand, if cars and bikes do not mix, then the traffic of one will not likely impact the traffic on another. However, there are plenty of reasons why a larger number of bikes may suggest a certain number of cars: (1) perhaps more people bike on the weekends and fewer people drive on the weekends; (2) perhaps a city with more bikers has more cars; (3) perhaps more bikes represents a busy time of the day,

which would also lead to more cars. It is challenging to know precisely where the impact stems from, it seems likely that there would be some relationship.

c. There is likely no impact of what Charles is doing and what is happening in a sporting match being watched on the TV. As a result, these two quantities should be independent.

d. The die roll from Charles has no impact on the die roll from Sadie, and vice versa. As a result, these two random variables will almost certainly be independent.

5.6 Manipulating Probabilities with Random Variables

Seeing as the joint, marginal, and conditional probability mass functions are exactly analogous to the corresponding concepts when they were introduced regarding events, it is reasonable to assume that we can extend the multiplication rule, the law of total probability, and Bayes' theorem to the framework of probability functions. Indeed, each of these relationships continues to hold for random variables in much the way that would be expected.

Multiplication Rule (with Probability Mass Functions)

Translating the multiplication rule to use probability mass functions gives

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x).$$

This can be seen by rearranging the relationship defining the conditional probabilities.

Bayes' Theorem (with Probability Mass Functions)

Bayes' Theorem can be rewritten using probability mass functions as well. Specifically,

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}.$$

This follows directly from the definition of conditional probabilities, as well as the multiplication rule.

These rules give the ability to compute the joint distribution and the other conditional information, when we have information regarding some of the marginals and some of the conditionals. These properties are used less *explicitly* when dealing with probability mass functions directly. Instead, they almost become absorbed into the fabric of the defining relationships themselves. That is to say, you are less likely to see Bayes' theorem invoked directly when moving between conditional distributions; however, moving between conditional distributions *is* an important

skill, and requires the use of Bayes' Theorem.

Example 5.14 (Charles and the Chores)**.** Charles has decided to take a break from having paid employment, instead taking time to help ensure that Sadie's house runs smoothly as Sadie has been busy trying to write a novel. Unfortunately, sometimes chores are not done in a timely manner, despite Charles's best efforts. While Sadie has been nothing but supportive, Charles decides to turn to probability to ease uncertainties.

Define $X = 1$ if a given chore was done by Charles, with $X = 0$ if it was done by Sadie. Let $Y = 1$ denote a chore being done on time with $Y = 0$ if it was late. Suppose that

$$\begin{aligned} p_{Y|X}(y|x) &= (0.5 + 0.4x)^y \times (0.5 - 0.4x)^{1-y}, \\ p_X(x) &= 0.85^x \times (0.15)^{1-x}, \\ p_Y(y) &= .85 \times (0.9)^y (0.1)^{1-y} + 0.075. \end{aligned}$$

a. What is the probability that a chore is late and done by Charles?
b. What is the probability that, given a chore was done by Charles, it was done late? What about with Sadie?
c. What is the probability that, given a chore was late, it was done by Charles? What about with Sadie?
d. Explain how the results of (b) and (c) may result in Charles feeling responsible for the late chores. Is that accurate?

Solution

a. For this question, we require the joint probability mass function we apply the product rule. That is,

$$p_{X,Y}(x, y) = p_{Y|X}(y|x)p_X(x) = (0.5 + 0.4x)^y \times (0.5 - 0.4x)^{1-y} \times 0.85^x \times (0.15)^{1-x}.$$

Then, to find the probability that a chore is late and done by Charles we take $p_{X,Y}(1, 0) = 0.1 \times 0.85 = 0.085$.

b. For Charles, we want $P(Y = 0|X = 1)$, which can be solved for directly using the provided conditional probability mass function. That is,

$$p_{Y|X}(0|1) = (0.5 + 0.4)^0 \times (0.5 - 0.4)^1 = 0.1.$$

For Sadie, we want $P(Y = 0|X = 0)$. This gives

$$p_{Y|X}(0|0) = (0.5)^0 \times (0.5)^1 = 0.5.$$

c. Here we require the alternative conditional distribution, $P(X|Y)$. For this we can leverage Bayes' Theorem. Specifically,

$$p_{X|Y} = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} = \frac{(0.5 + 0.4x)^y \times (0.5 - 0.4x)^{1-y} \times 0.85^x \times (0.15)^{1-x}}{.85 \times (0.9)^y (0.1)^{1-y} + 0.075}.$$

Then, the probability that Charles did the chore, given it was late, is $p_{X|Y}(1|0) = 0.53125$ and the probability that Sadie did the chore, given it was late, is $p_{X|Y}(0|0) = 0.46875$.

d. Two things are true with the given probabilities. First, if a chore was done late, it was more likely done by Charles than Sadie. Second, Sadie is more likely to do chores late than Charles. This explains why Charles may feel responsibility for the late chores: a higher proportion of late chores are Charles's chores. However, that is only because Charles does so many more chores to begin (the probability that Charles does a chore is 0.85).[14]

Unlike the multiplication rule and Bayes' theorem, the extension of the law of total probability is frequently cited when manipulating probability mass functions. It is a process which is important enough so as to warrant its own name: marginalization.

Marginalization

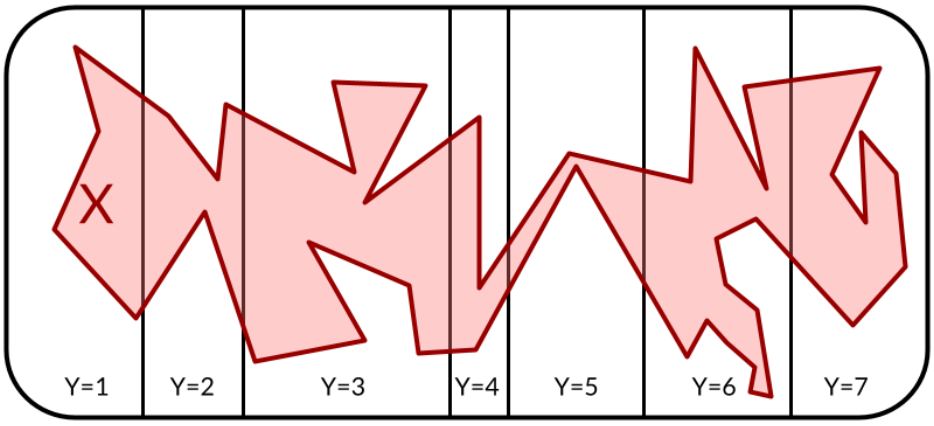
The idea with marginalization is that we are going to take a joint probability mass function and *marginalize it*, turning it into a marginal distribution. This is analogous to the law of total probability.

When dealing with a random variable, Y, there is some set of numeric values that Y can take on, namely the support of Y, which we denote $\mathcal{y}$. A natural partition of the space is to then take each possible value for Y as the event, and simply enumerate through the elements of $\mathcal{y}$.

Using this partition, we can ask about the ways that X can take on any particular value. In order for X to be some specific value, x', we know that Y must be one of the values, $y' \in \mathcal{y}$. Thus, if we add up all the possible combinations, $(X = x', Y = y_1)$, $(X = x', Y = y_2)$, and so forth, we will have covered every possible way of making $X = x'$. This is the exact same process we used when looking at contingency tables, where we summed a row or column to get the marginal probabilities.

[14] This is another example of *confusion of the inverse*.

Figure 5.1: This graphic shows the partitioning of the sample space using a random variable. It is entirely analogous to Figure 4.1, where in place of many different events, we partition over the set of values for Y. Here, $\mathcal{Y} = \{1, 2, 3, 4, 5, 6, 7\}$.



Taking this argument and encoding it with mathematical notation we get that

$$P(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \implies p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y).$$

The process of marginalization involves summing over one of the random variables in a joint distribution, leaving behind only the marginal of the other. This is often a very effective way of determining a marginal distribution when information about two random variables is easier to discern than information about only one.
To complete the analogy to the law of total probability, recall that the multiplication rule tells us that $p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$, and so we may also marginalize by taking

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x|y)p_Y(y).$$

This makes it clear that marginalization is often accomplished via arguments based on conditioning.

When confronted with questions from statistics and probability, it will often be the case that the natural answer to the question is "it depends." For instance, if asked "what is the probability that a student passes their next exam?" the likely response is "it depends." One very useful

technique for solving these questions in a satisfactory way is to continue that line of thought and explicitly specify "on what." In the previous example, you may say "it depends on how much they study." The conceit here is that, if you knew how much the student studied, then you would better understand the outcomes of the student's exam.

In mathematical terms this means you have a firm belief about the conditional distribution between the random quantities of *exam performance* and *study time*. The process of marginalization, and the law of total probability before that, provide useful ways of being able to translate these "it depends" statements into concrete beliefs about the marginal probabilities. Recall that marginal distributions are distributions which do not depend on any other quantity, and instead, thy capture the overall behaviour of a random variable. They have, in some sense, averaged out all other factors and give you the beliefs which do not depend on anything else at all. The technique for accomplishing this is marginalization.[15]

Example 5.15 (Photos and Bird Sightings on their Own)**.** Recall that when Charles and Sadie go on their bird watching adventures, Charles takes photos hoping for as many good ones as possible, and Sadie spots the birds. We take X to be the number of good photos taken on these trips, and Y to be the number of rare birds that are seen. We noted that

$$p_{X,Y}(x,y) = \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!},$$

with $0 \leq x \leq 10y$, and $y \geq 0$.

a. Write down an expression for the marginal probability mass function of Y.
b. Write down an expression for the marginal probability mass function of X.
c. **Challenge:** solve for the marginal probability mass function of Y.

Solution

a. To find $p_Y(y)$, we marginalize over X. That is, we sum the joint probability distribution function over all values for X. This gives

$$p_Y(y) = \sum_{x=0}^{10y} p_{X,Y}(x,y) = \sum_{x=0}^{10y} \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!}.$$

b. To find $p_X(x)$, we marginalize over Y. That is, we sum the joint probability distribution function over all values for Y. This gives

$$p_X(x) = \sum_{y=0}^{\infty} p_{X,Y}(x,y) = \sum_{y=0}^{\infty} \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!}.$$

[15] We shall see other forms of these "conditioning arguments" in the next chapter, where we try to summarize the behaviour of a random variable.

c. Note, that you will generally not be expected to manipulate these types of sums, however, it is possible to do. First, for $p_Y(y)$,

$$\begin{aligned} p_Y(y) &= \sum_{x=0}^{10y} \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x} \times \frac{e^{-3}3^y}{y!} \\ &= \frac{e^{-3}3^y}{y!} \underbrace{\sum_{x=0}^{10y} \binom{10y}{x} \times (0.25)^x \times (0.75)^{10y-x}}_{=1} \\ &= \frac{e^{-3}3^y}{y!}. \end{aligned}$$

5.7 Independent and Identically Distributed: A Framework for Interpretation

A very common assumption when addressing questions in statistics and in probability is that we have a set of random variables which are independent and identically distributed (often written iid). We now have the tools to understand concretely what this means.

Definition 5.11 (Independent and Identically Distributed (iid))**.** A set of random variables, $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributed (denoted iid) whenever:

i. every subset of random variables in the collection is independent of every other subset of random variables in the collection; and
ii. the marginal distribution for each of the random variables are exactly the same.

The assumption of iid random quantities will often come up when we are repeating a process many times over, and thinking about what observations will arise from this. Suppose that X_1 is a random variable that takes the value 1 if a flipped coin comes up heads, and 0 otherwise. If we imagine flipping this coin 100 times then it is reasonable to assume that each sequential coin flip will be independent of each other coin flip, since the result on one flip of a coin should not influence the result of any other flip of a coin. Moreover, every time the coin is flipped, it is reasonable to assume that the probability it shows up heads remains the same. As a result, these 100 random quantities, $(X_1, X_2, \ldots, X_{100})$, are said to be iid.

Example 5.16 (Reframing the Bird Photos)**.** Charles and Sadie had, to this point, been considering their bird photo taking adventures in terms of the joint and conditional distributions. After learning of independent and identically distributed random variables, Charles thinks that there is another way to frame the setup, and describes the following to Sadie:

Suppose it is known that there will be y birds seen on any given day. Then, the total number of good pictures of birds can be viewed as the sum of the number of good pictures of each bird. Thus, we can take a set of iid random variables, say $X_1, \dots, X_y$ which represent the number of good pictures taken of each bird.

a. Is this an accurate description? In reality, do we think that this will hold?
b. Supposing that this description is accurate, and that the probability mass function for each X_j is given by

$$p_X(x) = \binom{10}{x}(0.25)^x(0.75)^{10-x},$$

what is the joint probability mass function of $(X_1, \dots, X_y)$?

Solution

a. Yes, this is a reasonable explanation. In this case, if Charles is willing to assume that good photos for one bird do not impact the chances of good photos of another bird, then this is a perfectly valid way to frame the setting. Note that in this case, $Y = y$ is still random, so this whole description is conditioning on knowing $Y = y$. In the real world, it is likely that photos may not actually be independent. If Charles takes some really bad photos of one bird, that may impact the chances of taking good photos of the next. However, it is likely close enough to correct in order to be true.

b. Recall that for independent random variables, the joint probability mass function is given by the product of the marginal probability mass functions. To simplify the notation, we will define $T(X) = \sum_{i=1}^{y} x_i$. Thus the probability mass function of $X_1, \dots, X_y$ is given by

$$\begin{aligned}
p_{X_1,X_2,\dots,X_y}(x_1, \dots, x_y) &= \prod_{i=1}^{y} \binom{10}{x_i}(0.25)^{x_i}(0.75)^{10-x_i} \\
&= \left(\prod_{i=1}^{y} \binom{10}{x_i}\right) \times (0.25)^{\sum_{i=1}^{y} x_i}(0.75)^{10y-\sum_{i=1}^{y} x_i} \\
&= \left(\prod_{i=1}^{y} \frac{10!}{x_i!(10-x_i)!}\right) \times 0.25^{T(X)} \times 0.75^{10y-T(X)} \\
&= \left(\frac{(10!)^y}{\prod_{i=1}^{y} x_i!(10-x_i)!}\right) \times 0.25^{T(X)} \times 0.75^{10y-T(X)}.
\end{aligned}$$

While we will use the assumption of iid random variables explicitly at a later point.The iid assumption also provides an intuitive method for interpreting probability functions and distributions.

Suppose that we were to take a probability mass function, $p_X(x)$. If we were able to generate independent and identically distributed realizations from this probability mass function, then the function $p_X(x)$ describes the behaviour for these repeated realizations. Specifically, $p_X(x)$ will give the long-run proportion of realizations of the iid random variables which take the value x.

This type of statement is always the flavour of interpretation statements that are made with respect to probability and statistics. It is always be the case that, in order to understand what is meant by a statement of probability, we consider the repetition of some statistical experiment over and over again. When we were discussing sample spaces and experiments directly, we talked about repeating the experiment over and over again. When we begin to work with random variables instead, it becomes more natural to think about the replication procedures coming through the use of independent and identically distributed random variables.

As your study of probability progresses, you will begin to work with random quantities in a strictly theoretical sense. In introductory level problems, we are often holding in mind very concrete examples to illustrate the procedures and concepts. In this setting it is easy enough to hold in mind the experiment of interest: for instance, we may have a random variable representing the result of a coin toss, and you can envision repeatedly tossing a coin. As the concepts become less concrete, more abstract, and harder to draw direct parallels to tangible scenarios, it becomes more and more important to rely on the interpretations rooted in a series of independent and identically distributed random variables.

A large component of statistics as an area of study is making explicit the assumptions we are working with, and doing our best to ensure that these are reasonable. By interpreting probability mass functions as the proportion of independent and identically distributed random variables that take on a particular value when we repeatedly take realizations of these random variables, this philosophy is made clear and explicit.

Exercises

Exercise 5.1. Suppose that you sit in the library, observing the front desk until a patron takes out books. Describe at least 5 different random variables that could correspond to this experiment.

Exercise 5.2. Consider the data collection that goes on at a weather station. Describe as many different random variables as you can think of corresponding to this experiment.

Exercise 5.3. For each of the following random variables, identify whether they are discrete or continuous.

a. In a survey, the number of siblings each participant has is recorded.
b. A thermometer measures temperature in degrees Celsius.

c. The number of cars passing through a toll booth in an hour.
d. In a casino game, the amount of money won or lost on a single bet.
e. The weight of apples harvested from an orchard.
f. In a classroom, the number of students who own a laptop.
g. The time it takes for a light bulb to burn out.
h. When flipping a coin, the number of heads obtained in 10 flips.
i. The height of students in a class.
j. The number of emails received per day.
k. In a factory, the number of defective products in a batch.
l. The distance traveled by a car in a specific time interval.
m. When selecting a card from a standard deck, the card's face value.
n. The age at which individuals first learn to ride a bicycle.
o. The number of customers entering a store in one hour.
p. In a soccer match, the time it takes for the first goal to be scored.
q. The number of books a person reads in a month.
r. When rolling two dice, the sum of the numbers rolled.
s. The number of text messages sent in a day.
t. The volume of water in a reservoir.

Exercise 5.4. Suppose that a probability mass function is given by

$$p(x) = \begin{cases} k(7x+3) & x = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}.$$

Find the value of k.

Exercise 5.5. Consider a probability mass function given by

$$p(x) = \begin{cases} z & x = 1 \\ \frac{1}{x^2} & x = 2, 3, 4, 5, 6, 7, 8, 9, 10 \\ 0 & \text{otherwise} \end{cases}.$$

a. Find z.
b. What is the probability that X is an even number?

Exercise 5.6. Consider the probability mass function, defined for all non-negative integers, given by

$$p(x) = \frac{2^x e^{-2}}{x!}.$$

a. What is $P(X = 0)$?
b. What is $P(X = 5)$?
c. What is $P(X \geq 2)$?

Exercise 5.7. For each of the following, indicate and explain whether the following properties could belong to a valid probability mass function.

a. $p(x) < 0$ for some $x \in \mathbb{Z}$.
b. $p(x) > 1$ for some $x \in \mathbb{Z}$.
c. $p(x) > \frac{\pi}{\ell}$ for all elements of an ℓ element set.
d. $p(-|x|) > 0$ for all $x \in \mathbb{Z}$.

Exercise 5.8. Suppose that, for some fixed integer, y, you define the mathematical function $p(x) = \delta_y = I(x = y)$. That is, it takes a value of 1 if $x = y$ and 0 otherwise. Is this a valid probability mass function? Why?

Exercise 5.9. Consider the joint probability mass function of two random variables, X and Y, given by

$$P(X = x, Y = y) = \frac{1}{150}(x + y), 1 \leq x, y \leq 5.$$

a. Show that this is a valid joint probability mass function.
b. What is $P(X = 2, Y = 3)$?
c. What is the $P(X = 4)$?
d. What is the marginal probability mass function of Y?
e. Are X and Y independent? Explain.
f. What is the conditional probability mass function of X given Y?
g. What is the conditional probability mass function of Y given X?

Exercise 5.10. Consider the joint probability mass function of two random variables, X and Y, given by

$$P(X = x, Y = y) = \frac{1}{12}(y - x)^2, 1 \leq x, y \leq 3.$$

a. Show that this is a valid joint probability mass function.
b. What is $P(X = 1, Y = 2)$?
c. What is the $P(X = 2)$?
d. What is the marginal probability mass function of Y?
e. Are X and Y independent? Explain.
f. What is the conditional probability mass function of X given Y?
g. What is the conditional probability mass function of Y given X?

Exercise 5.11. Consider the following joint probability mass function represented as a contingency table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.1	0.2	0.3
$X = 2$	0.2	0.1	0.1

a. What is the probability that $X = 1$ and $Y = 2$?
b. Calculate $P(X = 2)$.
c. Find the marginal probability mass function of Y.
d. Are X and Y independent? Justify your answer.
e. What is $P(Y = 1|X = 2)$?
f. What is $P(X = 2|Y = 2)$?

Exercise 5.12. Suppose that a particular disease is associated with two common types of genetic mutations, say type A and type B. Let A and B correspond to the random variables which count the locations at which each type of mutation has occurred. In order for a type B mutation to occur, a type A must have also occurred at the same location, and so we can say that

$$P(B = b|A = a) = \binom{a}{b}(0.25)^b(0.75)^{a-b}, \quad b \in \{0, 1, 2, \dots, a\}.$$

Moreover, suppose that

$$P(A = a) = \frac{10 - a}{45} \quad a \in \{0, 1, 2, 3, 4, 5\}.$$

a. What is the joint probability mass function of A and B?
b. What is the marginal probability mass function of B?
c. What is the conditional probability mass function of A given B?
d. Is B independent of A? Explain.
e. Is A independent of B? Explain.
f. Suppose that you are at a substantially increased risk of the disease if the total, $A+B \geq 8$. What is the probability that an individual is at an increased risk?

Exercise 5.13. Suppose a factory produces two types of products: Widgets and Gadgets. Let W and G represent the random variables denoting the number of units produced for each type, in a particular hour. Suppose that the following is observed as the joint probability mass function

$$P(W = w, G = g) = \frac{1}{80}, \quad w \in \{1, 2, 3, \dots, 8\}, g \in \{1, 2, 3, \dots, 10\}.$$

a. What is the marginal probability mass function, $P(W)$?
b. What is the marginal probability mass function, $P(G)$?
c. Is $W \perp G$?
d. Let $T = W + G$. What is the probability mass function of T?
e. What is the conditional probability mass function, $P(W|T)$?
f. What is the conditional probability mass function, $P(G|T)$?
g. What is the conditional probability mass function, $P(T|G)$?
h. What is the conditional probability mass function, $P(T|W)$?

Exercise 5.14.

a. List an example of a real-world scenario where a set of random variables is likely to be iid. Explain.
b. List an example of a real-world scenario where a set of random variables is likely to be identically distributed, but not independent. Explain.
c. List an example of a real-world scenario where a set of random variables is likely to be independent, but not identically distributed. Explain.

Exercise 5.15. Suppose that X and Y are iid, with $p_X(x)$ and $p_Y(y)$. If Y takes on a specific value, y', with $p_Y(y') = 0.25$, what is the probability that X takes on the same value? Explain.

Exercise 5.16. Suppose that $X_1, \ldots, X_n$ are iid with a probability mass function, $p_X(x)$.

a. What is the joint probability mass function of $(X_1, \ldots, X_n)$?
b. What is the conditional probability mass function, $P(X_1 = x | X_2 = y)$?
c. What is the probability mass function of X_1 conditional on $\sum_{i=2}^{n} X_i$?

6 The Expected Value, Location Summaries, and Measures of Variability

6.1 Summarizing the Location of a Distribution

Until this point in our discussions of probability we have relied upon characterizing the behaviour of a random variable via the use of probability mass functions. In some sense, a probability mass function captures all of the probabilistic behaviour of a discrete random variable. Using the mass function you are able to characterize how often, in the long run, any particular value will be observed, and answer any questions associated with this. As a result, the mass function remains a critical area of focus for understanding how random quantities behave.

However, these functions need to be explored and manipulated in order for useful information to be extracted from them. They do not summarize this behaviour effectively, as they are not intended to be a summary tool. We may wish to have numeric quantities which are able to *concisely* express the behaviour of a distribution. Put differently, provided with a probability mass function it is hard to immediately answer "what do we expect to happen with this random variable?" despite the fact that this is a very obvious first question.

To address questions related to our expectations, we turn towards the statistical concept of central tendency or location.

Definition 6.1 (Location (Central Tendency))**.** The location of a distribution is a measure of a typical value, or a central value, for that distribution. A measure of location may also be referred to as a measure of central tendency.

With measures of location we are trying to capture, with one single number, what value is expected when we make observations from the random quantity. There are many ways one might think to describe our expectations, and it is worth exploring these concepts in some detail. In particular, we want to explore how we might think to define the **expected value** of a distribution.[1]

[1] As a general rule when learning it is often helpful to consider exercises of discovery. That is, try to determine *why* particular definitions are the way that they are, or where they have from.

6.2 Deriving the Expected Value

6.2.1 The Mode

One way that we may think to define our expected value is by asking what value is the most probable. This is a question which can be directly answered using the probability mass function. The process for this requires looking at the function and determining which value for x corresponds to the highest probability. This is the value that we are most likely to see. Sometimes this procedure is fairly straightforward, sometimes it is quite complicated. Regardless of the complexity of the specific scenario, the most probable value has a straightforward interpretation. As intuitive as this may seem, this is not the value that will be used as *the* expected value generally. Instead, this quantity is referred to as the mode.

Definition 6.2 (Mode)**.** The mode of a distribution is the most probable value of that distribution. Specifically, if X is a discrete random variable with mass function $p_X(x)$, then the mode is the value of x such that $p_X(x)$ is maximized.

Example 6.1 (Charles and Sadie Investigate Cupcake Sprinkles)**.** Charles and Sadie have noticed that their favourite coffee shop has been experiencing less foot traffic ever since *Probability Patisserie: Cupcake Conundrums* has opened up next door.[2] This cupcake shop has been attracting many of the regular customers, though, Charles and Sadie think that it has more to do with fancy marketing than with a better product. To this end, they go into the store and workout the probability mass function for X, the number of sprinkles on top of the cupcakes. They find the following

$$p_X(x) = \begin{cases} 0.25 & x = 0 \\ 0.3 & x = 1 \\ 0.1 & x = 2 \\ 0.05 & x \in \{3, 4, 5, 6, 7, 8, 9\} \\ 0 & \text{otherwise.} \end{cases}$$

What is the mode for the number of sprinkles on the cupcakes?

Solution

The mode is the value x such that $p_X(x)$ is a maximum. In this case the maximum occurs when $x = 1$, with $p_X(1) = 0.3$, and so the mode of the number of sprinkles is 1.

[2] You should picture the owner of this fictitious cupcake store as a giant, multinational, faceless corporation. A corporation that seems to take pleasure in stepping on the local businesses. That way, as Charles and Sadie try to fight back, we can remain on their side!

While the mode is a useful quantity, and for some decisions will be the most relevant summary value, there are some major issues with it as a general measure of location. For starters, consider that our most common probability model considered until this point has been that of equally likely outcomes. Here, there is no well-defined mode.[3] While the case of equally likely outcomes is a fairly strong example highlighting the issues with the mode, it need not be so dramatic to undermine its utility. It is possible for a distribution to have several modes which are quite distinct from one another, even if it's not all values in the support.

Moreover, it is quite common for the modal value to be not particularly likely itself. Consider a random variable that can take on a million different values. If all of the probabilities are approximately 0.000001 then presenting the mode as the most probable value does not translate to saying that the mode is particularly probable.

Example 6.2 (Charles and Sadie Investigate Cupcake Happiness)**.** After realizing that the cupcake shop did not provide very many sprinkles, Charles and Sadie decided to turn to the marketing material. In it, *Cupcake Conundrums* claims that, on a scale from 0 to 100,000 happiness points, most of their customers experience the maximum happiness after eating their cupcakes. Charles and Sadie are well aware of the ways in which these types of reports can be misleading, and so they investigate, finding the following probability mass function for the number of happiness points.

$$p_X(x) = \begin{cases} \frac{x}{5{,}000{,}050{,}000} & x \in \{1, \dots, 100000\} \\ 0 & \text{otherwise.} \end{cases}$$

a. Is the mode reported by the company correct?
b. Is the mode an accurate depiction of the distribution in this setting?

Solution

a. Yes. Note that $p_X(x)$ is increasing in x. As a result, the mode of the distribution is the highest value that the distribution can take on, which in this case is $x = 100,000$.

b. The mode is likely not a good indication of the distribution. Note that the maximal value happens only 100000/5000050000 of the time, which is a very very small value. If you look at the probability that a customer is only halfway happy, this is 50000/5000050000. This value is half the likelihood of the modal value, but the difference between these two probabilities is also only $50000/5000050000 \approx 0.000001$. That is, it is only 1 in 100,000 more likely to observe perfect happiness compared with half happiness. By extension, it is only about 2 in 100,000 more likely to observe perfect happiness compared with no happiness. The mode simply is not a good indicator of behaviour in this setting.

[3] When multiple modes exist, convention tends to report the set of all of the modes. In the equally likely outcome model, this is the entire support for the random variable, and as a result, the mode is exactly the probability mass function. There is no summary provided.

6.2.2 The Median

If the mode has these shortcomings, what else might work? Another intuitive concept is to try to select the "middle" of the distribution. One way to define the middle would be to select the value such that, in the long run, half of observations from the distribution are beneath it and half are above it. That way, when you are told this value, you immediately know that it is equally likely to observe values on either side of this mark. This is also a particularly intuitive definition for expected value, and is important enough to be named, the median.

Definition 6.3 (Median)**.** The median of a distribution is the value m which attempts to have $P(X \leq m) = 0.5$ and $P(X \geq m) = 0.5$. This value may not always exist, and so formally for discrete random variables, the median m is any value such that $P(X \leq m) \geq 0.5$ and $P(X \geq m) \geq 0.5$.

The median is the midpoint of a distribution and is very important for describing the behaviour of random variables. Medians are often the most helpful single value to report to indicate the typical behaviour of a distribution, and they are frequently used. When people interpret averages it is often the median that they are actually interpreting. It is very intuitive to be given a value and know that half of all realizations are above that point, and half of all realizations are below that point.

Example 6.3 (Charles and Sadie Investigate Cupcake Sprinkles (Again))**.** After seeing how the cupcake shop used the mode to misrepresent the happiness of its customers, Charles and Sadie worry that they may have been unfair with using the mode. As a result, they turn back to the cupcake sprinkle distribution and try to summarize it differently. Recall that the probability mass function for X, the number of sprinkles on top of the cupcakes, is

$$p_X(x) = \begin{cases} 0.25 & x = 0 \\ 0.3 & x = 1 \\ 0.1 & x = 2 \\ 0.05 & x \in \{3, 4, 5, 6, 7, 8, 9\} \\ 0 & \text{otherwise.} \end{cases}$$

What is the median number of sprinkles on the cupcakes?

Solution

Note that we are looking for a value, m, such that $P(X \geq m) \geq 0.5$ and $P(X \leq m) \geq 0.5$. To find this we can consider the cumulative probability sums. We know that $P(X \leq k) = \sum_{x=0}^{k} p_X(x)$, and as a result, stopping this value at the first time that the sum goes beyond 0.5 satisfies the second condition. In this case our cumulative sums are 0.25 then 0.55. As a result, $P(X \leq 1) = 0.55 > 0.5$. If we check $P(X \geq 1) = 1 - P(X <$

1) = 1 − P(X = 0) = 0.75 > 0.5. As a result, m = 1 satisfies the required conditions and is a median.

Despite the advantages of medians, they have their own drawbacks. For starters, the median can be exceptionally challenging to compute in certain settings. As a result, even when a median is appropriate, it may not be desirable if it is too challenging to determine.

Beyond the difficulties in computation, medians have a feature which is simultaneously a major benefit and a major drawback. Specifically, medians are less influenced by extreme values in the probability distribution. Consider two different distributions. The first is equally likely to take any value between 1 and 10. The second is equally likely to take any value between 1 and 9 or 1,000,000. In both of these settings, the median is 5 since $P(X \leq 5) = 0.5$ and $P(X \geq 5) = 0.5$. However, in the second setting we may observe a value as high as 1,000,000. Moreover, this value will be observed as often as the median will be.

The median, in some sense, ignores the extreme value in the probability distribution. In certain settings, this can be very desirable.[4] Consider the distribution of household incomes. There are a few households that earn an incredibly large amount, compared to the remaining households. If you are interested in understanding the "average household", the median may be a more appropriate measure, as those households with extreme incomes would otherwise distort the picture provided by most families. In this sense, the median's robustness to extreme values is a positive feature of it in terms of a summary measure for distributional behaviour.

Suppose instead that you work for an insurance company and are concerned with understanding the value of insurance claims that your company will need to pay out. The distribution will look quite similar to the income distribution. Most of the probability will be assigned to fairly small claims, with a small chance of a very large one. As an insurance company, if you use the median this large claim behaviour will be smoothed over, perhaps leaving you unprepared for the possibility of extremely large payouts. In this setting, the extreme values are informative and important, and as a result the median's robustness becomes a hindrance to correctly describing the important behaviour.[5]

Example 6.4 (Charles and Sadie Investigate Cupcake Public Relations)**.** Charles and Sadie feel as though they may be finally catching a break when word gets around that the cupcake shop was using spoiled ingredients, making the patrons sick! The cupcake shop, in hearing this, sent their massive public relations team into damage control mode. They claimed that the median number of illnesses per week associated with the cupcake store was the same as

[4]You will find many introductory sources that say that you should always use the median when you are concerned with extreme values. This is a great oversimplification of the truth, and points to a general rule in Statistics: there are very few general rules. The guidance comes from the fact that often extreme values skew our perceptions of the underlying truth. While this may be true in general, it is not true often enough to warrant being given as universal guidance.

[5]An insurance company ignoring these massive claims would almost certainly go out of business very, very quickly.

most local food services businesses. Doing some digging, Charles and Sadie find the following two probability mass functions, for X and Y, where X is the number of weekly illnesses at the cupcake shop, and Y is the number from a local business that has been around for a while.

$$p_X(x) = \begin{cases} 0.1 & x = 0 \\ 0.1 & x = 1 \\ 0.35 & x = 2 \\ 0.05 & x = 3 \\ 0.4 & x = 25 \\ 0 & \text{otherwise;} \end{cases}$$

$$p_Y(y) = \begin{cases} 0.45 & x = 0 \\ 0.04 & x = 1 \\ 0.5 & x = 2 \\ 0.01 & x = 3 \end{cases}.$$

a. What is the median of X?
b. What is the median of Y?
c. Is the claim made by the public relations team an accurate depiction of the world? Why or why not?

Solution

a. We can consider cumulative sums here again. In order we get, 0.1, 0.2, 0.55. This suggests that $x = 2$ is a candidate. If we check $P(X \geq 2) = 1 - 0.2 = 0.8$, and so $x = 2$ is a median.

b. Using cumulative sums we get 0.45, 0.49, 0.99. Again, this gives $x = 2$ as a candidate. If we check $P(X \geq 2) = 1 - 0.49 = 0.51$, and so $x = 2$ is a median.

c. The claim is accurate in that the median number for both establishments in 2. However, here the median is a bad representation for X. The probability that more than 2 cases occur in a week at the cupcake shop is 0.8, compared with 0.51 at the local business. What's more, the probability that X is as high as 25 is 0.4, where Y is *never* higher than 3.

Between the median and the mode we have two measures which capture some sense of expected value, each with their own set of strengths and drawbacks. Neither capture what it is that is referred to as *the* expected value. For this, we need to take inspiration from the median, and consider another way that we may think to find the center of the distribution.

6.2.3 The Mean

If the median gives the middle reading along the values sequentially, we may also wish to think about trying to find the *center of gravity* of the numbers. Suppose you take a pen, or marker, or small box of chocolates, and you wish to balance this object on a finger or an arm. To do so, you do not place the item so that half of its length sits on one side of the appendage and half on the other. You adjust the location so that half of the *mass* sits on either side of the appendage.

Throughout our discussion of discrete random variables we have referred to probability as *mass*. We use the *probability mass function* to generate our probability values. This metaphor can be extended when we try to find the center of the distribution. If we imagine placing a mass with weight equal to the probability mass at each value that a random variable can take on, we may ask, "where would we have to place a fulcrum to have this number line be balanced?" The answer to this question serves as another possible measure of center. It turns out that this notion of center is the one that we are all most familiar with, the simple average, or mean.

Definition 6.4 (Mean)**.** The mean of a distribution is the center of mass of the distribution. For a random variable, X with support $\mathcal{X}$, and probability mass function $p_X(x)$, the mean of X is given by

$$\sum_{x \in \mathcal{X}} x p_X(x).$$

It is this measure of location which ends up being called the **expected value** in statistics. We will use average, expected value, mean, and expectation interchangeably. In terms of notation, the expected value of a random variable X is denoted $E[X]$. Mathematically, the expected value is desirable for many reasons, some of which we will study in more depth later on. One of these desirable features, which stands in contrast with the median, is the comparative ease with which expected values can be computed. The summation for the expected value is easy to write down, and typically can be solved (either analytically, or readily with a computer).

Example 6.5 (Charles and Sadie Investigate Cupcake Sprinkles (One Last Time))**.** While the median and mode number of sprinkles on the cupcakes were the same, Charles and Sadie realize that this does not end up connecting well with the total number of sprinkles given out. Perhaps customers like the possibility of getting a large number of sprinkles, if they get lucky. As a result, they decide to round out the summary of the distribution by considering the mean number of sprinkles as well. Recall that the probability mass function for X, the number of sprinkles on top of the cupcakes, is

$$p_X(x) = \begin{cases} 0.25 & x = 0 \\ 0.3 & x = 1 \\ 0.1 & x = 2 \\ 0.05 & x \in \{3, 4, 5, 6, 7, 8, 9\} \\ 0 & \text{otherwise.} \end{cases}$$

What is the mean number of sprinkles?

Solution

Recall that the mean of a distribution is given by $E[X] = \sum_{x \in \mathcal{X}} x P_X(x)$. In this case, this results in

$$\begin{aligned}
&\sum_{x=0}^{9} x p_X(x) \\
&= 0(0.25) + 1(0.3) + 2(0.1) + (3+4+5+6+7+8+9)(0.05) \\
&= 2.6.
\end{aligned}$$

As a result, the mean number of sprinkles used is 2.6. This is higher than the median and mode, but is still not particularly high.

In the case of an equally likely probability model, the expected value becomes the standard average that is widely used. Suppose that there are n options in the support with $\mathcal{X} = \{x_1, \ldots, x_n\}$. We can write

$$E[X] = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is the formula for the average that is most commonly applied. When the probability models are more complex, the formula is not precisely the standard average – instead, it becomes a weighted average, where the weights are the probabilities.[6] The frequency with which expected values are used make them attractive as a quick summary for the center of a distribution.

6.2.4 How is the Mean "Expected"?

While the mean provides a useful, intuitive measure of center of the distribution, it is perhaps counterintuitive to name it the "expected value." To understand the naming convention it is easiest to consider the application which has likely spurred more development of statistics and probability than any other: gambling.

Suppose that there is some game of chance that can pay out different amounts with different probabilities. A critical question for a gambler in deciding whether or not to play such a game is "how much can I expect to earn, if I play?" This is crucial to understanding, for instance, how much you should be willing to pay to participate, or if you are the one running the game, how much you should charge to ensure that you make a profit.

[6] While less commonly applied than the simple average, a weighted average is familiar to most students for a crucial purpose: grade calculations. If you view the weight of each item in a course as a probability mass, and the grade you scored as the value, then your final grade in the course is exactly the expected value of this distribution.

If you want to understand what you expect to earn, the intuitive way of accomplishing this is to weight each possible outcome by how likely it is to occur. This is exactly the expected value formula that has been provided, and so the expected value can be thought of as the expected payout of a game of chance where the outcomes are payouts corresponding to each probability.

Cost for a Game of Chance

Suppose that a game of chance is being run. It will cost $\$m$ to play, and the game will pay out values from $\mathcal{X}$, according to a random variable X with probability mass function $p_X(x)$. The question that we want to know is what should the value be for m, such that, in the long term, a player playing the game will earn no money?
Note that if $X = x$ then a player playing the game will earn $x - m$ for that play, and this will happen with probability $p_X(x)$. Thus, in the long run, in $p_X(x)$ proportion of games the player earns $x - m$. If the player plays N games, then as $N \to \infty$, the total contribution from these winnings will be $(x-m) \cdot N \cdot p_X(x)$. If we add up this contribution for all possible results, x, we get

$$\begin{aligned}\sum_{x\in\mathcal{X}} N(x-m)p_X(x) &= N\left[\sum_{x\in\mathcal{X}} xp_X(x) - \sum_{x\in\mathcal{X}} mp_X(x)\right] \\ &= N\left[\sum_{x\in\mathcal{X}} xp_X(x) - m\sum_{x\in\mathcal{X}} p_X(x)\right] \\ &= N\left[E[X] - m\right].\end{aligned}$$

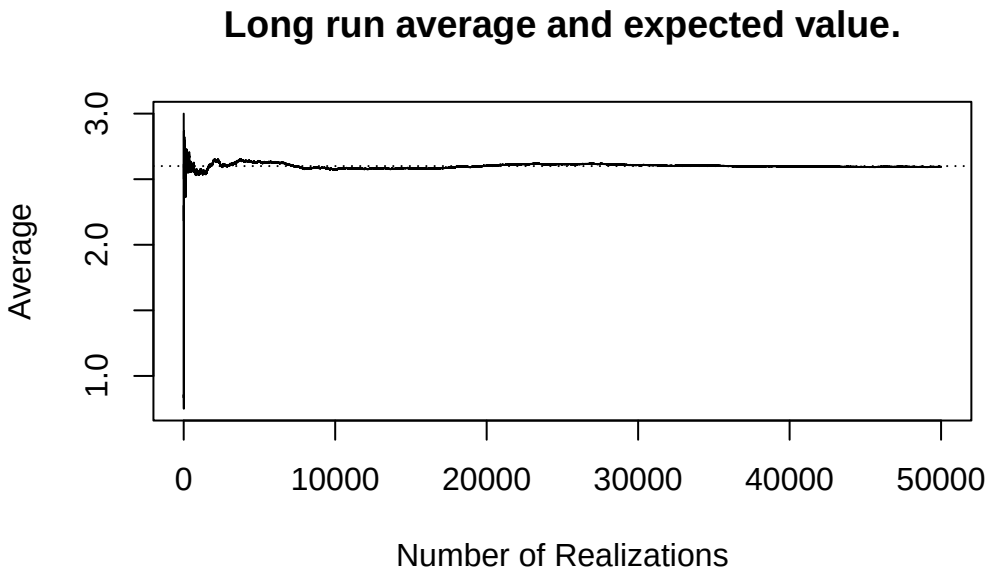
Thus, in the long run, the total that the player wins, as $N \to \infty$ will be $N(E[X] - m)$. Thus, in order for the long run earnings to be zero, m should be set to be equal to $E[X]$. If m is set to be 0, then per game the player *expects* to earn $E[X]$. This also represents the cost at which a rational actor should be willing to pay to participate. If a game of chance costs more than the expected value to play, in the long run you will lose money. If a game of chance costs less than the expected value, in the long run you will earn money. It is hard to overstate the utility of gambling in developing probability theory, and as such these types of connections are expected.

To interpret the expected value of a random variable, one possibility is using the intuition that we used to derive the result. Notably, the expected value is the center of mass of the distribution, where the masses correspond to probabilities. This means that it is not necessarily an actual central number over the range, but rather that it sits in the weighted middle. While this interpretation is useful in many situations, there are times where the point of balance is a less intuitive description. For these, it can sometimes be useful to frame the expected value as the long term simple average from the distribution.

If we imagine observing many independent and identically distributed random variables, then

as the number of samples tends to infinity, the expected value of X and the simple average will begin to coincide with one another. That is the distance between $E[X]$ and $\frac{1}{n}\sum_{i=1}^{n} X_i$ will shrink to 0. As a result, we can view the expected value as the average over repeated experiments. This interpretation coincides nicely with the description based on games of chance. Specifically, if you were to repeatedly play the same game of chance, the average payout per game will be equal to the expected value, if you play for long enough.

Figure 6.1: This plot produces 50000 realizations of the random variable described in Example 6.5. As computed in the example, the expected value is 2.6. As many repeated observations are averaged, the empirical average converges to this computed value.



6.3 Which Measure of Central Tendency Should be Used?

Where the median demonstrated robustness against extreme values in the distribution, the mean does not. For instance, if we consider the distribution of incomes across a particular region, the mean will be much higher than the median, since those families with exceptionally high incomes will not be smoothed over as they were with medians. In this case, the lack of robustness for the expected value will render the mean a less representative summary for the true behaviour of the random quantity.

To see this concretely consider a random variable which with equal probability takes a value between 1 and 9. This will have $E[X] = 5$. Now, if the 9 is made to be $1,000,000$, the expected

value will now be $E[X] = 111115.\dot{1}$. This is a far cry from the median which does not change from 5 in either case. This lack of robustness is desirable in the event of the insurance example from the median discussion, but will be less desirable in other settings.

The mean, median, and mode are the three standard measures of central tendency. They are single values which describe the standard behaviour of a random quantity. Each of the three has merits as a measure, and each has drawbacks for certain settings. The question of which to use and when depends primarily on the question of interest under consideration, rather than on features of the data alone.[7] Often, presenting more than one measure can give a better sense of the distributional behaviour that any one individual will.

Example 6.6 (Charles and Sadie Reflect on the Cupcake Adventures)**.** Charles and Sadie decide it is worth stepping back and summarizing all that has happened with regards to their cupcake adventures, trying to ensure that distributions are always summarized fairly.

a. For the number of sprinkles per cupcake is the mean, median, or mode the best measure of central tendency?
b. For the amount happiness in customers, is the mean, median, or mode the best measure of central tendency?
c. For the number of customers who become ill eating at establishments, is the mean, median, or mode the best measure of central tendency?

Solution

a. There is an argument that any of the measures here could work the best, it would depend on the individual's preferences. The mode makes a lot of sense as, if you are a regular customer, the mode is what you will experience day-to-day; most days you will have the modal number of sprinkles. The median may make sense as it provides a benchmark for measuring what constitutes a lot of sprinkles and a little. Half of the time you'll end up with a more sprinkled donut, half the time a less. The mean would be particularly interesting to note for the store owners themselves, since the mean is directly tied to totals: if the store knows that they sell 100 cupcakes per day, most days, then they also can expect to use 100 times the mean number of sprinkles in a day. This is helpful for planning.

b. The median is likely the most useful measure in this setting. The mode could be useful if happiness were not measured on a $100,000$ point scale. The mean is likely a less relevant value here as we are not particularly concerned with the total happiness, and if there was a large skew or major outliers in this distribution, we would want to determine where the majority fall. This is more analogous to the income example rather than the insurance example.

[7] The data is one consideration for which measure to use, but not the only one (and not the most important one).

c. In this case either the mode or mean are likely the best gauges. Here we do care about totals, and in particular, we do not want to smooth over outliers. It is very relevant if sometimes a lot of individuals become ill after eating an establishment, and the median would hide this information.

Despite the utility of all three measures, the expected value holds a place of more central importance in probability and statistics. A lot of this has to do with further mathematical properties of the mean. Because of its central role, it is worth studying the expected value in some more depth.

6.4 Expected Values of Functions of Random Variables

Sometimes the value of a random variable needs to be mapped through a function to give the value which is most relevant to us. Consider, for instance, a situation wherein the side lengths of boxes being manufactured by a specific supplier are random, due to incorrectly calibrated tolerances in the machines. The resulting boxes are perfect cubes. Suppose we are interested in the volume of the produced box not the side length. If a box has side length x, then its volume will be x^3, and so we may desire some way of computing $E[X^3]$ rather than $E[X]$.

Generally, for a function $g(X)$, we may want to compute $E[g(X)]$. It is important to recognize that $E[g(X)] \neq g(E[X])$. This is a common mistake.[8] If we are unable to apply the function to the expected value, then the question of how to compute the expected value remains. Instead of applying the function to overall expected value, instead, we apply the function to *each* value in the defining relationship for the expected value. That is,

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x).$$

This is sometimes referred to as the "law of the unconscious statistician," a name which may be aggressive enough to help remember the correct way to compute the expectation.[9]

The Law of the Unconscious Statistician

The law of the unconscious statistician (LOTUS) states that, for a random variable X, if we wish to find $E[g(X)]$, then we compute

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x).$$

[8] and an attractive one, but a mistake nonetheless.
[9] Some statisticians dislike this name. I find it to be rather cute.

Example 6.7 (The Happiness Scale Inversion)**.** Charles and Sadie have made really great strides working to protect their favourite coffee shop from the new cupcake store. One day when digging through the material more, they realize that the happiness report produced by the company is even less accurate than they had originally reported! The company reported the following probability mass function for happiness points

$$p_X(x) = \begin{cases} \frac{x}{5{,}000{,}050{,}000} & x \in \{1, \dots, 100000\} \\ 0 & \text{otherwise.} \end{cases}$$

Charles and Sadie track down the source of this expression and they find that, in fact, this does not measure happiness points at all. Instead, the number of happiness points is a function of X, specifically, $Z = 1/X$.

a. What is the expected value of X? Note, it may be helpful to recall that $\sum_{x=1}^{k} x^2 = \frac{k(k+1)(2k+1)}{6}$.
b. What is the expected value of Z?

Solution

a. Using directly the formula for $E[X]$ gives

$$\begin{aligned} E[X] = \sum_{x=1}^{100000} x p_X(x) &= \sum_{x=1}^{100000} x \frac{x}{5000050000} \\ &= \frac{1}{5000050000} \sum_{x=1}^{100000} x^2 \\ &= \frac{1}{5000050000} \cdot \frac{100000(100000+1)(2(100000)+1)}{6} \\ &= 66667. \end{aligned}$$

b. Applying the law of the unconscious statistician, gives

$$\begin{aligned} E[Z] = E[1/X] = \sum_{x=1}^{100000} \frac{1}{x} p_X(x) &= \sum_{x=1}^{100000} \frac{1}{x} \frac{x}{5000050000} \\ &= \frac{1}{5000050000} \sum_{x=1}^{100000} 1 \\ &= \frac{100000}{5000050000} \\ &= \frac{2}{100001}. \end{aligned}$$

These functions applied to random variables are often thought of as "transformations" of the random quantities. For instance, we *transformed* a side length into a volume. While the law of the unconscious statistician will apply to any transformation for a random variable, we can sometimes use shortcuts to circumvent its application. In particular, when $g(X) = aX + b$, for constant numbers a and b, we can greatly simplify the expected value of the transformation. To see this note

$$\begin{aligned} E[aX + b] &= \sum_{x \in \mathcal{X}} (ax + b) p_X(x) \\ &= \sum_{x \in \mathcal{X}} axp_X(x) + bp_X(x) \\ &= a \sum_{x \in \mathcal{X}} xp_X(x) + b \sum_{x \in \mathcal{X}} p_X(X) \\ &= aE[X] + b. \end{aligned}$$

That is, in general, we have that $E[aX + b] = aE[X] + b$. Note that part of the property of the linearity of expectation that we can immediately see if that the expected value of any constant is always that constant. If we take $a = 0$, then we see that $E[aX + b] = E[b] = b$. Thus, any time that we need to take the expected value of any constant number, we know that it is just that number.

This is particularly useful as linear transformations like $aX + b$ arise very commonly. For instance, most unit conversions are simple linear combinations. If a random quantity is measured in one unit then this result can be used to quickly convert expectations to another.

Example 6.8 (Sadie's Trip to America)**.** Sadie has recently returned from a long trip to America. The trip was long enough that temperatures measured in Fahrenheit started to make sense. When Sadie and Charles begin to talk about the weather, Charles brings up the temperature distribution of a possible summer vacation spot. Unfortunately for Sadie, these temperatures are all in Celsius. The distribution Charles provides is

$$p_X(x) = \begin{cases} 0.1 & x \in \{10, 11, 12, 13, 14\} \\ 0.05 & x \in \{15, 16, 17, 18, 19, 20, 21, 22, 23, 24\} \\ 0 & \text{otherwise} \end{cases}$$

a. What is the expected temperature, in Celsius?
b. Supposing that the temperature in Fahrenheit is given by $Y = 1.8X + 32$, what is the expected temperature in Fahrenheit?

Solution

a. Here we find

$$\begin{aligned}\sum_{x=10}^{24} x p_X(x) &= 0.1 \sum_{x=10}^{14} x + 0.05 \sum_{x=15}^{24} x \\ &= 15.75.\end{aligned}$$

b. Since $E[X] = 15.75$ then $E[Y] = E[1.8X + 32] = 1.8E[X] + 32 = 60.35$.

This type of linear transformation also frequently comes up with games of chance and payouts, or with scoring more generally.[10] Beyond being linear over simple transformations, summations in general behave nicely with expectations. Specifically, for any quantities separated by addition, say $g(X) + h(X)$, the expected value will be the sum of each expected value. Formally,

$$\begin{aligned}E[g(X) + h(X)] &= \sum_{x \in \mathcal{X}} (g(x) + h(x)) p_X(x) \\ &= \sum_{x \in \mathcal{X}} g(x) p_X(x) + h(x) p_X(x) \\ &= \sum_{x \in \mathcal{X}} g(x) p_X(x) + \sum_{x \in \mathcal{X}} h(x) p_X(x) \\ &= E[g(X)] + E[h(X)].\end{aligned}$$

Behaving well under linearity is one of the very nice properties of expectations. It will come in useful when dealing with a large variety of important quantities, and as we will see shortly, this linearity will also extend to multiple different random quantities.

Measures of central tendency are important to summarize the behaviour of a random quantity. Whether using the mean, median, or mode, these measures of location describe, on average, what to expect from observations of the random quantity. However, understanding a distribution requires understanding far more than simply the measures of location. As was discussed previously, the probability mass function captures the complete probabilistic behaviour of a discrete random variable, it is only intuitive that some information would be lost with a single numeric summary.

[10] For instance, suppose you are betting a certain amount on the results of a coin toss, or that you are taking a multiple choice test that gives 2 points for a correct answer.

Equal Location Across Different Distributions

Consider the following three distributions for three random variables, X, Y, and Z:

$$p_X(x) = \begin{cases} 0.25 & x = -1 \\ 0.5 & x = 0 \\ 0.25 & x = 1 \\ 0 & \text{otherwise.} \end{cases} \qquad p_Y(y) = \begin{cases} 0.05 & x = -5 \\ 0.1 & x = -4 \\ 0.1 & x = -3 \\ 0.1 & x = -2 \\ 0.1 & x = -1 \\ 0.15 & x = 0 \\ 0.1 & x = 1 \\ 0.1 & x = 2 \\ 0.1 & x = 3 \\ 0.1 & x = 4 \\ 0.05 & x = 5 \\ 0 & \text{otherwise.} \end{cases} \qquad p_Z(z) = \begin{cases} 0.25 & x = -400 \\ 0.5 & x = 0 \\ \frac{1}{3196} & x \in \{1, \ldots, 799\} \\ 0 & \text{otherwise.} \end{cases}.$$

In each of these distributions we have the mean, median, and mode all equalling 0.[11] However, even just from a quick glance, these distributions are all very, very differently behaved. The location summaries here clearly miss much of the important information about these different distributions. Spend some time trying to think of the ways in which they differ from one another, and see if you can determine *what* is missing from relying solely on measures of location.

6.5 Summarizing the Variability of a Random Variable

A key characteristic of the behaviour of a random variable which is not captured by the measures of location is the variability of the quantity. If we imagine taking repeated realizations of a random variable, the variability of the random variable captures how much movement there will be observation to observation. If a random variable has low variability, we expect that the various observations will cluster together, becoming not too distant from one another. If a random variable has high variability, we expect the observations to jump around each time.

[11] Try working this out!

6.6 The Range

Just as was the case with measures of location, there are several measures of variability which may be applicable in any given setting. One fairly basic measure of this variability is the range of possible values: what is the highest possible value, what is the lowest possible value, and how much distance is there between those two points?

Definition 6.5 (Range)**.** For a random variable, X, the range of the random variable is defined as $\text{Range}(X) = \max(X) - \min(X)$. That is, it is the distance between the maximum value that the random variable can take on, and the minimum value that the random variable can take on. Sometimes the range is reported with these outpoints explicitly specified.

This is a fairly intuitive notion, and is particularly useful in the equal probability model over a sequence of numbers. Consider dice. Dice are typically defined by the range of values that they occupy, say 1 to 6, or 1 to 20. Once you know the values present on any die, you have a sense for how much the values can move observation to observation.

Example 6.9 (Charles and Sadie Explore Ice Cream Flavors)**.** Charles and Sadie have decided to spend their sunny afternoon exploring various ice cream flavors at their local parlor, *Symmetric Scoops*. They notice that Scoops Galore offers a wide variety of flavors, from classic vanilla to exotic dragon fruit swirl. Intrigued by the selection, they decide to investigate the probability distribution of a random variable, Y, representing the number of unique flavors a customer selects.

After discreetly observing several customers, Charles and Sadie jot down their findings:

$$p_Y(y) = \begin{cases} 0.2 & y = 1 \\ 0.35 & y = 2 \\ 0.3 & y = 3 \\ 0.15 & y = 4 \\ 0 & \text{otherwise} \end{cases}$$

What is the range of the random variable Y, representing the number of unique ice cream flavors a customer selects at Scoops Galore?

Solution

Looking at the defining relationship for this random variable, we can see that the maximum value the distribution can take on is 4 (with probability 0.15) and the minimum value is 1 (with probability 0.2). As a result, we say that $\text{Range}(Y) = 4 - 1 = 3$.

While the range is an important measure to consider to determine the behaviour of a random variable, it is a fairly crude measurement. It may be the case that, while the extreme values are possible, they are sufficiently unlikely so as to come up very infrequently and not remain representative of the likely spread of observations. Alternatively, many random variables have a theoretically infinite range. In these cases, providing the range will likely not provide much utility.

Example 6.10 (Charles and Sadie Explore Ice Cream Flavors, Again)**.** Having had so much fun at *Symmetric Scoops* the first time, Charles and Sadie return once more to continue to investigate customers behaviour. When they arrive they notice a temporary promotion going on, which happens on one randomly selected Sunday a year, where customers can buy a *Mega Sunda(y)e Extravaganza*, a wonderfully extravagant creation that features scoops from all 50 flavours on offer. Working on this today they find the following probability mass function for Y.

$$p_Y(y) = \begin{cases} 0.2 & y = 1 \\ 0.35 & y = 2 \\ 0.3 & y = 3 \\ 0.145 & y = 4 \\ 0.005 & y = 50 \\ 0 & \text{otherwise} \end{cases}$$

What is the range of the random variable Y now? Does this accurately represent the dispersion we expect from Y?

Solution

Looking at the defining relationship for this random variable, we can see that the maximum value the distribution can take on is 50 (with probability 0.005) and the minimum value is 1 (with probability 0.2). As a result, we say that $\text{Range}(Y) = 50 - 1 = 49$. This does **not** accurately represent the dispersion of the random variable. Not only do very few customers order this when it is available, it is also almost never available. As a result, the maximum value (while possible) is not reflective of the usual behaviour of this random quantity, and this displays one of the concerns with the range as a measure of spread.

6.7 The Interquartile Range

To remedy these two issues, we may think of techniques to modify the range. Instead of taking the minimum and maximum possible values, we can instead consider ranges of values which remain more plausible. A common way to do this is to extend our concept of a median beyond the half-way point. The median of a random variable X, is the value, m, such that

$P(X \leq m) = 0.5$[12]. While there is good reason to care about the midpoint, we can think of generalizing this to be *any* probability.

That is, we could find a number z, such that $P(X \leq z) = 0.1$. We could then use this value to conclude that the probability of observing a value less than z is 10%. These values are referred to, generally, as **percentiles** and they are the natural extension of medians.

Definition 6.6. The $100p$th percentile (e.g., 70th percentile for $p = 0.7$, or 20th percentile for $p = 0.20$), is denoted $\zeta(p)$ and is the value such that $P(X \leq \zeta(p)) = p$. Thus, the median is given by $\zeta(0.5)$ and is also called the 50th percentile.[13]

Example 6.11 (Perplexed at the Ice Cream Parlour)**.** Charles and Sadie remained somewhat disappointed by their lack of ability to accurately capture the behaviour of a random variable using the range. To this end, they spend a lot more time at the ice cream parlor, and come up with, what they believe, is the correct probability mass function for how many flavours customers order, in total.

$$p_Y(y) = \begin{cases} 0.25 & y = 1 \\ 0.25 & y = 2 \\ 0.1 & y = 3 \\ 0.05 & y = 4 \\ 0.05 & y = 5 \\ 0.05 & y = 6 \\ 0.15 & y = 7 \\ 0.05 & y = 8 \\ 0.04 & y = 9 \\ 0.005 & y = 10 \\ 0.005 & y = 50 \\ 0 & \text{otherwise} \end{cases}$$

Using this probability mass function, find $\zeta(0.25)$, $\zeta(0.5)$, $\zeta(0.75)$, and $\zeta(0.99)$. What do each of these values mean?

Solution

a. $\zeta(0.25)$ is found by looking for $P(Y \leq \zeta(0.25)) = 0.25$. We note that $P(Y \leq 1) = P(Y = 1) = 0.25$, and so $\zeta(0.25) = 1$. This means that there is a 25% chance that a randomly selected individual will order 1 or fewer flavours.

[12] And, as a result, $P(X > m) = 0.5$ as well.

[13] Note that, when dealing with discrete random variables, it may not be possible to find a value $\zeta(p)$ such that $P(X \leq \zeta(p)) = p$ exactly. Instead, we typically define the percentile here to be such that $P(X < \zeta(p)) < p \leq P(X \leq \zeta(p))$.

b. $\zeta(0.5)$ is found by looking for $P(Y \leq \zeta(0.5)) = 0.25$. Note that $P(Y \leq 2) = P(Y = 1) + P(Y = 2) = 0.5$, and so $\zeta(0.5) = 2$. This means that 50% of customers order 2 or fewer flavours, and 50% of customers order more than 2 flavours.
c. $\zeta(0.75)$ is found by looking for $P(Y \leq \zeta(0.75)) = 0.75$. Note that we can continue the cumulative sums from the probability mass function. This gives, in order, $0.25, 0.5, 0.6, 0.65, 0.7, 0.75, 0.9, 0.95, 0.99, 0.995, 1$. As a result, $P(Y \leq 6) = 0.75$ and so $\zeta(0.75) = 6$. This means that 75% of customers order 7 or fewer flavours (the remaining 25% order more than 7).
d. $\zeta(0.99)$ can be found through the same cumulative sums as in (c). This is given by $y = 9$, such that $P(Y \leq 9) = 0.99$, so $\zeta(0.99) = 9$. This means that 99% of individuals order 9 or fewer flavours.

We can leverage percentiles to remedy some of the issues with the range as a measure of variability. Framed in terms of percentiles, the minimum value is $\zeta(0)$, and the maximum value is $\zeta(1)$. Instead of considering the extreme endpoints, we can consider the difference between more moderate percentiles. Doing so allows us to overcome the major concerns outlined with the range. If we take $\zeta(p_1)$ and $\zeta(p_2)$, for $p_1 < p_2$, then the difference between $\zeta(p_2) - \zeta(p_1)$ can be seen as a measure of variability, analogous to the range.

The most common choices would be to take $\zeta(0.25)$ and $\zeta(0.75)$, the 25th and 75th percentiles, respectively. These are also referred to as the first and third quartiles, respectively. They are named as, taking $\zeta(0.25)$, $\zeta(0.5)$ and $\zeta(0.75)$, the distribution is cut into quarters.

With the first and third quartiles computed, we can compute the **interquartile range**, which is given by $\zeta(0.75) - \zeta(0.25)$.

Definition 6.7 (Interquartile Range (IQR))**.** The interquartile range, or IQR, is defined as $\zeta(0.75) - \zeta(0.25)$, the difference between the third and first quartiles. It is a measure of spread, and is typically denoted as $\text{IQR} = Q3 - Q1$, where Q stands for quartiles.

Like the range, the IQR gives a measure of how much spread there tends to be in a distribution. Unlike the range, however, we can be more certain that both the first and third quartiles are reasonable values around which repeated observations of the random variable would be observed. Specifically, there is a probability of 0.5 that a value between the first and third quartile will be observed. The larger the IQR, the more spread out these moderate observations will be, and as a result, the more variable the distribution is.

Example 6.12 (Addressing the Ice Cream Perplexity)**.** Having understood many of the percentiles of the distribution of ice cream flavours ordered at *Symmetric Scoops*, Charles and Sadie decide to have another shot at capturing the spread. Recall that the probability mass

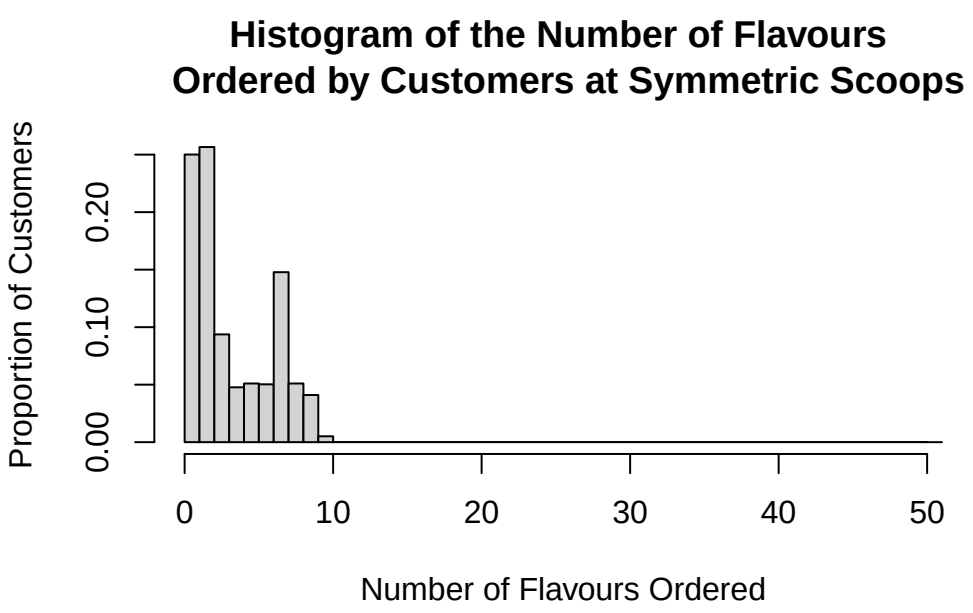
function for Y, the number of ice cream flavours ordered by a customer, is given by

$$p_Y(y) = \begin{cases} 0.25 & y = 1 \\ 0.25 & y = 2 \\ 0.1 & y = 3 \\ 0.05 & y = 4 \\ 0.05 & y = 5 \\ 0.05 & y = 6 \\ 0.15 & y = 7 \\ 0.05 & y = 8 \\ 0.04 & y = 9 \\ 0.005 & y = 10 \\ 0.005 & y = 50 \\ 0 & \text{otherwise} \end{cases}$$

What is the interquartile range for this population? How does this compare to the range?

Solution

We know that $\zeta(0.25) = 1$ and $\zeta(0.75) = 6$. As a result, the IQR $= 6 - 1 = 5$. By contrast, the range is given by $50 - 1 = 49$. It is incredibly rare to observe a customer ordering 50 flavours, as a result, using 5 is a better measure of spread of the distribution. To see this, we could consider plotting a histogram (more on these later on) that show the number of customers who order each type, and then intuitively as for a good measure of how *spread out* the data are.



6.8 The Variance and Mean Absolute Deviation

Both the range and the interquartile range give a sense of the variation in the distribution irrespective of the measures of location for that distribution. Another plausible method for assessing the variability of a distribution is to assess how far we expect observations to be from its center. Intuitively, if observations of X are near the center with high probability, then the distribution will be less variable than if the average distance to the center is larger.

This intuitive measure of variability is useful for capturing the behaviour of a random variable, particularly when paired with a measure of location. However, we do have to be careful: not all measures of dispersion based on this notion will be useful. Consider the most basic possibility, $X - E[X]$. We might ask, for instance, what is the expected value of this quantity. If we take $E[X - E[X]]$ then note that this a linear combination in expectation since $E[X]$ is just some number. Thus, $E[X - E[X]] = E[X] - E[X] = 0$. In other words, the expected difference between a random variable and its mean is exactly 0. We thus need to think harder about how best to turn this intuition into a useful measure of spread.

The issue with this procedure is that some realizations are going to be below the mean, making the difference negative, and some will be above the mean, making the difference positive. Our defining relationship for the mean relied on balancing these two sets of mass. However, when discussing the variability of the random variable, we do not much care whether the observations

are lower than expected or higher than expected, we simply care how much variability there is around what is expected. To remedy this, we should consider only the distance between the observation and the expectation, not the sign. That is, if X is 5 below $E[X]$ we should treat that the same as if X is 5 above $E[X]$.

There are two common ways to turn value into its magnitude in mathematics generally: squaring the number and using absolute values. Both of these tactics are useful approaches to defining measures of spread, and they result in the **variance** when using the expected value of the squared deviations, and the **mean absolute deviation** when using the absolute value. While $E[|X - E[X]|]$ is perhaps the more intuitive quantity to consider, generally speaking it will not be the one that we use.

Definition 6.8 (Variance)**.** The variance of a random variable, typically denoted $\text{var}(X)$, is given by the expected value of the squared deviations of a random variable from its mean. That is,

$$\text{var}(X) = E\left[(X - E[X])^2\right].$$

Definition 6.9 (Mean Absolute Deviation)**.** The mean absolute deviation of a random variable, typically denoted $\text{MAD}(X)$, is given by the expected value of the absolute value of the deviations of a random variable from its mean. That is,

$$\text{MAD}(X) = E\left[|X - E[X]|\right].$$

In general when we need a positive quantity in mathematics it will typically be preferable to consider the square to the absolute value.[14] The variance is **the** central measure of deviation for random variables. When discussing the variability of a random variable, it will almost universally be in reference to the variance.

Note that if we take $g(X) = (X - E[X])^2$, then the variance of X is the expected value of a transformation. We have seen that to compute these we apply the law of the unconscious statistician, and substitute $g(X)$ into the defining relationship for the expected value, which for the variance gives

$$\text{var}(X) = \sum_{x \in \in \mathcal{X}} (x - E[X])^2 p_X(x).$$

In order to compute the variance, we must also find the mean as the function $g(X)$ relies upon this value.

Example 6.13 (Variation in the Number of Ice Cream Flavours Ordered)**.** Noting how different the range and IQR were of the distribution of number of different flavours ordered by customers at the ice cream parlour, Charles and Sadie decide to turn to the variance and mean absolute deviation to try understand the distribution's variability once-and-for-all. Recall that

[14] The reasons for this are plentiful, but generally squares are easier to handle than absolute values, and as a result become more natural quantities to consider.

the probability mass function for Y, the number of ice cream flavours ordered by a customer, is given by

$$p_Y(y) = \begin{cases} 0.25 & y = 1 \\ 0.25 & y = 2 \\ 0.1 & y = 3 \\ 0.05 & y = 4 \\ 0.05 & y = 5 \\ 0.05 & y = 6 \\ 0.15 & y = 7 \\ 0.05 & y = 8 \\ 0.04 & y = 9 \\ 0.005 & y = 10 \\ 0.005 & y = 50 \\ 0 & \text{otherwise} \end{cases}$$

Find the variance and the mean absolute variation for this distribution.

Solution

For both $\text{var}(Y)$ and $\text{MAD}(Y)$ we require the expected value of Y. To this end note that

$$\begin{aligned} E[Y] = (0.25)(1) + (0.25)(2) + (0.1)(3) + (0.05)(4) + (0.05)(5) + (0.05)(6) + (0.15)(7) \\ + (0.05)(8) + (0.04)(9) + (0.005)(10) + (0.005)(50) = 3.91. \end{aligned}$$

Then, it can be useful to define a table with the probabilities, absolute, and squared deviations, so that the summations can be made easier.

Y	$p_Y(y)$	$(Y - E[Y])^2$	$\lvert Y - E[Y]\rvert$
1	0.25	8.4681	2.91
2	0.25	3.6481	1.91
3	0.1	0.8281	0.91
4	0.05	0.0081	0.09
5	0.05	1.1881	1.09
6	0.05	4.3681	2.09
7	0.15	9.5481	3.09
8	0.05	16.7281	4.09
9	0.04	25.9081	5.09
10	0.005	37.0881	6.09
50	0.005	2124.2881	46.09

With this table we can compute both the variance and mean absolute deviation by taking the second column times by the third, and adding up those results (for the variance) and the second column times the fourth, and adding up those results (for the MAD). Doing so results in $\text{var}(Y) = 17.5019$ and $\text{MAD}(Y) = 2.592$.

6.8.1 Standard Deviation

The higher that an individual random variable's variance is, the more spread we expect there to be in repeated realizations of that quantity. Specifically, the more spread out around the mean value the random variable will be. A random variable with a low variance will concentrate more around its mean value than one with a higher variance. One confusing part of the variance of a random variable is in trying to assess the units. Suppose that a random quantity is measured in a particular set of units – dollars, seconds, grams, or similar. In this case, our interpretations of measures of location will all be in the same units, which aids in drawing connections to the underlying phenomenon that we are trying to study. However, because the variance is squared, we cannot make the same extensions to it. Variance is not measured in the regular units, but in the regular units, *squared*.

Suppose you have a random time being measured, perhaps the reaction time for some treatment to take effect in a treated patient. Finding the mean or median will give you a result that you can read off in seconds. The range and interquartile range both give you the spread in seconds. However, if you work out the variance of this quantity it will be measured in seconds squared – a unit that is challenging to have much intuition about. To remedy this we will often use a transformed version of the variance, called the **standard deviation**, returning the units to be only the original scale.

Definition 6.10. The standard deviation of a random variable is the square root of the variance, which is to say

$$\text{SD}(X) = \sqrt{\text{var}(X)}.$$

We do not often consider computing the standard deviation directly, and so will most commonly refer to the variance when discussing the behaviour of a random variable, but it is important to be able to move seamlessly between these two measures of spread.

Example 6.14 (Variation in the Number of Ice Cream Flavours Ordered, the Finale)**.** Having realized that the IQR and range are not directly comparable to the variance, as they are measured in different units, Charles and Sadie decide to end their investigation in the variability of the number of ice cream flavours ordered by calculating the standard deviation of Y. Recall

that the probability mass function for Y, the number of ice cream flavours ordered by a customer, is given by

$$p_Y(y) = \begin{cases} 0.25 & y = 1 \\ 0.25 & y = 2 \\ 0.1 & y = 3 \\ 0.05 & y = 4 \\ 0.05 & y = 5 \\ 0.05 & y = 6 \\ 0.15 & y = 7 \\ 0.05 & y = 8 \\ 0.04 & y = 9 \\ 0.005 & y = 10 \\ 0.005 & y = 50 \\ 0 & \text{otherwise} \end{cases}.$$

What is the standard deviation of this random variable?

Solution

The standard deviation is given by the square root of the variance. This is given by $\sqrt{17.5019} = 4.18353$.

Example 6.15 (The Dolphin Olympics)**.** Sadie and Charles, on a trip to a beach in a particularly quirky town, find themselves watching the **dolphin Olympics**. These are a set of games put on by the local dolphin populations where they see how high they can jump from the water, and how many flips they can do while they do it. Charles and Sadie get to work determining the probability distributions related to the heights and the number of flips, with Sadie taking the heights and Charles taking the number of flips. At the end of the day they are discussing their findings and Sadie indicates that the standard deviation of the height that was jumped was $1.6m$. Charles is intrigued, saying, "I could have sworn that there was more variation in the jump heights than in the number of flips, but the variance of the flips was 2.56, greater than the variability in the heights!"

Is Charles correct? Why or why not?

Solution

Charles has compared the standard deviation of the first quantity to the variance of the second. We must compare the same measure, and as a result, conclude that the variance of the heights was $1.6^2 = 2.56$, exactly the same as the second.

6.8.2 Computing the Variance

When computing the variance of a random quantity, we often use a shortcut for the formula,

$$\text{var}(X) = E[X^2] - E[X]^2.$$

Generally, this is moderately more straightforward to calculate since X^2 is an easier transformation than $(X - E[X])^2$. This identity will come back time and time again, with a lot of versatility in the ways that it can be used. Typically, when a variance is needed to be calculated the process is to simply compute $E[X]$ and $E[X^2]$, and then apply this relationship.

Proof of the Variance Identity

To determine the variance identity, we need only remember the definition of the variance (being $E[(X-E[X])^2]$), and then use summation techniques to manipulate the expression. To this end consider,

$$\begin{aligned}
\text{var}(X) &= \sum_{x\in\mathcal{X}} (x - E[X])^2 p_X(x) \\
&= \sum_{x\in\mathcal{X}} (x^2 - 2xE[X] + E[X]^2) p_X(x) \\
&= \sum_{x\in\mathcal{X}} x^2 p_X(x) - 2E[X] \sum_{x\in\mathcal{X}} x p_X(x) + E[X^2] \sum_{x\in\mathcal{X}} p_X(x) \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

Example 6.16 (Choosing the Outcome of a Die)**.** Charles and Sadie come across a new game of chance involving the rolling of a die. In it a player chooses a number between 1 and 6. They then roll a die (up to) 6 times. If they get their chosen number on the first throw for the first time, they get \$1. If they get their chosen number on the second throw, they get \$2. The same goes for the first time seeing their chosen number from throws 3 through 6. If they never get the chosen number, they have to pay \$1.

What is the expected value and variance of a player playing this game?

Solution

Let X represent the winnings of a play of this game. We know that X takes a value in $\{-1, 1, 2, 3, 4, 5, 6\}$. The $p_X(-1)$ is given by the probability that the die never shows the chosen number. We can view this as 6 independent trials, where each trial is the result of a die roll. If A_j is the event that the die shows the selected number, then we can calculate $P(A_1^C, A_2^C, A_3^C, A_4^C, A_5^C, A_6^C) = P(A_1^C)P(A_2^C)P(A_3^C)P(A_4^C)P(A_5^C)P(A_6^C)$. Note that each of these probabilities is equivalent, and equivalently equal to $\frac{5}{6}$, and so this

probability is $\left(\frac{5}{6}\right)^6 \approx 0.335$. The remainder of the outcomes happen if the first time that the number shows up is on toss j. This means that we have $A_1, \ldots, A_{j-1}^C$ and then A_j occurring. As a result, for $j = 1, \ldots, 6$ we get $P(A_1^C, \ldots, A_{j-1}^C, A_j) = P(A_1^C)^{j-1}P(A_j)$. Note that $P(A_j) = \frac{1}{6}$, no matter j, and so the total probability here is $\left(\frac{5}{6}\right)^{j-1}\left(\frac{1}{6}\right)$. This fully defines the probability mass function.
To get the $E[X]$ we can apply the standard expectation formula. To get the variance we can note that $\text{var}(X) = E[X^2] - E[X]^2$, and so while we compute $E[X]$ we can also compute $E[X^2]$. To this end

$$E[X] = (-1)\left(\frac{5}{5}\right)^6 + \frac{1}{6}\left[1 + 2\left(\frac{5}{6}\right) + 3\left(\frac{5}{6}\right)^2 + 4\left(\frac{5}{6}\right)^3 + 5\left(\frac{5}{6}\right)^4 + 6\left(\frac{5}{6}\right)^5\right].$$

Solving this gives approximately \$0.98, so every play of the game you would expect to earn 98 cents, approximately. To find $E[X^2]$, we use essentially the same setup, this time squaring the values,

$$E[X] = (-1)^2\left(\frac{5}{5}\right)^6 + \frac{1}{6}\left[1^2 + 2^2\left(\frac{5}{6}\right) + 3^2\left(\frac{5}{6}\right)^2 + 4^2\left(\frac{5}{6}\right)^3 + 5^2\left(\frac{5}{6}\right)^4 + 6^2\left(\frac{5}{6}\right)^5\right].$$

This is 8.728, approximately. Then the variance will be given by

$$\text{var}(X) = E[X^2] - E[X]^2 \approx 8.728 - (0.98)^2 = 7.7676.$$

6.8.3 The Variance of Transformations

With expectations, we saw that $E[g(X)]$ needed to be directly computed from the definition. The same is true for variances of transformations. Specifically, $\text{var}(g(X))$ is given by $E[(g(X) - E[g(X)])^2]$ which can be simplified with the previous relationship as $E[g(X)^2] - E[g(X)]^2$. Just as with expectations, it is important to realize that $\text{var}(g(X)) \neq g(\text{var}(X))$, and so dealing with transformations requires further work.

With expectations, we highlighted linear transformations as a special case, with $g(X) = aX + b$. For the variance, the linear transformations are also worth distinguishing from others. In particular,

$$\text{var}(aX + b) = a^2\text{var}(X).$$

In the same way that the linearity of expectation demonstrates that the expected value of any constant is that constant, we can use this identity to show that the variance of constant is zero. However, we can also reason to this based on our definitions so far. Suppose that we

have a random variable which is constant.[15] A constant b can be seen as a random variable with probability distribution $p_X(x) = 1$ if $x = b$ and $p_X(x) = 0$ otherwise. The expected value is going to be $E[X] = 1(b) = b$, and $E[X^2] = 1(b)^2 = b^2$. Thus, $\text{var}(b) = E[X^2] - E[X]^2 = b^2 - b^2 = 0$. From an intuitive perspective, there is no variation around the mean of a constant. It is always the same value. As a result, when taking the variance, we know that it should be 0.

Proof of the Variance of Linear Transformations

To calculate the variance of a linear transformation of a random variable we can apply the standard identity for the variance, giving

$$\begin{aligned} E[(aX+b)^2] &= E[a^2X^2 + 2abX + b^2] \\ &= E[a^2X^2] + E[2abX] + E[b^2] \\ &= a^2E[X^2] + 2abE[X] + b^2. \end{aligned}$$

Next, we note that $E[aX + b] = aE[X] + b$ and so

$$\begin{aligned} E[aX+b]^2 &= (aE[X] + b)^2 \\ &= a^2E[X]^2 + 2abE[X] + b^2. \end{aligned}$$

Differencing these two quantities gives

$$a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 = a^2(E[X^2] - E[X]^2).$$

By noting that $E[X^2] - E[X]^2$, we can complete the statement that

$$\text{var}(aX + b) = a^2\text{var}(X).$$

Thus, when applying a linear transformation, only the multiplicative constant matters, and it transforms the variance by a squared factor. This should make some intuitive sense that the additive constant does not change anything. If we consider that variance is a measure of spread, adding a constant value to our random quantity will not make it more or less spread out, it will simply shift where the spread is located. This is not true of the mean, which measures where the center of the distribution is, which helps explain why the result identities are different.

Example 6.17 (Sadie's Trip to America)**.** When Sadie had recently returned from a long trip to America, Charles and Sadie worked out how to convert the expected values of temperatures from Celsius to Fahrenheit. They now wish to do the same, with the variances. The

[15] This seems to be an oxymoron, but it is perfectly well defined.

distribution of daily temperatures, in Celsius, is

$$p_X(x) = \begin{cases} 0.1 & x \in \{10, 11, 12, 13, 14\} \\ 0.05 & x \in \{15, 16, 17, 18, 19, 20, 21, 22, 23, 24\} \\ 0 & \text{otherwise} \end{cases}$$

a. What is the variance of temperatures, in Celsius?
b. Supposing that the temperature in Fahrenheit is given by $Y = 1.8X + 32$, what is the variance of temperatures in Fahrenheit?

Solution

a. Here we find

$$\begin{aligned} E[X] = \sum_{x=10}^{24} x p_X(x) &= 0.1 \sum_{x=10}^{14} x + 0.05 \sum_{x=15}^{24} x \\ &= 15.75. \\ E[X^2] = \sum_{x=10}^{24} x^2 p_X(x) &= 0.1 \sum_{x=10}^{14} x^2 + 0.05 \sum_{x=15}^{24} x^2 \\ &= 267.25 \\ \text{var}(X) &= 267.25 - 15.75^2 \\ &= 19.1875. \end{aligned}$$

b. Since $\text{var}(X) = 19.1875$, then we know that $\text{var}(Y) = \text{var}(1.8X + 32) = 1.8^2 \text{var}(X) = 62.1675$.

Unlike the expectation, the variance of additive terms will not generally be the addition of the variances themselves. That is, we cannot say that $\text{var}(g(X) + h(X)) = \text{var}(g(X)) + \text{var}(h(X))$. Writing out the definition shows issue with this,

$$E[(g(X) + h(X))^2] = E[g(X)^2] + 2E[g(X)h(X)] + E[h(X)^2].$$

The first and third terms here are nicely separated and behave well. However, the central term is not generally easy to simplify. You can view $g(X)h(X)$ as a function itself, and so

$$E[g(X)h(X)] \neq E[g(X)]E[h(X)].$$

Instead, this will typically need to be worked out for any specific set of functions.

Exercises

Exercise 6.1. Find the variance and standard deviation of the sum obtained in tossing a pair of standard dice.

Exercise 6.2. In a lottery there are 200 prizes of \$5, 20 prizes of \$25, and 5 prizes of \$100. Assuming that $10,000$ lottery tickets are to be issued and sold, what is the fair prices to pay for a ticket?

Exercise 6.3. Suppose that X is a random variable with mean 9.5 and variance 0.16. For each of the following, identify whether you can determine the mean and variance of the listed quantity, and if so, find it.

a. $3X$.
b. $2X + 4$.
c. X^2.
d. $\frac{X^3}{X^2}$.

Exercise 6.4. Consider the following pmf.

$$p(x) = \begin{cases} \frac{1}{5} & X \in \{0, 1, 2\} \\ \frac{1}{10} & X \in \{3, 4\} \\ \frac{1}{15} & X \in \{5, 6, 7\} \\ 0 & \text{otherwise.} \end{cases}$$

Calculate:

a. $E[X]$.
b. $\text{var}(X)$.
c. $\zeta(0.5)$.
d. $\text{Range}(X)$.
e. $\text{MAD}(X)$.
f. $E[X^3]$.
g. $\text{var}(X^3)$.
h. $E[2X + 3X^2]$.
i. $\text{var}(e^X)$.

Exercise 6.5. Peter and Paula play a game of chance that consists of several rounds. Each individual round is won, with equal probabilities of 1/2, by either Peter or Paula; the winner then receives one point. Successive rounds are independent. Each has staked \$50 for a total of \$100, and they agree that the game ends as soon as one of them has won a total of 5 points; this player then receives the \$100. After they have completed four rounds, of which Peter has won three and Paula only one, a fire breaks out so that they cannot continue their game.

a. How should the \$100 be divided between Peter and Paula?
b. How should the \$100 be divided in the general case, when Peter needs to win a more rounds and Paula needs to win b more rounds?

Exercise 6.6. Suppose that X is a random variable with mean μ and variance σ^2. Prove that

$$Z = \frac{X - \mu}{\sigma},$$

is a random variable with mean 0 and variance 1.

Exercise 6.7. Describe several measures of location indicating the strengths and drawbacks of each.

Exercise 6.8. Describe several measures of variability indicating the strengths and drawbacks of each.

Exercise 6.9.

a. Suppose that two random quantities have the same mode, median, and expected value. Is their distribution guaranteed to be the same? Is it guaranteed to be similar? Why?
b. Suppose that two random quantities have the same range, IQR, variance, and mean absolute deviation. Is their distribution guaranteed to be the same? Is it guaranteed to be similar? Why?
c. Suppose that two random quantities have the same percentiles for all values for which the percentiles are defined. Is their distribution guaranteed to be teh same? Is it guaranteed to be similar? Why?

Exercise 6.10. Is it possible for a random variable, which only takes on positive values, to have a standard deviation which is larger than its mean? Explain.

Exercise 6.11. Is it possible to, knowing that $\text{var}(X) > \text{var}(Y)$, conclude anything regarding $\text{MAD}(X)$ and $\text{MAD}(Y)$? Explain.

Exercise 6.12. Consider a bag containing three red marbles, two green marbles, and four blue marbles. You randomly draw two marbles without replacement. Let Y represent the number of green marbles drawn.

The probability mass function is given by,

$$p_Y(y) = \begin{cases} \frac{6}{36} & y = 0 \\ \frac{12}{36} & y = 1 \\ \frac{3}{36} & y = 2 \\ 0 & \text{otherwise} \end{cases}$$

a. What is the expected value of X?
b. What is the mode of X?
c. What is the variance of X?
d. What is the range of X?
e. What is the standard deviation of X?

Exercise 6.13. Suppose you toss a biased coin three times. Let Z represent the number of heads obtained.

The probability mass function is given by,

$$p_Z(z) = \begin{cases} \frac{1}{8} & z = 0 \\ \frac{3}{8} & z = 1 \\ \frac{3}{8} & z = 2 \\ \frac{1}{8} & z = 3 \\ 0 & \text{otherwise} \end{cases}$$

a. What is the expected value of X?
b. What is the median of X?
c. What is the variance of X?
d. What is the standard deviation of X?

Exercise 6.14. Suppose you observe the number of cars passing a traffic light in a minute. Let W represent the number of cars observed.

The probability mass function is given by,

$$p_W(w) = \begin{cases} \frac{1}{10} & w = 0 \\ \frac{2}{10} & w = 1 \\ \frac{3}{10} & w = 2 \\ \frac{2}{10} & w = 3 \\ \frac{1}{10} & w = 4 \\ 0 & \text{otherwise} \end{cases}$$

a. What is the expected value of X?
b. What is the mode of X?
c. What is the variance of X?
d. What is $\zeta(0.3)$?
e. What is $\zeta(0.8)$?
f. What is the standard deviation of X?

Exercise 6.15. Consider a book containing 200 pages. You randomly select a page number. Let V represent the sum of the digits in the page number selected.

a. What is the expected value of X?
b. What is the variance of X?
c. What is the range of X?
d. What is the standard deviation of X?

7 Expectations and Variances with Multiple Random Variables

7.1 Conditional Expectation

Up until this point we have considered the marginal probability distribution when exploring the measures of central tendency and spread. These help to summarize the marginal behaviour of a random quantity, capturing the distribution of X alone. When introducing distributions, we also made a point to introduce the conditional distribution as one which is particularly relevant when there is extra information. The question "what do we expect to happen, given that we have an additional piece of information?" is not only well-defined, but it is an incredibly common type of question to ask.[1] To answer it, we require **conditional expectations**.

Definition 7.1. The conditional expectation of a random variable, X, given a second random variable, Y, is the average value of X when we know the value of Y. Specifically, we write $E[X|Y]$, and define this to be

$$E[X|Y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x|y),$$

which is exactly analogous to the defining relationship for $E[X]$, replacing the marginal probability mass function with the conditional probability mass function.

In principle, a conditional expectation is no more challenging to calculate than a marginal expectation. Suppose we want to know the expected value of X assuming that we know that a second random quantity, Y has taken on the value y. We write this as $E[X|Y = y]$, and we replace $p_X(x)$ with $p_{X|Y}(x|y)$ in the defining relationship. That is

$$E[X|Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x|y).$$

[1] For instance, you might ask "how long do we expect a patient to live, given that they received a particular treatment?" or "how much do we expect this house to sell for, given it has a certain square footage?" or "how many goals do we expect this hockey team to score, given their current lineup?" A large number of questions which we may hope to answer using data can be framed as a question of conditional expectation.

We can think of the conditional distribution of $X|Y = y$ as simply being a distribution itself, and then work with that no differently. The conditional variance, which we denote $\text{var}(X|Y = y)$ is defined in an exactly analogous manner, giving

$$\text{var}(X|Y) = E[(X - E[X|Y])^2|Y].$$

Above we supposed that we knew that $Y = y$. However, sometimes we want to work with the conditional distribution more generally. That is, we want to investigate the behaviour of $X|Y$, without yet knowing what Y equals. We can use the same procedure as above, however, this time we leave Y unspecified. We denote this as $E[X|Y]$, and this expression will be a function of Y. Then, whenever a value for Y is observed, we can specify $Y = y$, deriving the specific value. We will typically compute $E[X|Y]$ rather than $E[X|Y = y]$, since once we have $E[X|Y]$ we can easily find $E[X|Y = y]$ for *every* value of y.

Example 7.1 (Charles Commences Crocheting)**.** Charles has recently taking up crocheting, but as it is a new skill, is still in the phase of learning where mistakes are somewhat common. When sitting down to practice, the number of rows that Charles can complete in an hour is being recorded by Sadie, as a random quantity X. After these have been completed, Charles goes back through and counts the number of mistakes that were made, recording this as Y. In their experiments they find that

$$p_{X,Y}(x, y) = \frac{44800}{854769} \frac{1}{(x - y)!y!} \left(\frac{21}{10}\right)^x \left(\frac{3}{7}\right)^y,$$

for $x \in \{1, 2, 3, \dots, 10\}$ and $y \in \{0, 1, 2, \dots, y\}$. Sadie works out that

$$p_X(x) = \frac{44800}{854769} \frac{3^x}{x!}, \quad x \in \{1, 2, 3, \dots, 10\}.$$

a. How could Sadie have worked out $p_X(x)$? You do not need to actually compute it.
b. If we know that $X = 3$, what is the expected value of Y?
c. Generally, given X, write down an expression for the expected value of Y.
d. **Challenge:** Can you simplify the expression in (c)? It may be useful to know that $k\binom{n}{k} = n\binom{n-1}{k-1}$.
e. What is the variance of Y, when $X = 3$?

Solution

a. Sadie could have used the process of marginalization. That is,

$$p_X(x) = \sum_{y=0}^{x} p_{X,Y}(x, y) = \sum_{y=0}^{x} \frac{44800}{854769} \frac{1}{(x - y)!y!} \left(\frac{21}{10}\right)^x \left(\frac{3}{7}\right)^y.$$

b. We want $E[Y|X = 3]$. For this, we can use $p_{Y|X}(y|3)$ as the distribution, which is

expressible as

$$p_{Y|X}(y|3) = \frac{\frac{44800}{854769}\frac{1}{(3-y)!y!}\left(\frac{21}{10}\right)^3\left(\frac{3}{7}\right)^y}{\frac{44800}{854769}\frac{3^3}{(3)!}} = \left(\frac{7}{10}\right)^3\binom{3}{y}\left(\frac{3}{7}\right)^y,$$

where $y \in \{0,1,2,3\}$. Then,

$$\begin{aligned}
E[Y|X=3] &= \sum_{y=0}^{3} y\left(\frac{7}{10}\right)^3\binom{3}{y}\left(\frac{3}{7}\right)^y \\
&= (1)\left(\frac{7}{10}\right)^3\binom{3}{1}\left(\frac{3}{7}\right)^1 + (2)\left(\frac{7}{10}\right)^3\binom{3}{2}\left(\frac{3}{7}\right)^2 + (3)\left(\frac{7}{10}\right)^3\binom{3}{3}\left(\frac{3}{7}\right)^3 \\
&= \frac{9}{10} = 0.9.
\end{aligned}$$

c. Following the same process as above, we first get that the general conditional distribution is given by

$$p_{Y|X}(y|x) = \frac{\frac{44800}{854769}\frac{1}{(x-y)!y!}\left(\frac{21}{10}\right)^x\left(\frac{3}{7}\right)^y}{\frac{44800}{854769}\frac{3^x}{(x)!}} = \binom{x}{y}\left(\frac{7}{10}\right)^x\left(\frac{3}{7}\right)^y,$$

for $y \in \{0,\dots,x\}$. Then the expected value of $E[Y|X]$ can be found as

$$E[Y|X] = \sum_{y=0}^{x} y\binom{x}{y}\left(\frac{7}{10}\right)^x\left(\frac{3}{7}\right)^y.$$

d. We have

$$\begin{aligned}
E[Y|X] &= \sum_{y=0}^{x} y\binom{x}{y}\left(\frac{7}{10}\right)^x\left(\frac{3}{7}\right)^y \\
&= \sum_{y=1}^{x} y\binom{x}{y}\left(\frac{7}{10}\right)^x\left(\frac{3}{7}\right)^y \\
&= \sum_{y=1}^{x} x\binom{x-1}{y-1}\left(\frac{7}{10}\right)^x\left(\frac{3}{7}\right)^y \\
&= \frac{3x}{10}\sum_{y=1}^{x}\binom{x-1}{y-1}\left(\frac{7}{10}\right)^{x-1}\left(\frac{3}{7}\right)^{y-1} \\
&= \frac{3x}{10}\sum_{k=0}^{r}\binom{r}{k}\left(\frac{7}{10}\right)^r\left(\frac{3}{7}\right)^k \\
&= \frac{3x}{10}\sum_{k=0}^{r} p_{Y|X}(k|r) \\
&= \frac{3x}{10}.
\end{aligned}$$

e. We have seen that $E[Y|X=3] = 0.9$. As a result, using the previously derived conditional probability distribution,

$$\begin{aligned}
\text{var}(Y|X=3) &= \sum_{y=0}^{3}(y-0.9)^2 \left(\frac{7}{10}\right)^3 \binom{3}{y} \left(\frac{3}{7}\right)^y \\
&= 0.9^2 \left(\frac{7}{10}\right)^3 \binom{3}{0} \left(\frac{3}{7}\right)^0 + (0.1)^2 \left(\frac{7}{10}\right)^3 \binom{3}{1} \left(\frac{3}{7}\right)^1 \\
&\quad + (1.1)^2 \left(\frac{7}{10}\right)^3 \binom{3}{2} \left(\frac{3}{7}\right)^2 + (2.1)^2 \left(\frac{7}{10}\right)^3 \binom{3}{3} \left(\frac{3}{7}\right)^3 \\
&= 0.63
\end{aligned}$$

7.2 Conditional Expectations as Random Variables

Since $E[X|Y]$ is a function of an unknown random quantity, Y, $E[X|Y]$ is also a random variable.[2] It is a transformation of Y, and as such, it will have some distribution, some expectation, and some variance itself. This is often a confusing concept when it is first introduced, so to recap:

- X and Y are both random variables;
- $E[X]$ and $E[Y]$ are both constant, numerical values describing the distribution of X and Y;
- $E[X|Y=y]$ and $E[Y|X=x]$ are each numeric constants which summarize the distribution of $X|Y=y$ and $Y|X=x$ respectively;
- $E[X|Y]$ and $E[Y|X]$ are functions of Y and X, respectively, and can as such be seen as transformations of (and random quantities depending on) Y and X respectively.

We do not often think of the distribution of $E[X|Y]$ directly, however, there are very useful results regarding its expected value and its variance, which will commonly be exploited. If we take the expected value of $E[X|Y]$ we will find that $E[E[X|Y]] = E[X]$. Note that since $E[X|Y] = g(Y)$ for some transformation, g, the outer expectation is taken with respect to the distribution of Y. Sometimes when this may get confusing we will use notation to emphasize this fact, specifically, $E_Y[E_{X|Y}[X|Y]] = E_X[X]$. This notation is not necessary, but it can clarify when there is much going on, and is a useful technique to fallback on.

[2] It is useful to keep in mind that anytime we do *anything* with a random variable, mathematically, we produce an additional random variable. If we think of a random variable as being some mathematical variable whose value depends on the results of an experiment, then if we take that value and apply a function to it we have a *new* value whose results also depend on the results of an experiment.

The Law of Total Expectation

For any random quantities, X and Y, the Law of Total Expectation states that

$$E[X] = E[E[X|Y]].$$

That is, if we first compute the conditional expectation of X given Y, then take the expected value of this quantity, we compute $E[X]$.

In the same way that it is sometimes easier to first condition on Y in order to compute the marginal distribution of X via applications of the law of total probability, so too can it be easier to first work out conditional expectations, and then take the expected value of the resulting expression.

Proof of the Law of Total Expectation

To prove that law of total expectation, we note that $E[X|Y]$ is a random function of Y. As a result, we can apply the LOTUS to $E[X|Y]$ as a function of Y when we take $E[E[X|Y]]$. Doing so yields,

$$\begin{aligned}
E_Y[E[X|Y]] &= \sum_{y\in\mathcal{Y}} E[X|Y]p_Y(y) \\
&= \sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} x p_{X|Y}(x|Y) \right) p_Y(y) \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} x \frac{p_{X,Y}(x,y)}{p_Y(y)} p_Y(y) \\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} x p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} x p_X(x) \\
&= E[X].
\end{aligned}$$

The remainder of the proof, following an application of the LOTUS relies upon manipulating summations.

Example 7.2 (Charles Crochet Mistakes)**.** While Charles came to understand the expected number of mistakes being made given a certain number of crochet lines being complete, it is easier for Charles to consider this on the basis of hourly errors than conditional hourly errors. Knowing that

$$p_{X,Y}(x,y) = \frac{44800}{854769} \frac{1}{(x-y)!y!} \left(\frac{21}{10}\right)^x \left(\frac{3}{7}\right)^y,$$

for $x \in \{1, 2, 3, \dots, 10\}$ and $y \in \{0, 1, 2, \dots, y\}$, that

$$p_X(x) = \frac{44800}{854769}\frac{3^x}{x!}, \quad x \in \{1, 2, 3, \dots, 10\},$$

and that $E[Y|X] = \frac{3X}{10}$, what is $E[Y]$?

Solution

Here we can apply the law of total expectation. We have that $E[Y] = E[E[Y|X]] = E\left[\frac{3X}{10}\right] = \frac{3}{10}E[X]$. Thus, we need to work out $E[X]$, which can be done via the probability mass function of X. Specifically,

$$\begin{aligned} E[X] &= \sum_{x=1}^{10} x\frac{44800}{854769}\frac{3^x}{x!} \\ &= \frac{44800}{854769}\sum_{x=1}^{10}\frac{3^x}{(x-1)!} \\ &= \frac{44800}{854769} \times \frac{67413}{1120} \\ &= \frac{898840}{284923} \approx 3.155. \end{aligned}$$

Thus, in total, the expected value of Y will be

$$\frac{3}{10} \times \frac{898840}{284923} = \frac{269652}{284923} \approx 0.946.$$

7.3 Conditional Variance

While the conditional expectation is used often, the conditional variance is less central to the study of random variables. As discussed, briefly, the conditional variance is given by the same variance relationship, replacing the marginal probability distribution with the conditional one. Just as with expectations $\text{var}(X|Y = y)$ is a numeric quantity given by $E[(X - E[X|Y = y])^2|Y = y]$ and $\text{var}(X|Y)$ is a random variable given by $E[(X - E[X|Y])^2|Y]$. This means that we can consider the distribution, and critically the expected value of, $\text{var}(X|Y)$. A core result relating to conditional expectations and variances connects these concepts.

The Law of Total Variance

For any random variables X and Y, we can write

$$\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y]).$$

This result can be viewed as decomposing the variance of a random quantity into two separate components, and comes up again in later statistics courses. At this point we can view this as a method for connecting the marginal distribution through the conditional variance nad expectation.

Example 7.3 (Charles' Crochet Consistency)**.** Charles understands that the number of mistakes made per hour (Y) given the number of rows crocheted per hour (X) has $E[Y|X] = 0.3X$. Moreover, the variability in this estimate is given by $\text{var}(Y|X) = 0.21X$. Sadie has worked hard to find out that

$$E[X] = \frac{898840}{284923} \quad \text{and} \quad \text{var}(X) = \frac{214410323010}{81181115929}.$$

Can Charles use this information to understand $\text{var}(Y)$?

Solution

We can apply the law of total variance. Specifically,

$$\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}(E[Y|X]) = E[0.21X] + \text{var}(0.3X) = 0.21E[X] + 0.09\text{var}(X).$$

Plugging in the marginal values gives

$$\text{var}(Y) = 0.21\frac{898840}{284923} + 0.09\frac{214410323010}{81181115929} = \frac{730779688281}{811811159290} \approx 0.90.$$

7.4 Joint Expectations

The final set of techniques to consider[3] relate to making use of the joint distribution between X and Y. Specifically, if we have any function of two random variables, say $g(X, Y)$ and we wish to find $E[g(X, Y)]$. This follows in an exactly analogous derivation to what we have seen so far. In this case, we replace the marginal distribution with the joint distribution. The variance extends in the same manner as well.

Definition 7.2 (Joint Expectation)**.** The joint expectation of a function (g) of two random variables, X and Y, is written $E[g(X, Y)]$. This is an expectation computed with respect to the joint distribution of X and Y, giving

$$E[g(X, Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) p_{X,Y}(x, y).$$

The joint expectation captures the location of a multivariate function, and is readily extended to more than two random variables.

[3] At least, for now.

Definition 7.3 (Joint Variance)**.** The joint variance of a function (g) of two random variables, X and Y, is written $\text{var}(g(X,Y))$. This is a variance computed with respect to the joint distribution of X and Y, giving

$$\text{var}(g(X,Y)) = E[(g(X,Y) - E[g(X,Y)])^2].$$

The joint variance captures the spread of a multivariate function, and is readily extended to more than two random variables.

For instance, if we want to consider the product of two random variables, we could use this technique to determine $E[XY]$ and $\text{var}(XY)$.

Example 7.4 (Door-to-Door Charity Chocolate Bars)**.** Charles and Sadie are helping to raise money for a local charity, and to do so, they are going around house-to-house to sell chocolate bars. As they walk between the homes, they realize that depending on where in the city they are, the number of houses that they visit in a day is going to be vary. Moreover, each time they stop by a house, whether or not they will make a sale is uncertain. If, in any given hour, they take Y to be the number of houses that they visit, and X to be the number of chocolate bars that they sell, then they work out that the joint probability mass function of X and Y is given by

$$p_{X,Y}(x,y) = \frac{2y-1}{36(y+1)}, \quad y \in \{1,\dots,6\}, x \in \{0,\dots,y\}.$$

What is the expected number of chocolate bars per house that the visit?

Solution

We want $E[g(X,Y)]$ where $g(X,Y) = \dfrac{X}{Y}$. Thus, using the defining relationship for joint

probabilities we get

$$\begin{aligned}
E[g(X,Y)] &= E\left[\frac{X}{Y}\right] \\
&= \sum_{y=1}^{6}\sum_{x=0}^{y}\frac{x}{y}\cdot\frac{2y-1}{36(y+1)} \\
&= \sum_{y=1}^{6}\frac{2y-1}{36y(y+1)}\sum_{x=0}^{y}x \\
&= \sum_{y=1}^{6}\frac{2y-1}{36y(y+1)}\cdot\frac{y(y+1)}{2} \\
&= \sum_{y=1}^{6}\frac{2y-1}{72} \\
&= \frac{1}{72}\left[2\sum_{y=1}^{6}y-\sum_{y=1}^{6}1\right] \\
&= \frac{1}{72}\left[42-6\right] = \frac{1}{2}.
\end{aligned}$$

As a result, they sell 0.5 chocolate bars per house that they visit, on average.

It is worth considering, briefly, the ways in which conditional and joint expectations interact. Namely, if we know that $Y = y$, then the transformation $g(X, y)$ only has one random component, which is X. As a result, taking $E[g(X,Y)|Y = y] = E[g(X, y)|Y = y]$. If instead we use the conditional distribution without a specific value, we still have that Y is fixed within the expression, it is just fixed to an unknown quantity. That is $E[g(X,Y)|Y]$ will be a function of Y. We saw before that $E[E[X|Y]] = E[X]$, and the same is true in the joint case. Thus, one technique for computing the joint expectation, $g(X,Y)$ is to first compute the conditional expectation, and then compute the marginal expectation of the resulting quantity.

Example 7.5 (Door-to-Door Charity Chocolate Bars, Marginally Easier)**.** While walking around selling chocolate bars for charity, Charles and Sadie realize that it is fairly straightforward[4] to marginalize the joint probability mass function for the number of houses that they visit and the number of chocolate bars that they sell, since X does not actually appear in the equation. That is, when

$$p_{X,Y}(x,y) = \frac{2y-1}{36(y+1)}, \quad y \in \{1,\dots,6\}, x \in \{0,\dots,y\},$$

[4]Comparatively speaking!

taking the sum $\sum_{x=0}^{y} p_{X,Y}(x,y) = (y+1)p_{X,Y}(x,y) = \dfrac{2y-1}{36}$. This gives the marginal probability distribution of Y. They also realize that this has greatly simplified finding the conditional probability distribution of X given Y.

a. Find the expected value of the number of chocolate bars per house that they sell, given the number of houses they visit.
b. Use this result to determine the expected number of chocolate bars sold per visited house.

Solution

a. Note that

$$p_{X|Y}(x|y) = \frac{1}{y+1} \quad x \in \{0, 1, \dots, y\}.$$

As a result, we can compute

$$E[\frac{X}{Y}|Y]\frac{1}{Y}E[X|Y] = \frac{1}{Y}\sum_{x=0}^{Y}\frac{x}{Y+1} = \frac{1}{Y(Y+1)}\cdot\frac{Y(Y+1)}{2} = \frac{1}{2}.$$

b. Note that,

$$E\left[\frac{Y}{X}\right] = E\left[E\left[\frac{Y}{X}\middle| Y\right]\right] = E[0.5] = 0.5,$$

just as before.

7.4.1 Linear Combinations of Random Variables

With this relationship, we can ask about taking combinations of random variables. For instance, if we have two random variables X and Y, we can use this framework to understand how $X+Y$ behaves. An application of these rules with the function $g(X,Y) = X+Y$ gives $E[X+Y] = E[X]+E[Y]$, and that $\text{var}(X+Y) = \text{var}(X)+\text{var}(Y)+2E[(X-E[X])(Y-E[Y])]$. Thus, we see that expectations are linear over combinations of random variables, however, variances are not. The term $E[(X-E[X])(Y-E[Y])]$ is called the **covariance** of X and Y, and it is a measure of how related X and Y happen to be.

Definition 7.4 (Covariance)**.** The covariance of two random variables, X and Y, is given by $\text{cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$. The covariance measures the relationship between X and Y, where a positive covariance value means that as X increases, Y will also increase on average (and vice versa). A negative covariance means that as X increases, Y will decrease on average (and vice versa).

The covariance behaves similarly to the variance. We can see directly from the definition that $\text{cov}(X,X) = \text{var}(X)$. Moreover, using similar arguments to those used for the variance, we

can show that

$$\text{cov}(aX + b, cY + d) = ac\text{cov}(X, Y).$$

Covariances remain linear, so that

$$\begin{aligned}\text{cov}(X + Y, X + Y + Z) = \text{cov}(X, X) + \text{cov}(X, Y) + \text{cov}(X, Z)\\ + \text{cov}(Y, X) + \text{cov}(Y, Y) + \text{cov}(Y, Z).\end{aligned}$$

These make covariances somewhat nicer to deal with than variances, and on occasion it may be easier to think of variances as covariances with themselves.

Proofs for the Expectation and Variance of Linear Combinations of Random Variables

With $g(X, Y) = X + Y$, we can consider applying the defining relationship for joint expectations. That is

$$\begin{aligned}E[X + Y] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x + y)p_{X,Y}(x, y)\\ &= \sum_{x\in\mathcal{X}} x\sum_{y\in\mathcal{Y}} p_{X,Y}(x, y) + \sum_{y\in\mathcal{Y}} y\sum_{x\in\mathcal{X}} p_{X,Y}(x, y)\\ &= \sum_{x\in\mathcal{X}} xp_X(x) + \sum_{y\in\mathcal{Y}} yp_Y(y)\\ &= E[X] + E[Y].\end{aligned}$$

For the variances, we apply the variance relationship, giving

$$\begin{aligned}E[(X + Y - E[X] - E[Y])^2] &= E[((X - E[X]) + (Y - E[Y]))^2]\\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2]\\ &\quad + 2E[(X - E[X])(Y - E[Y])]\\ &= \text{var}(X) + \text{var}(Y) + 2E[(X - E[X])(Y - E[Y])].\end{aligned}$$

Rewriting the covariance in more common terms gives,

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y).$$

Example 7.6 (Charles and Sadie's Orchard Trip)**.** Charles and Sadie adore visiting orchards when the season is right. They are happy to go pick fruit, and then combine everything that they manage together at the end. On one trip to a favourite orchard of theirs they decide to split up and pick separately. This works well enough that on the trip home they decide to start analyzing this behaviour. They take X to be the quantity of fruit picked by Sadie, and Y to be the quantity of fruit picked by Charles. Suppose that they figure that the number of kilograms of fruit jointly picked by them is represented by the probability mass function

$$p_{X,Y}(x, y) = \frac{14xy}{251(x + y)}, \quad x, y \in \{1, \ldots, 4\}.$$

a. What quantity of fruit does Sadie pick on average? Charles?
b. What is the variance of fruit picked by Sadie? Charles?
c. What is the covariance between the amount of fruit that Sadie and Charles each pick?
d. What is the expected total fruit picked between both Charles and Sadie?
e. What is the variance of the total fruit picked between both Charles and Sadie?

Solution

a. Note that the joint distribution is symmetric in X and Y so they will have equal expected value. We solve for $E[X]$ and note that $E[Y]$ will be the same.

$$\begin{aligned} E[X] &= \sum_{x=1}^{4} x p_X(x) \\ &= \sum_{x=1}^{4} x \sum_{y=1}^{4} p_{X,Y}(x,y) \\ &= \sum_{x=1}^{4} \sum_{y=1}^{4} \frac{14x^2 y}{251(x+y)} \\ &= \frac{700}{251} \approx 2.789. \end{aligned}$$

b. For the variance we note that they will also be equivalent between the two individuals. Since we have $E[X]$ already, we simply compute $E[X^2]$ which is given by

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{4} x^2 p_X(x) \\ &= \sum_{x=1}^{4} x^2 \sum_{y=1}^{4} p_{X,Y}(x,y) \\ &= \sum_{x=1}^{4} \sum_{y=1}^{4} \frac{14x^3 y}{251(x+y)} \\ &= \frac{11169}{1255} \approx 8.900. \end{aligned}$$

Thus, the variance is going to be given by

$$\text{var}(X) = \frac{11169}{1255} - \left(\frac{700}{251}\right)^2 = \frac{353419}{315005}.$$

c. The covariance is computable using the joint distribution directly, giving

$$\begin{aligned}\text{cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= \sum_{x=1}^{4}\sum_{y=1}^{4}\left(X-\frac{700}{251}\right)\left(Y-\frac{700}{251}\right)\frac{14xy}{251(x+y)} \\ &= \frac{17581}{315005} \approx 0.059\end{aligned}$$

d. We want $E[X+Y] = E[X] + E[Y] = \frac{1400}{251}$. This could also be computed directly,

$$\begin{aligned}E[X+Y] &= \sum_{x=1}^{4}\sum_{y=1}^{4}(x+y)\frac{14xy}{251(x+y)} \\ &= \frac{14}{251}\sum_{x=1}^{4}\sum_{y=1}^{4}xy \\ &= \frac{14}{251}\times 100 = \frac{1400}{251}.\end{aligned}$$

e. We want

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) = 2\times\frac{353419}{315005} + 2\times\frac{17581}{315005} = \frac{148400}{63001}.$$

7.5 Expectations when Random Variables are Independent

Whenever we can assume independence of random quantities, we can greatly simplify the expressions we are dealing with. Recall that the key defining relationship with independence is that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. Suppose then that we can write $g(X,Y) = g_X(X)h_Y(Y)$. For instance, for the covariance we have $g(X,Y) = (x-E[X])(Y-E[Y])$ and so $g_X(X) = X-E[X]$ and $h_Y(Y) = Y - E[Y]$. If we want to compute $E[g(X,Y)]$ then we get

$$\begin{aligned}E[g(X,Y)] &= E[g_X(X)h_Y(Y)] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g_X(x)h_Y(y)p_{X,Y}(x,y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g_X(x)h_Y(y)p_X(x)p_Y(y) \\ &= \sum_{x\in\mathcal{X}} g_X(x)p_X(x)\sum_{y\in\mathcal{Y}} h_Y(y)p_Y(y) \\ &= E[g_X(X)]E[h_Y(Y)].\end{aligned}$$

Thus, whenever random variables are independent, we have the ability to separate them over their expectations. Stated succinctly, whenever $X \perp Y$, then

$$E[g_X(X)h_Y(Y)] = E[g_X(X)]E[h_Y(Y)].$$

Example 7.7 (Sadie and Charles Turn back To Dice)**.** With a more thorough understanding of joint distributions, Sadie and Charles turn back to games of chance. They are considering games where dice are rolled a set number of times, and then the total sum is recorded across all of the rolls. They want to understand both what happens in expectation, and the variability of these trials.

a. Suppose that X_1 is a single roll of a die. What is the mean and variance of the roll?
b. Suppose that X_1 and X_2 are the results from two independent rolls of a die. What is the mean and variance of $X_1 + X_2$.
c. Suppose that $X_1, \dots, X_n$ are the results from n independent rolls of a die. What is the mean and variance of $\sum_{i=1}^n X_i$?
d. Suppose that $X_1, \dots, X_n$ are the results from n independent rolls of a die. Moreover, take Z to be the result of an additional independent die roll. What is the mean and variance of $Z \times \sum_{i=1}^n X_i$?

Solution

a. We know that X_1 takes on the values in $\{1, \dots, 6\}$ with equal probability each. Thus we have

$$\begin{aligned} E[X_1] &= \sum_{x=1}^{6} \frac{x}{6} = \frac{21}{6} = 3.5 \\ E[X_1^2] &= \sum_{x=1}^{6} \frac{x^2}{6} = \frac{91}{6} \\ \text{var}(X_1) &= \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12}. \end{aligned}$$

b. Note that because X_1 and X_2 are independent, we get

$$\begin{aligned} E[X_1 + X_2] &= E[X_1] + E[X_2] \\ &= 2\frac{21}{6} = 7. \\ \text{var}(X_1 + X_2) &= \text{var}(X_1) + \text{var}(X_2) \\ &= 2\frac{35}{12} = \frac{35}{6}. \end{aligned}$$

c. Note that because X_1 and X_2 are independent, we get

$$\begin{aligned} E[\sum_{i=1}^{n} X_i] &= \sum_{i=1}^{n} E[X_i] \\ &= \frac{21n}{6} \\ \text{var}(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} \text{var}(X_i) \\ &= \frac{35n}{12}. \end{aligned}$$

d. Note that Z and $X_1, \ldots, X_n$ are all independent. As a result if we take $T = \sum_{i=1}^{n} X_i$, then $T \perp Z$ and so $E[TZ] = E[T]E[Z]$. Thus, from (a), (b), and (c) we get $E[ZT] = \frac{21}{6} \times \frac{21n}{6} = \frac{49n}{4}$.

For the variance note that we get

$$\text{var}(TZ) = E[(TZ)^2] - E[TZ]^2 = E[T^2]E[Z^2] - E[TZ]^2.$$

For $E[T^2]$ and $E[Z^2]$ note that $E[T^2] = \text{var}(T) + E[T]^2$, and similarly for Z. Thus

$$E[T^2] = \frac{35n}{12} + \left(\frac{21n}{6}\right)^2,$$

and

$$E[Z^2] = \frac{91}{6},$$

thus

$$\text{var}(TZ) = \left(\frac{35n}{12} + \left(\frac{21n}{6}\right)^2\right)\left(\frac{91}{6}\right) - \left(\frac{49n}{4}\right)^2 = \frac{245n(21n+26)}{144}.$$

Consider what this means for the covariance between independent random variables. If $X \perp Y$ then

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[X - E[X]]E[Y - E[Y]].$$

Note that $E[X - E[X]] = E[X] - E[X] = 0$, and the same for $E[Y - E[Y]]$. Thus, if $X \perp Y$ then $\text{cov}(X, Y) = 0$. That is to say, if X and Y are independent, then $\text{cov}(X, Y) = 0$. It is critical to note that this relationship does **not** go both ways. You are able to have $\text{cov}(X, Y) = 0$ even if $X \not\perp Y$.

While the covariance is interesting in and of itself, the result allows us to simplify the expression for the variance of a sum of two random variables. Specifically, for independent random variables X and Y we also must have that $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$. This further

extends to more than two random variables, where if (for instance) we have $X_1, X_2, \ldots, X_n$ all independent, we get both that

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i],$$

and that

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i).$$

These are relationships that we will use **heavily** once we begin to consider statistics. Note, this extension to more than two random variables applies to all of the concepts discussed throughout this chapter.

In order to do so, the relevant joint distribution, or conditional distribution would need to be substituted into the definitions. Often the complexity here becomes a matter of keeping track of which quantities are random, and which are not. For instance, if we have X, Y, Z as random variables, then $E[X|Y, Z]$ is a random function of Y and Z. We will still have that $E[E[X|Y, Z]] = E[X]$, however, the outer expectation is now the joint expectation with respect to Y and Z. As a result, we can also write $E[E[X|Y, Z]|Y]$. The first expectation will be with respect to $X|Y, Z$, while the outer expectation is with respect to $Z|Y$. This is a useful demonstration for when making the distribution of the expectation explicit may help clarify what is being computed. In general, the innermost expectations will always have more conditioning variables than the outer ones. Each time we step out, we peel back one of hte conditional variables until the outermost is either a marginal (or joint). This may help to keep things clear.

Exercises

Find the variance and standard deviation of the sum obtained in tossing a pair of standard dice. (Note, this was Exercise 6.1 as well; however, now we can use a different technique for it.)

Exercise 7.1. Consider the joint probability mass function of two random variables, X and Y, given by

$$P(X = x, Y = y) = \frac{1}{150}(x + y), 1 \leq x, y \leq 5.$$

a. Find $E[X|Y = 2]$.
b. Find $E[Y|X]$.
c. Find $\text{var}(Y|X)$.

Exercise 7.2. Consider the joint probability mass function of two random variables, X and Y, given by

$$P(X = x, Y = y) = \frac{1}{12}(y - x)^2, 1 \leq x, y \leq 3.$$

a. Find $E[X|Y = 1]$.
b. Find $E[Y|X]$.
c. Find $\text{var}(Y|X)$.
d. Find $\text{cov}(X, Y)$.

Exercise 7.3. Consider the following joint probability mass function represented as a contingency table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.1	0.2	0.3
$X = 2$	0.2	0.1	0.1

a. Find $E[X]$.
b. Find $E[Y]$.
c. Find $\text{var}(X)$.
d. Find $\text{var}(Y)$.
e. Find $\text{cov}(X, Y)$.
f. Find the expected value and variance of $X + Y$.

Exercise 7.4. Suppose that a particular disease is associated with two common types of genetic mutations, say type A and type B. Let A and B correspond to the random variables which count the locations at which each type of mutation has occurred. In order for a type B mutation to occur, a type A must have also occurred at the same location, and so we can say that

$$P(B = b|A = a) = \binom{a}{b}(0.25)^b(0.75)^{a-b}, \quad b \in \{0, 1, 2, \ldots, a\}.$$

Moreover, suppose that

$$P(A = a) = \frac{10 - a}{45} \quad a \in \{0, 1, 2, 3, 4, 5\}.$$

a. Find $E[B|A]$.
b. Find $E[B]$.
c. Find $\text{var}(B|A)$.
d. Find $\text{var}(B)$.

Exercise 7.5. Suppose a factory produces two types of products: Widgets and Gadgets. Let W and G represent the random variables denoting the number of units produced for each

type, in a particular hour. Suppose that the following is observed as the joint probability mass function

$$P(W = w, G = g) = \frac{1}{80}, \quad w \in \{1, 2, 3, \dots, 8\}, g \in \{1, 2, 3, \dots, 10\}.$$

Find both the mean and variance of $W + G$.

8 The Named Discrete Distributions

8.1 General Named Distributions and the Discrete Uniform

So far, our discussion of probability distributions and their summaries has centered on general results for arbitrary probability mass functions. The basic premise has been that, by knowing a probability mass function, you are able to understand the complete behaviour of a random quantity. Directly from this mass function we are able to derive summaries for the behaviour, for instance describing the location and variability of the random variable. In short by knowing the probability mass function[1] we immediately understand how the random variable behaves. We have not, however, spent much time discussing where the probability mass functions actually *come* from.

We have seen one fairly general probability mass function, the one deriving from the equally likely outcomes model. This probability mass function is completely defined by the set of possible values that the random variable can take on. Suppose that we restrict our attention to the sample space being a set of k integers from a through to $a+k$.[2] When setup in this way, this distribution is often referred to as the **discrete uniform distribution**. Typically, we will use the values for the lower bound (a) and the upper bound ($b = a+k$) to define *which* discrete uniform we are discussing. If we say that X follows a discrete uniform distribution with parameters a and b we are say that X is a random variable which has an equal probability of taking any of the integers from a to b. Put differently, we have that

$$p_X(x) = \begin{cases} \frac{1}{b-a+1} & x \in \{a, a+1, ..., b\} \\ 0 & \text{otherwise.} \end{cases}$$

Whenever we want to say that X follows a particular distribution, wwe use a mathematical shorthand to do so. Specifically, we write $X \sim$ Distribution(parameters) to mean "X follows the Distribution with parameters." For instance, if X represents the results of a fair six-sided die roll, we can write $X \sim$ Discrete Uniform$(1, 6)$. We will typically shorten this to something like $X \sim$ D.Unif$(1, 6)$. Knowing the probability mass function of X, we also immediately can work out the expectation and variance for the random variable. Doing this results in $E[X] = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a+1)^2-1}{12}$. This means that simply by knowing that a random

[1] ... and to a lesser extent, the expected value and variance.

[2] Note that this assumption is not actually restrictive: if you have any k items you can simply label each of the items one of the numbers between 1 and k and take $a = 1$.

variable follows a discrete uniform distribution we also immediately know[3] any of the properties that we have discussed up until this point.

Mean and Variance of Discrete Uniform

Recall that $E[X]$ is given by

$$E[X] = \sum_{x \in \mathcal{X}} x p_X(x).$$

Thus, if we take X to be from the discrete uniform then

$$\begin{aligned} E[X] &= \sum_{x=a}^{b} x \frac{1}{b-a+1} \\ &= \frac{1}{b-a+1} \sum_{x=a}^{b} x \\ &= \frac{1}{b-a+1} \times \frac{1}{2}(b-a+1)(a+b) \\ &= \frac{a+b}{2}. \end{aligned}$$

For the variance, we can take the same derivation as above, this time solving $E[X^2]$. For this we get

$$\begin{aligned} E[X^2] &= \sum_{x=a}^{b} x^2 \frac{1}{b-a+1} \\ &= \frac{1}{b-a+1} \sum_{x=a}^{b} x^2 \\ &= \frac{1}{b-a+1} \times \frac{1}{6}(b-a+1)(2a^2+2ab-a+2b^2+b) \\ &= \frac{2a^2+2ab-a+2b^2+b}{6}. \end{aligned}$$

[3] Or, if we forget, can lookup.

Then,

$$\begin{aligned}
\text{var}(X) &= E[X^2] - E[X]^2 \\
&= \frac{2a^2 + 2ab - a + 2b^2 + b}{6} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{8a^2 + 8ab - 4a + 8b^2 + 4b - 6a^2 - 12ab - 6b^2}{24} \\
&= \frac{2a^2 - 4ab - 4a + 4b + 2b^2}{24} \\
&= \frac{a^2 - 2ab - 2a + 2b + b^2}{12} \\
&= \frac{a^2 - 2ab - 2a + 2b + b^2 + 1 - 1}{12} \\
&= \frac{(b - a + 1)^2 - 1}{12}.
\end{aligned}$$

This realization is particularly powerful. There are many real-world quantities which, through inspection, must follow a discrete uniform distribution. For instance, consider rolling a die. In the case of a die roll we take $a = 1$, $b = 6$, and immediately understand that $E[X] = 3.5$, that $\text{var}(X) = \frac{17}{6}$, and that the probability of each value is $\frac{1}{6}$. We could do the same calculations for any die with any sides labeled in consecutive order.

Example 8.1 (Charles and Sadie find Discrete Uniform Quantities)**.** As Charles and Sadie begin to learn about named distributions, they decide that it is worthwhile to try to find examples of the named distributions around them. To do so, they start discussing a number of possibilities, each of them trying to describe whether and how different quantities obey the distribution.

For each of the following, help Charles and Sadie provide a justification of why a discrete uniform would (or would not be) appropriate, including a description of the parameters.

a. At a recent hockey game, Charles and Sadie were spectators and the 50-50 raffle was selected by choosing the seat number between 1 and 4000.
b. Charles' sibling is having a child, and Charles has been thinking about what day of the week the child is going to be born on.
c. Sadie enjoys playing the birthday-guessing-game when in public, wherein Sadie attempts to guess what day of the year random patrons of the coffee shop were born on.
d. In their free time, Charles and Sadie enjoy playing role-playing games. These games often necessitate the rolling of dice, from the standard six-sided ones, through to large twenty-sided dice, or smaller three-sided ones.
e. Charles' struggles to wake up in the mornings and as a result uses an alarm that plays a random song on a streaming music service. The song is selected from Charles' library at random, keeping each morning exciting!

Solution

a. The winning seat can be treated as a random variable, X, which is distributed according to D.Unif$(1, 4000)$. The mean and variance are not particularly meaningful in this setting, however, each individual will have a constant probability of winning.
b. If the days of the week are labelled 1 through 7 (say, Sunday through Saturday), then it is reasonable to suspect that Y representing the numerically-encoded day of the week will follow approximately a D. Unif$(1, 7)$. It may be arguable that some days are less likely than others to be born on (empirically, Sundays appear to have about 0.12 which is less than 1 in 7), however, it is probably a reasonably close approximation.
c. Here we can either view Sadie's guess *or* the true birthday as the random quantity. If Sadie truly guesses at random then it is likely this will follow a D. Unif$(1, 365)$, being well-represented by the discrete uniform. For one's true birthdays, if we are willing to ignore seasonal birth effects and leap years, then it is reasonable to assume that it too will be D. Unif$(1, 365)$, however, these are both effects that do truly exist and would slightly alter from the uniform probabilities.
d. Each roll of the die gives a discrete uniform based on the number of sides that it has. For any d sided die, we can represent it via a D. Unif$(1, d)$. In these types of games, the mean and variance are more pertinent than in most of the other examples we have discussed. Importantly, if we roll multiple dice and take a sum (or a maximum, minimum, etc.) these no longer will be described by the uniform distribution.
e. If Charles were to enumerate the songs on the streaming service with 1 to n, then the played song may follow a D. Unif$(1, n)$ distribution.[4]

While the discrete uniform can be useful for real-world applications, it is also a comparatively simple distribution. The main point of this discussion is not actually to introduce the discrete uniform, but rather to introduce the concept of a **named distribution**. There are processes in the world which occur frequently enough, in a wide array of settings, with the same underlying structure for their uncertainty. If we study one version of these general processes we can derive the mass function, expectation, and variance for them. Then, we are easily able to describe the probabilistic behaviour of other quantities following a similar process. At that point, understanding the uncertainty of random quantities becomes a matter of matching the processes to the correct distribution, and then applying what we know about that distribution directly. While not every process will directly correspond to a known, named distribution, we can often get very close using just a handful of these.[5]

[4] Interestingly, shuffle algorithms employed by streaming companies typically deviate far from the discrete uniform distribution. When Apple introduced shuffle to the iPod it originally had the next song be chosen as a discrete uniform over the as-yet unplayed songs. However, people found that this did not suitably feel *random* as maybe you would hear the same artist a few times in a row; they moved a way from an equally-likely model of selection to have people *feel* like it was more random.

[5] Somewhat interestingly, this phenomenon of getting far with a small amount of effort comes up in a large

For each named distribution there is an underlying structure describing the scenarios in which it arises. For instance, for the discrete uniform, this is when there is a set of equally likely outcomes which can be described using consecutive integers. Once matched, there will is a probability mass function, an expected value, and a variance associated with the distribution. Importantly, all of these quantities will depend on some **parameters**. In the discrete uniform we used the parameters a and b. These parameters specify which *version* of the distribution is relevant for the underlying scenario.

It is best to think of the named distributions as families of distributions, with specific iterations being dictated by the parameter values. If two processes follow the same distribution with different parameters they will not be identically distributed, they are simply drawn from the same family. If two processes have the same parameter values and the same underlying distribution, they are identically distributed and their probabilistic behaviour will be exactly the same. For instance, there is no probabilistic difference between rolling a fair, six-sided die or drawing a card at random from a set of 6 cards labelled 1 through 6. There may be real-world differences which matter, but from a probabilistic point of view, they are exactly the same. This is a useful realization as it allows the use of simple models to understand more complex phenomenon.

8.2 The Bernoulli Distribution

Perhaps the best way to demonstrate the effectiveness of these simple models is to introduce one of the most basic named probability distribution, **the Bernoulli distribution**.[6] The Bernoulli distribution characterizes any statistical experiment with a *binary outcome* when these results are denoted 0 and 1. The parameter that indexes the distribution is p, which gives the probability of observing a 1. The most straightforward application of a Bernoulli random variable is the flip of a coin. Take $X = 1$ if a head is shown, and $X = 0$ if a tails is shown. Then $X \sim \text{Bern}(p)$[7], with $p = 0.5$. If $X \sim \text{Bern}(p)$ then we know that

$$p_X(x) = \begin{cases} p^x(1-p)^{1-x} & x \in \{0,1\} \\ 0 & \text{otherwise.} \end{cases}$$

Further, we can show that that $E[X] = p$ and $\text{var}(X) = p(1-p)$. We typically call $X = 1$ a "success" and $X = 0$ a "failure" when discussing Bernoulli random variables.

A coin flip is, by itself, not particularly interesting. However *any* statistical experiment with binary outcomes coded this way can be seen as a Bernoulli random variable. Suppose, for

number of contexts. Often referred to as the **80-20 rule** the idea is that 80% of results can be achieved via 20% of efforts. So, for instance, approximately 80% of processes in the world can be described by 20% of probability distributions. This phenomenon is, itself, represented by a named distribution: the **Pareto distribution**. This, for me, is beautifully ironic.

[6]The Bernoulli distribution is named after Jacob Bernoulli. The Bernoulli family was an incredibly prominent family of mathematicians from Switzerland, with their hands all over much of mathematics.

[7]We use Bern is used for Bernoulli distributions

instance, you are interested in whether you will pass a particular course or not. There are two options, a "success" (passing) and a "failure" (failing), and the chances of this are governed by some probability p. Alternatively, suppose you want to know whether the next flight you take will land safely, or whether a particular medical treatment will effectively treat an illness. These are the same situations. Each of these scenarios is analyzed in exactly the same way as a coin toss: the probabilities change, but the underlying functions and mathematical objects do not. There is no probabilistic difference between determining whether a coin will come up heads or whether a plane will safely land.

Example 8.2 (Charles and Sadie find Bernoulli Quantities)**.** In their ongoing quest to best understand the named distributions, Charles ans Sadie continue their discussions of various quantities, and how they may fit a Bernoulli distribution.

For each of the following, help Charles and Sadie provide a justification of why a Bernoulli distribution would (or would not be) appropriate, including a description of the parameter.

a. Charles and Sadie think about entering their favourite coffee shop, and seeing whether or not there is seating available for them.
b. Charles thinks about the games that they have played deciding on who would have to pay, based on the outcomes of a few coin tosses.
c. Sadie is going on a trip shortly, and wonders whether or not the plane will be running on schedule.
d. Charles and Sadie are big sports fans, and want to know whether the local team will win the upcoming match.
e. Sadie wants to take out a book from the library, but is not sure whether it is available.

Solution

a. In this case we can treat $X = 1$ as there being seating available ("success") and $X = 0$ as there not being seating available. The parameter, p would be given by the probability that there is seating available, which is likely not a known probability directly.
b. In this case we can treat $X = 1$ as Sadie having to pay, with $X = 0$ as Charles having to pay. The parameter p corresponds to the probability that Sadie pays, which in most of the original examples would have $p = 0.5$.
c. In this case we can treat $X = 1$ as the plane being on time, and $X = 0$ as it being late. The parameter p would correspond to the probability that the plane is late.
d. In this case we can treat $X = 1$ as the team winning the game, and $X = 0$ as them losing. The parameter p would correspond to the probability that they win.
e. In this case we can treat $X = 1$ as the book being available, and $X = 0$ as it being not available. The parameter p would correspond to the probability that it is available.

8.3 The Binomial Distribution

A natural extension to tossing a coin once and seeing if it comes up heads or not is tossing a coin n times and counting how many times it comes up heads. If we take X to be the number of "successes" in n independent and identically distributed Bernoulli trials, then we say that X has a **binomial distribution**. The binomial distribution is characterized by two different parameters, the number of trials that are being performed, denoted n, and the probability of a success on each trial, p. We write $X \sim \text{Bin}(n, p)$.

Example 8.3 (Missing All the Good Games)**.** Charles and Sadie often will attend hockey games together for their local team. In order to keep their costs down, they try not to go to every single game. Instead, each time a game comes around that they can both attend, they roll a die. If it is a 3 or higher, they go, otherwise they stay home. One season they are looking back over the 30 games, and seeing which they attended and which they did not. The team won 19 of the games, and lost the other 11. However, of the 18 games that Charles and Sadie attended that season, they saw all 11 losses and only 7 of the wins! They cannot help but feel like they got very unlucky.

a. How many wins should Charles and Sadie have expected to see?
b. What is the variance in the number of wins that they could have expected?
c. What is the probability that, for a season record of 19 wins and 11 losses, they would have been present for 7 or fewer wins?

Solution

Suppose that we take X to represent the number of winning games that Sadie and Charles are present for. In this case, looking back on the season, we know that X is going to be binomial with $n = 19$ and $p = \frac{2}{3}$, since they go if a 3 or greater is rolled on the die. Thus, we can take $X \sim \text{Bin}(19, \frac{2}{3})$.

a. $E[X] = np = \frac{2}{3}(19) = 12.666$. Thus they should have expected to see around 12 or 13 wins.
b. $\text{var}(X) = np(1-p) = 4.222$.

c. We want $P(X \leq 7)$. For this we get

$$\begin{aligned} P(X \leq 7) &= \sum_{x=0}^{7} p_X(x) \\ &= \sum_{x=0}^{7} \binom{19}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{19-x} \\ &= \frac{2876233}{387420489} \\ &\approx 0.0074. \end{aligned}$$

This leaves them very unlucky.

If we know that $X \sim \text{Bin}(n, p)$ then we also know that

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, 1, \ldots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

This is the first distribution we have seen which knowing the underlying distribution would not have immediately translated into knowing the probability mass function, which begins to illustrate why this is a useful area of study. We can work out that $E[X] = np$ and $\text{var}(X) = np(1-p)$.

Example 8.4 (Charles and Sadie find Binomial Quantities)**.** The adventures in understanding named distributions continue, this time Charles ans Sadie wonder where they may find Binomial distributions.

For each of the following, help Charles and Sadie provide a justification of why a Binomial distribution would (or would not be) appropriate, including a description of the parameters.

a. Charles rolls ten, six-sided dice, counting up the number of them which show a 1.
b. Charles and Sadie reflect on the record of who has paid during their past 20 visits to the coffee shop.
c. Sadie starts playing on a baseball league that uses a pitching machine, and wants to know, without practicing, how many hits occur in the first 20 *at bats*.
d. Charles and Sadie distribute flyers for an upcoming concert that their folk-punk band is putting on. They hand out 50 flyers and are wondering how many people will show.
e. Charles and Sadie form a pub trivia team with some friends. One day, their friend who specializes in history is gone one day, and 5 multiple choice history questions come up. They want to know what their score will be randomly guessing on those questions.

Solution

a. The number of dice showing a 1 is represented by a $\text{Bin}(10, \frac{1}{6})$ distribution. Each die has an equal chance of showing a 1, independently from all other rolls, and each roll will definitely be either a 1 or not a 1.
b. We can take X to represent the number of times that Charles has paid. This will be given by a $\text{Bin}(20, 0.5)$ distribution. Each time both are equally likely to pay, independent of all other times, and one of them will always pay.
c. We can take X to be the number of hits, and then claim that this will follow a $\text{Bin}(20, p)$ distribution, where p is the probability that Sadie makes a hit. The caveats that the pitching machine is used and that Sadie does not practice are important since if different pitches were used, or different skill levels emerged as time progresses, then it may not satisfy the conditions of a binomial distribution.
d. Here we can take X, the number who attend the show, as a $\text{Bin}(50, p)$, where p is the probability that any individual who saw the flyer attends the show. In order for this to hold we need to assume that all individuals choose to attend or not individually, without influence from anyone else, and that the probability is constant across different people.
e. Suppose that there are 4 multiple choice options, then the number they answer correctly will follow a $\text{Bin}(5, 0.25)$ distribution. Assuming that their guesses are independent across the different questions, and that they do not have a better sense for some questions than others.

Note that the binomial distribution can be constructed by summing independent and identically distributed Bernoulli variables. Specifically, if $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(p)$[8] then taking

$$Y = \sum_{i=1}^{n} X_i$$

gives a binomial distribution, with n and p. To understand this intuitively, note that if a Bernoulli comes up 1 when we get a heads on a single flip of the coin, then if we flip the coin n times and count the number of heads this is the same as counting the number of 1s from each corresponding Bernoulli trial. Once we know this construction, we can use the properties we have previously seen about independent random variables to work out the mean and variance for the distribution.

[8]Note, $\overset{iid}{\sim}$ means "independent and identically distributed according to…"

Construction of Binomial Random Variables

Suppose that we take $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(p)$. Then, if we define

$$Y = \sum_{i=1}^{n} X_i$$

we can work out $E[Y]$ and $\text{var}(Y)$ directly. First,

$$E[Y] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = nE[X_i] = np.$$

Moreover, owing to independence we get

$$\text{var}(Y) = \text{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p).$$

These are the same as what was listed for the binomial directly. We are also able to derive the probability mass function of the binomial using this construction.
If we note that $P(X_1 = x) = p^x(1-p)^{1-x}$, then be independence we cna write the joint probability mass function as the product of these, which gives:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = p^y(1-p)^{n-y}. \end{aligned}$$

Consider how we could have $Y = k$, for some $k \in \{0, \dots, n\}$. In order for this to be the case we would require exactly k of the n Bernoulli random variables to be 1 and the other $n-k$ to be 0. So, for instance, we may have $\{X_1 = 1, X_2 = 1, \dots, X_k = 1, X_{k+1} = 0, \dots, X_n = 0\}$. If we consider choosing which of the n trials should be 1, we know that there will be $\binom{n}{k}$ of these, and so that gives the number of disjoint combinations that satisfy $Y = k$. Combining this with the joint probability mass function we worked out gives

$$P(Y = y) = \binom{n}{k} p^k (1-p)^{n-k},$$

as required.

It is important to note that, to get the binomial distribution, we have made several key assumptions. First, we are counting the number of successes in a **fixed** number of trials. In order for something to be binomially distributed, we must know in advance how many trials there are under consideration. Second, each of these trials must be independent of one another. The outcome on one cannot impact any of the others. Third, there must be a constant probability of success across all trials. If the probabilities are shifting overtime, then

a binomial is no longer appropriate.

Example 8.5 (Charles' and Sadie's Binomial Mistakes)**.** In trying to learn about the binomial distribution, Charles and Sadie identified several candidates for quantities which did not satisfy a binomial distribution. For each of the following, help Charles and Sadie understand why a Binomial distribution would not be appropriate.

a. Charles rolls ten dice, of various different sizes, counting up the number of them which show a 1.
b. Charles and Sadie are considering the weather for the next week, thinking about how many days there will be rain.
c. Charles and Sadie consider handing out flyers for their band's concert, and give 10 flyers to a group of 10 friends walking by.
d. Sadie is considering a particular major league baseball player, considering the number of hits in a set number of *at bats*.
e. Charles and Sadie are playing a game where they roll a die, then flip a coin the number of times that is shown on the die, counting up the total number of heads.

Solution

a. The issue in this case is that, if the dice have different numbers of sides, the probability that they show a 1 will not be constant. As a result, there is no value of p that gives the probability of success on each trial.
b. The issue in this case is two-fold. First, it is unlikely that there is going to be a constant probability of rain. Even if there is, or we treat it as though it will be, it is unlikely that one day is independent of the next, all the time. That is to say, today's weather likely impacts tomorrows, removing independence.
c. In this case, it is unlikely that the group of 10 friends will independently decide to go or not. Likely if one goes, many will, and vice versa.
d. The issue here is that batters will bat against different pitches, which changes the probability of a hit. That is, some pitchers are less talented, and thus there is a higher probability of getting a hit against them.
e. The issue here is that there is not a set number of trials. Instead, the number of trials is random and dependent on the outcome of a previous statistical experiment.

8.4 The Geometric Distribution

The binomial counted the number of successes in a fixed number of trials. We may be interested in a related question, namely "how many trials would be needed to see a success?" Instead of "how many heads in n flips of a coin?" we may ask "how many flips of a coin to get a head?" Much like the binomial, these quantities will be intimately tied to the Bernoulli distribution.

Once more we are envisioning a sequence of independent and identically distributed trials being performed. However, instead of knowing that we will stop after n trials have been conducted, here we will only stop once we see a particular result. Any random quantities following this process are said to follow a **geometric distribution**. The geometric distribution is parameterized with a single parameter, p, the probability of success. We write $X \sim \text{Geo}(p)$, and have that

$$p_X(x) = \begin{cases} (1-p)^{x-1}p & x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $X \sim \text{Geo}(p)$, then $E[X] = \frac{1}{p}$ and $\text{var}(X) = \frac{1-p}{p^2}$.

Example 8.6 (Charles Plays Darts)**.** Sadie is a very accomplished darts player. Charles is not. Despite Sadie's best efforts, when Charles plays darts it is essentially randomly choosing an area on the dartboard. In one friendly game, Charles decides to only aim at the highest scoring region of the board - the triple twenty. This region occupies about 0.9% of the total area, and the two friends play with a rule where if Charles misses entirely, another dart is thrown. Charles is curious as to how many darts are going to be needed to be thrown until a triple twenty is hit.

a. Find the expected number of darts required.
b. Find the variance in the total number of darts required.
c. What is the probability that it takes **more** than 5 total throws to get a triple twenty?

Solution

This is a geometric distribution with $p = 0.009$. Let $X \sim \text{Geo}(p)$ represent the number of darts that Charles sees.

a. The expected number of darts required, $E[X]$ is given by $1/p$, which is $\dfrac{1}{0.009} = 111.\dot{1}$.
b. The variance of the expected number of darts, $\text{var}(X)$ is given by $\frac{1-p}{p^2} = 12234.5679$.
c. To find $P(X \geq 5)$ it is simpler to work with $P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$. Then we take

$$\begin{aligned} P(X \geq 5) &= 1 - P(X < 5) \\ &= 1 - (P(X=1) + P(X=2) + P(X=3) + P(X=4)) \\ &= 1 - (0.009)(1-0.009)^{1-1} - (0.009)(1-0.009)^{2-1} \\ &\quad - (0.009)(1-0.009)^{3-1} - (0.009)(1-0.009)^{4-1} \\ &= 1 - (0.009)[1 + 0.991 + 0.991^2 + 0.991^3] \\ &= 0.964483... \end{aligned}$$

Thus, with more than 96% probability, Charles will take X or more dart throws to hit the triple twenty.

The geometric distribution differs from other named distributions that we have considered in that the random variable can take on an infinite number of possible values. The probability that X exceeds a very large threshold shrinks to 0, however, there is no maximum value that can be observed. To form the geometric random variable we assume that we are performing independent and identically distributed Bernoulli trials, and that we stop only after the first observed success.[9]

Example 8.7 (Charles and Sadie find Geometric Quantities)**.** Still working through their distributional knowledge, Charles and Sadie are now hoping to identify geometric quantities.

For each of the following, help Charles and Sadie provide a justification of why a Geometric distribution would (or would not be) appropriate, including a description of the parameter.

a. Charles continues to roll a six-sided die until a 6 is gotten.
b. Sadie has paid for coffee for the last several times, they are both wondering how many more times until Charles will have to pay.
c. Charles is considering a new job which would require a whole lot of plane travel, and so Charles begins to wonder how many flights will run on time before the first delay.
d. Sadie, in a round of guess-their-birthday, starts to count the number of people it takes (guessing one birthday for each person) until a birthday is guessed correctly.
e. Charles and Sadie very much enjoy a particular brand of vegan peach yogurt, but it is hard to find. They want to consider how many stores they need to visit before they find it in stock.

Solution

a. This will be $\text{Geo}(\frac{1}{6})$, as the trials are all independent and identically distributed, and the rolling stops at the first success.
b. If we define "Charles has to pay" as a success, then the number of visits to the coffee shop, starting today, until Charles pays will be $\text{Geo}(\frac{1}{2})$. Each trial is independent and identically distributed, and we stop counting when Charles pays.
c. In order for this to follow a geometric distribution we would require that each plane is equally likely to be delayed, and that there is no influence from one plane to another. If so, we take p to be the probability that a plane is delayed, and then X is the number of flights up-to and including the first delayed flight, and $X \sim \text{Geo}(p)$.
d. Assuming that Sadie makes guesses independently of one another, each with a probability of $\frac{1}{365}$ of being correct, then this will be $X \sim \text{Geo}(\frac{1}{365})$.

[9] In the framing we are using here, the random variable X counts the total number of trials *including* the trail upon which the first success was reached. Sometimes you may see this distribution parameterized slightly differently, taking X to instead count the number of failures before the first success. To convert between the two framings we need only subtract 1. There is no meaningful difference in the underlying behaviour.

e. If each store stocks with probability, p, then the total number of stores until they purchase will be Geo(p). However, this will assume that each store is out-of-stock independently of each other store, and this seems unlikely to be the case supposing (for instance) a common supplier.

8.5 The Negative Binomial Distribution

A natural way to make the geometric distribution more flexible is to not stop after the first success, but rather after a set number of successes. That is, instead of flipping a coin until we see a head, we flip a coin until we see r heads. Any random quantity which follows this general pattern is said to follow a **negative binomial distribution**. We use two parameters to describe the negative binomial distribution, r the number of successes we are looking to achieve, and p the probability of a success on any given trial. We write $X \sim \text{NB}(r, p)$. If we know that $X \sim \text{NB}(r, p)$, then we immediately get

$$p_X(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & x \geq r \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we have $E[X] = \frac{r}{p}$ and $\text{var}(X) = \frac{r(1-p)}{p^2}$. Setting $r = 1$, we get the same quantities explored in the case of the geometric distribution. That is, if $X \sim \text{NB}(1, p)$ then we can also say that $X \sim \text{Geo}(p)$.[10]

Example 8.8 (Charles and Sadie Try to Balance the Dart Game)**.** Charles and Sadie have been continuing to play darts quite often together, and though Charles is improving, Sadie's expertise still makes it an unfair game. As a result, they consider playing alternative versions of the game in which Sadie is at a disadvantage in order to make things more *fair*. Charles has improved so that 20% of the time, the aimed for area of the dart board is hit. Sadie, on the other hand, hits with 60% accuracy. Suppose that they are considering a game where Charles has to hit a spot 3 times, and then Sadie has to hit the same spot r times. They are trying to figure out r to make the game fair.

a. How many tosses will it take, on average, for Charles to hit the area on the target 3 times?
b. What should Charles and Sadie make r, if they want it to take Sadie the same number of tosses on average as Charles?
c. If they set r as in (b), who has a higher variance in the number of tosses it takes?
d. What is the probability that Charles wins before Sadie has a chance to win?
e. What is the probability that Sadie wins on the first available turn?

[10] Like with the geometric distribution, we have taken X to represent the total number of trials considered, including the r successes. There are alternative parameterizations which would count how many *failures* occur prior to the rth success, which can be viewed as the value we consider minus r.

Solution

Let X be the number of tosses it takes for Charles and Y to be the number of tosses it takes for Sadie. We have that $X \sim \text{NB}(3, 0.2)$, while $Y \sim \text{NB}(r, 0.6)$.

a. We know that $E[X] = \dfrac{3}{0.2} = 15$.

b. We have that $E[Y] = \dfrac{r}{0.6}$. Setting this equal to 15 gives

$$\frac{r}{0.6} = 15 \implies r = 15(0.6) = 9.$$

Thus, they should set $r = 9$.

c. We have that $\text{var}(X) = \dfrac{3(0.8)}{0.2^2} = 60$ while $\text{var}(Y) = \dfrac{9(0.4)}{0.6^2} = 10$. Thus, Charles has a far greater variance in the total number of tosses than Sadie.

d. In order for Charles to win before Sadie has a chance, this would require $X < 9$. Thus,

$$\begin{aligned} P(X < 9) &= \sum_{x=3}^{8} p_X(x) \\ &= \sum_{x=3}^{8} \binom{x-1}{2} (0.2)^3 (0.8)^{x-3} \\ &= \left(\frac{0.2}{0.8}\right)^3 \left[\binom{2}{2} + \binom{3}{2}(0.8) + \binom{4}{2}(0.8)^2 + \binom{5}{2}(0.8)^3 \right. \\ &\quad \left. + \binom{6}{2}(0.8)^4 + \binom{7}{2}(0.8)^5\right] \\ &= \frac{79329}{390625} \approx 0.203. \end{aligned}$$

Thus, approximately 20% of the time Charles will win before Sadie even can.

e. The probability that Sadie wins on the first turn is $P(Y = 9)$. We get

$$p_Y(9) = \binom{8}{8}(0.6)^9(0.4)^{9-9} = \frac{19683}{1953125} = 0.010078.$$

Thus it is only around a 1% chance that Sadie wins on turn 9, the earliest possible turn for it.

Example 8.9 (Charles and Sadie find Negative Binomial Quantities)**.** Having mastered many different distributions, Charles and Sadie fix their attention of finding the negative binomial quantities in the world around them.

For each of the following, help Charles and Sadie provide a justification of why a negative

binomial distribution would (or would not be) appropriate, including a description of the parameters.

a. Charles continues to roll a six-sided die until there have been a total of six 6s seen.
b. Charles sets aside \$100 for coffee purchases in a different account. If Charles and Sadie's order comes to \$6.25 per trip, how many trips until the account needs to be restocked.
c. Sadie decided that poker is very fun to play after the first time getting a royal flush.[11] Sadie is very interested how many more hands of poker are likely to be needed until this feeling comes around a few more times.
d. Charles is continuing to crochet. Charles wants to know how many granny squares are needed to be made until there are enough squares without any errors to give as holiday gifts this year.
e. Sadie and Charles need to sell a total of ten chocolate bars for their fundraiser. They want to know how many houses they will need to visit in order to achieve this.

Solution

a. Each trial here is independent, identically distributed with a constant probability. Charles stops after 6 successes, and as a result the negative binomial is satisfied. X will follows a NB$(6, \frac{1}{6})$ distribution.
b. Note that at \$6.25 per order, \$100 buys 16 orders. As a result, this process will end after 16 "successes", which is to say, times when Charles has to pay. We know from past examples that the probability Charles pays is 0.5, independent of all other cases, and as a result this will follow a NB$(16, 0.5)$ distribution.
c. The probability of getting a royal flush is constant in poker hands[12] Moreover, Sadie is interested in a set number of royal flushes being achieved (a few more maybe means 3 or so), and as a result, this falls into the category of negative binomial with r set to the number Sadie wants to see, and p being the probability of seeing a royal flush, which is approximately $\frac{4}{\binom{52}{5}}$, depending on the type of poker being played.
d. Assuming that there are a set number of holiday gifts that Charles wants to give, then we can take this to be r from the negative binomial distribution. Assuming that Charles' odds of making a mistake on any granny square are constant at p, then we can take the probability of success as $1 - p$. Under these assumptions, and assuming the squares are all independent of one another, then this will be NB$(r, 1 - p)$.
e. If we assume that each house, independently, buys chocolate bars with a constant probability, p, then this will follow a NB$(10, p)$ distribution.

[11] Note, a royal flush is the set of cards 10 through ace of a single suit. This is the most rare poker hand in a standard game, occurring in only 4 ways of the $\binom{52}{5}$ total possible hands.
[12] Depending on the type of poker being played, but it is close enough to warrant the assumption likely.

8.6 The Hypergeometric Distribution

One of the use cases demonstrated for the binomial distribution is drawing *with* replacement. In order for the binomial distribution to be relevant it must be the case that the probability of a success is unchanging, and correspondingly, if the process under consideration is random draws from a population then these draws must be with replacement. Otherwise the probabilities would shift.[13] Suppose that we are wish to draw the ace of spades from a standard, shuffled deck of 52 cards. If we begin drawing cards without returning them to the deck after each draw, the probability that the next draw is the ace of spades is increasing over the draws. At first, the probability is $\frac{1}{52}$. If the first card is not the ace of spades, then the next draw it will be $\frac{1}{51}$. This continues until eventually the probability will grow to be 1. As a result, this type of scenario does not fit into the independent and identically distributed Bernoulli trials that we have been exploring.

We require a different setup to model drawing **without** replacement from a finite population. Suppose that our population consists of two types of items, "successes" and "failures". If we are interested in counting how many successes we see in a set number of draws, then this random quantity will follow a **hypergeometric distribution**. The hypergeometric distribution is parameterized using three different parameters: the number of items in the population, N, the number of these which are considered successes, M, and the total number of items that are to be drawn without replacement, n. We write $X \sim \text{HG}(N, M, n)$. If $X \sim \text{HG}(N, M, n)$ then

$$p_X(x) = \begin{cases} \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} & x \in \{\max\{0, M-N+n\}, \ldots, \min\{n, M\}\} \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E[X] = \frac{nM}{N}$ and the variance is given by

$$\text{var}(X) = n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}.$$

Example 8.10 (Charles Sock Drawer)**.** Charles firmly believes that socks should not be sold in pairs.[14] As a result, Charles has decided to keep socks unpaired, free floating in the drawers. In Charles' sock drawer there are 13 individual white socks, 8 individual black socks, and 9 individual red socks. One day, Charles is trying to wear matching socks because of a fancy dinner party, however, it is dark when the socks are being selected.

a. If Charles takes two socks from the drawer, what is the probability that there is a matching pair of white socks?

[13] Note, if the population size is large enough drawing with or without replacement is essentially the same thing. Sometimes, in sufficiently large populations, we do not differentiate between draws with and without replacement.

[14] What happens if you lose on sock? Why must you replace both of them? You can wear either sock on either foot. It really seems to be unnecessary that you cannot buy one at a time.

b. If Charles takes two socks from the drawer, what is the probability that there is a matching pair of **any** colour?
c. If Charles takes 5 socks, how many of them are expected to be red? What is the variance of the number of red socks?
d. How many socks should Charles take out in order to expect to receive at least 2 black socks?

Solution

We can view this as a hypergeometric distribution, where the specific parameters depend on what we are defining as a success. There will always be $N = 30$, and n will correspond to the number of socks that are drawn.

a. Here there are $M = 13$ successes, $N = 30$, and $n = 2$. We wish to know $p_X(2)$, and so we find
$$p_X(2) = \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} = \frac{\binom{30-13}{0}\binom{13}{2}}{\binom{30}{2}} = \frac{26}{145} \approx 0.1793.$$
b. We can say that $P(\text{Pair}) = P(\text{Pair of White}) + P(\text{Pair of Red}) + P(\text{Pair of Black})$. We found $P(\text{Pair of White})$ in (a), and the remaining terms will be similar, changing $M = 9$ for the red socks, and $M = 8$ for the black socks. Thus,
$$\begin{aligned} P(\text{Pair of Red}) &= \frac{\binom{30-9}{0}\binom{9}{2}}{\binom{30}{2}} \\ &= \frac{12}{145} \\ P(\text{Pair of Black}) &= \frac{\binom{30-8}{0}\binom{8}{2}}{\binom{30}{2}} \\ &= \frac{28}{435} \\ P(\text{Pair}) &= \frac{26}{145} + \frac{12}{145} + \frac{28}{435} \\ &= \frac{142}{435} \\ &= 0.3264. \end{aligned}$$
Thus, there is approximately a 33% chance that Charles draws a pair of socks.
c. Here we have $M = 9$ and $n = 5$. Thus, we find
$$E[X] = \frac{(5)(9)}{30} = 1.5,$$
and so 1.5 red socks are to be expected.[15] For the variance, we get
$$\text{var}(X) = 5 \cdot \frac{9}{30} \cdot \frac{30-9}{30} \cdot \frac{30-5}{30-1} = \frac{105}{116} \approx 0.905.$$

d. In this case we want $M = 8$. In order to have $E[X] \geq 2$ we require

$$\frac{n(8)}{30} \geq 2 \implies n \geq \frac{2(30)}{8} = 7.5,$$

and so Charles should draw at least 8 socks in order to expect to have at least 2 black socks in the set of drawn ones.

The Hypergeometric Distribution, Binomial Distribution, and Survey Sampling

The hypergeometric is closely linked to the binomial distribution. If we consider the population described in the hypergeometric setup then the probability of a success on the first draw is $p = \frac{M}{N}$. Note that $E[X] = np$, exactly the same as in the binomial. However, plugging this in for the variance we get $\text{var}(X) = np(1-p)\frac{N-n}{N-1}$. Notice that if $n = 1$, this extra term is simply 1, and for $n > 1$ it will be less than 1. As a result, the variance of the hypergeometric is smaller than the variance of the corresponding binomial. This makes intuitive sense. In the hypergeometric setup, the fact that draws are without replacement means that as more draws go on the probability of observing a success increases, reducing the likelihood of long runs of no observed successes. There is a cap on the behaviour of the random quantity thanks to the finiteness of the population. Correspondingly, the multiplicative term by which the variance shrinks, $\frac{N-n}{N-1}$ is referred to as the **finite population correction**. This factor differentiates the behaviour of the hypergeometric and the binomial.

Note that as N becomes very, very large, as long as n is small by comparison, the finite population correction will approach 1. In other words, drawing without replacement in a large enough sample behaves almost exactly the same as drawing with replacement in the same sample. The binomial distribution can be used to approximate the hypergeometric distribution, so long as the population is very large. Again, this makes sense intuitively. If you have a deck with a million cards in it, and you are going to draw 2, whether or not you return the first one to the deck has very little bearing on the probabilities associated with this scenario. Generally, the binomial distribution is easier to work with, so this approximation can be useful in some settings.

These types of considerations are deeply important in the field of **survey sampling**. In survey sampling, researchers select individuals from a population of interest, and have them respond to questions, or collect information on the respondents. You can think of, for instance, surveys that the university sends out or Canada-wide surveys trying to gauge sentiment on various policies. Generally in a survey you are going to be sampling *without* replacement: you will not include the same person in your sample more than once. However, it may also be the case that you are taking a fairly small sample (say

[15] Now, 1.5 socks can never be observed. As a result, you may say that Charles expects to see either 1 or 2 red socks, though, recall our way of interpreting the expected value.

$n = 1000$) from a very large population (say $N = 40000000$), at which point the finite population correction factor equals 0.999975. This may as well be 1 and you can simplify matters by using binomial probabilities.

Example 8.11 (Charles and Sadie find Hypergeometric Quantities)**.** Charles and Sadie have almost made it through the set of distributions that they want to learn, now moving on to the hypergeometric.

For each of the following, help Charles and Sadie provide a justification of why a hypergeometric distribution would (or would not be) appropriate, including a description of the parameters.

a. When streaming music, Charles will often shuffle the entire library. While listening through, Charles keeps tracks of the number of songs that come on which are favourites.
b. Sadie is considering the number of spades that are likely to show up in different poker hands, from different versions of the game where hands may not be 5 cards.
c. Charles and Sadie run a book club with some of their closest friends. Before each meeting, they take an anonymous vote as to whether the book was enjoyed or not, so that they know the total number of individuals who actually enjoyed the reading. If they know how many people will show up to a meeting, they are interested in how many of those people will have enjoyed the book.
d. Sadie learns of black swans, and wants to understand how many are likely to be seen if Sadie starts to view swans at a swan sanctuary.
e. Charles and Sadie are invited to partake in a survey. The survey is concerned with the number of people living in their town who support investments into transit infrastructure.

Solution

a. This will be hypergeometric assuming that the shuffle algorithm does not play the same songs multiple times in a session. Here N is the total number of songs in the library, M is the number of songs which Charles considers to be favourites, and n is the number that Charles decides to listen to.
b. Assuming that the carts are dealt without replacement this will be hypergeometric. There are $N = 52$ total cards, $M = 13$ spades, and then n will correspond to the size of the hand (such as 5 or 3 or 8).
c. This will be hypergeometric, assuming that there is actually random attendance at the meetings. Here, M will correspond to the number of individuals who enjoyed the book, N is the total size of the book club, and n is the number of people who can attend any given meeting.
d. In order for this to be hypergeometric, we would need to assume that Sadie is able to tell apart the different swans, so as to not count them more than once. In this case, M is the total number of black swans at the sanctuary, N is the total number of swans at the sanctuary, and n is the number of swans that Sadie looks at.

e. This will typically be a hypergeometric, as most surveys are without replacement. Here, N is the size of the possible population of individuals being surveyed, n is the size of the survey, and M is the number of people in the town who support investments into transit infrastructure.

8.7 The Poisson Distribution

The hypergeometric strayed from the pattern of the previously introduced distributions by not being represented as a sequence of Bernoulli trials. However, it was still characterized by a sequence of repeated trials. While many statistical experiments can be framed in this way, there are of course processes which are not described by repeated trials. Consider, for instance, any process where something is observed for a set period of times and events may or may not occur during this interval. Perhaps you sit on the side of the road and count the number of cars traveling by a particular intersection over the course of an hour. Each car going by is an event, but in this setting, the number of events is the random quantity itself. None of the distributions discussed until this point are suited to this type of process.

When we have events which occur at a constant rate, and our interest is in the number of events which occur, then we can make use of the **Poisson distribution**.[16] The Poisson distribution takes a single parameter, λ, which is the average rate of occurrence of the events over the time period we are interested in. We write $X \sim \text{Poi}(\lambda)$. If $X \sim \text{Poi}(\lambda)$ then

$$p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E[X] = \lambda$ and $\text{var}(X) = \lambda$.[17]

Example 8.12 (Charles' Novella Mistakes)**.** Charles has decided to write a short novella. Hard at work, the novella turns out to be 105 pages, at the time of completion. Charles sends a copy off to the printer excited to share it with Sadie. After printing, Charles realizes that the spellcheck on the program indicates that there is a total of 215 errors. Mortified that these were not corrected before handing it over to Sadie, Charles starts to work out just how bad the situation is likely to be.

a. What is the average number of mistakes per page of the novella?
b. What is the variance for the number of mistakes per page in the novella?
c. What is the probability that there were no mistakes on the first page?
d. What is the probability that there were five or more mistakes on the first page?

[16] The Poisson distribution gets it name from the French mathematician, Siméon Poisson, rather than from the french word for *fish*. Though, I suppose it is likely that the surname came from the french word for fish, so in a roundabout way, the distribution is sort of the *Fish Distribution*.

[17] The Poisson distribution is interesting in that the mean and variance are always equal to one another.

Solution

a. On average there will be $\frac{215}{105} = \frac{43}{21}$ errors per page. We can take the distribution to be Poi($\frac{43}{21}$).

b. The variance for a Poisson distribution is given by λ. Thus, it will be $\frac{215}{105} = \frac{43}{21}$.

c. Here we want $p_X(0)$. This is given by

$$p_X(0) = \frac{e^{-43/21}(43/21)^0}{0!} = e^{-43/21} \approx 0.129.$$

Thus, there is about a 13% chance that there are no errors on the first page.

d. Here we want $P(X \geq 5) = 1 - P(X < 5) = 1 - (p_X(0) + p_X(1) + p_X(2) + p_X(3) + p_X(4))$. Solving this directly gives

$$\begin{aligned} P(X \geq 5) = 1 - P(X < 5) &= 1 - \sum_{x=0}^{4} p_X(x) \\ &= 1 - \sum_{x=0}^{4} \frac{e^{-43/21}(43/21)^x}{x!} \\ &\approx 0.0571. \end{aligned}$$

Thus, there is just over a 5% chance of there being 5 or more errors on the first page.

While the most common applications for the Poisson distribution have to do with the occurrences of events throughout time, it is also possible to view this as the occurrences of events throughout space. For instance, if there is a manufacturer producing rope, then the number of defects in a set length of rope is likely to follow the Poisson distribution. Similarly, in a set geographic area, the number of birds of a particular species is likely to follow a Poisson distribution. For the Poisson distribution, we are typically thinking that there is a rate at which events of interest occur, and we can use the Poisson distribution to model the total number of occurrences over some specified interval.

Example 8.13 (Charles and Sadie find Poisson Quantities)**.** As the last named discrete distribution, Charles and Sadie are excitedly exploring quantities around them which may be explained by the Poisson distribution.

For each of the following, help Charles and Sadie provide a justification of why a Poisson distribution would (or would not be) appropriate, including a description of the parameter.

a. Charles got a small injury which required a trip to the emergency room. While there,

Charles begins to count the number of people that arrive over the course of the next hour.

b. Sadie, while out exploring birds, decides to count the number of cardinals in a particular good viewing spot which is about $100m^2$ in size.
c. Charles, while crocheting, realizes that sometimes the yarn in use has defects itself. Charles begins to think about the number of defects in 200 yards of yarn.
d. Charles and Sadie are sitting at a public park overlooking the water. They start to count the number of boats that pass by over the course of the two hours they sit there.
e. Sadie starts a blog sharing recipes for baking vegan breads, and begins to consider the number of visitors that show up to the site each day.

Solution

a. Taking X to represent the number of individuals arriving to the emergency room in the next hour, this will reasonably follow a Poisson distribution. The parameter λ would represent the number of individuals who arrive per hour, on average. It is counting arrivals per hour.
b. If Sadie has a sense of the number of cardinals on average per $100m^2$ area, then a Poisson will be appropriate, supposing that the birds are fairly uniformly spread out. The parameter λ would be the average number of birds in that size of region. It is counting occurrences per area.
c. If Charles knows the average number of defects per 200 yards of yarn, then this value can be taken to be λ. Supposing that the defects occur uniformly, without dependence (e.g., one defect makes more likely around it) then a Poisson distribution is appropriate. It is counting defects per length of yarn.
d. We would take λ to represent the average number of boats in a two hour period passing by the location that they are sitting. Supposing that these happen independently, at roughly uniform rates, then a Poisson is likely appropriate. This would be counting arrivals per (two) hour(s).
e. If Sadie gets a measure of the average number of visitors per day to the website, then setting this to be λ gives a reasonable representation of traffic patterns. We would likely want to assume that the arrivals are not more on some days (for instance, the weekends) than others, but under these assumptions having $X \sim \text{Poi}(\lambda)$ is likely reasonable.

8.8 Using Named Distributions

While many other named, discrete distributions exist, these are the most common. When confronted with a problem in the real-world for which you wish to understand the uncertainty associated with it, a reasonable first step is to determine whether a named distribution is well-suited to representing the underlying phenomenon. Is it a situation with enumerated events

which are equally likely? Use the discrete uniform. Is it a binary outcome? Use the Bernoulli. Are you counting the number of success in a fixed number of trials? Use the binomial. Are you running repeated trials until a (certain number of) success(es)? Use the geometric (or negative binomial). Are you sampling without replacement? Use the hypergeometric. Are you counting events over a fixed space? Use the Poisson.

Once identified, the distribution can be used in exactly the same way as any probability mass function. That is, we still require all the probability rules, event descriptions, and techniques from before. The difference in these cases is that we immediately have access to the correct form of the probability mass function, the expected value, and the variance.

An additional utility with this approach to solving probability questions is that, over time and repeated practice, you can build an intuition as to the behaviour of random variables following these various distributions. Probabilities in general can be deeply unintuitive. It can be hard to assess, without formally working it out, whether an event is likely or unlikely, let alone how likely an event is. However, the lack of intuition from our wider experience can be negated almost entirely by building of intuition through the repeated application of these distributions. You can start to gain a sense of how binomial random variables behave, being able to determine just from inspection whether events seem plausible or not. Much of the study of probability and statistics is about building a set of tools that can overcome the flaws in our intuitive reasoning regarding uncertainty. This comes only through practice, however, this framework of named distributions provides a very solid foundation to perform such practice.

8.8.1 Named Distributions in R

In R all of the named distributions that we have discussed, and in fact, many that we have not discussed, are implemented to make calculations easier. In particular, there are R functions which evaluate the probability mass function for the various distributions. Alongside these, there are also functions which calculate what we refer to as the *cumulative probability*[18], which is to say the $P(X \leq k)$ for some value k. These functions generally are called `d{distname}`, where `{distname}` is the name of the relevant distribution. For instance, `dbinom` for the binomial, `dpoi` for the Poisson, and so forth. These will evaluate the probability mass function at the relevant values. In order to evaluate the cumulative probability at the specified values you would call `p{distname}`. These functions take in a parameter for the value to evaluate at, and then parameters that correspond to the various parameters from the distributions themselves.

```
# Binomial Distribution
dbinom(3, size = 10, p = 0.5)     # P(X = 3) if X~Bin(10, 0.5)
pbinom(3, size = 10, p = 0.5)     # P(X <= 3)
```

[18] More on this in the coming chapter.

```r
# Geometric Distribution
dgeom(10, prob = 0.1)              # P(X=10) if X~Geo(0.1)
pgeom(10, prob = 0.1)              # P(X<=10)

# Negative Binomial
dnbinom(6, size = 3, prob = 0.2)  # P(X=6) if X~NB(6, 0.2)
pnbinom(6, size = 3, prob = 0.2)  # P(X<=6)

# Hypergeometric
# In this case, X ~ HG(N=n+m, M = m, n = k)
dhyper(2, m = 6, n = 9, k = 4)    # P(X=2) X ~ HG(15, 6, 4)
phyper(2, m = 6, n = 9, k = 4)    # P(X<=2)

# Poisson
dpois(5, lambda = 4)               # P(X=5) X ~ Poi(4)
ppois(5, lambda = 4)               # P(X<=5)
## [1] 0.1171875
## [1] 0.171875
## [1] 0.03486784
## [1] 0.6861894
## [1] 0.05872026
## [1] 0.2618025
## [1] 0.3956044
## [1] 0.8571429
## [1] 0.1562935
## [1] 0.7851304
```

Note that we have seen the `sample` function before, which is an implementation of the discrete uniform. To implement the Bernoulli, we can use the binomial distribution with $n = 1$. It is a worthwhile exercise to see if you can use R to start answering the questions from this chapter, numerically. This is the first prominent use case for R programming which can save a tremendous amount of time, and is likely the first use case that becomes directly relevant to the course material.

Exercises

Exercise 8.1. For each of the following scenarios, describe which distribution the indicated random variable most closely resembles, and why.

a. In a manufacturing process, you're interested in the number of trials required to produce the first defective item.

b. A hospital wants to study the number of patients arriving at the emergency room in a fixed time interval, where the arrival rate is low and the events are rare.
c. A quality control team inspects a batch of items and classifies each batch as either defective or non-defective.
d. In a survey, you want to determine the number of people who prefer online shopping over in-store shopping.
e. Researchers are studying a group of patients with a very rare disease. They wish to know whether a particular gene mutation is associated with the disease or not. They sample members of the population who have the disease to further study.
f. An online streaming platform wants to model the number of times a specific video is watched in an hour, given a known average view rate.
g. A baseball player wants is interested in the number of home runs they hit in a game, where the probability of hitting a home run is the same for each at-bat.
h. A call center manager wants to model the number of calls received per minute during a quiet period of the day.
i. In a network setup, multiple backup routers are used, and the network remains functional as long as a certain number of them are operational. A network engineer is interested in how many backup routers are needed to ensure network reliability under various failure scenarios.

Exercise 8.2. Show that if $X \sim \text{Bern}(p)$ then $E[X] = p$ and $\text{var}(X) = p(1-p)$.

Exercise 8.3. Show that if $X_1, \ldots, X_r \overset{iid}{\sim} \text{Geo}(p)$ and $Y = \sum_{i=1}^{r} X_i$, then Y follows a negative binomial distribution. Find the mean and variance of this distribution, and indicate which negative binomial distribution it is.

Exercise 8.4. Suppose that X follows a Binomial distribution with 8 trials and probability of success 0.4. Find:

a. $P(X = 2)$.
b. $P(X = 4)$.
c. $P(X < 2)$.
d. $P(X > 6)$.
e. $E[X]$.
f. $\text{var}(X)$.

Exercise 8.5. If 20% of bolts produced by a machine are defective, determine the probability that out of 4 bolts chosen at random:

a. 1 is defective.
b. 0 are defective.

c. fewer than 2 bolts are defective.

Exercise 8.6. Of all the weld failures in a certain assembly, 85% of them occur in the weld metal itself, and the remaining 15% occur in the base metal. A sample of 20 weld failures is examined.

a. What is the probability that exactly five of them are base metal failures?
b. What is the probability that fewer than four of them are base metal failures?
c. What is the probability that none of them are base metal failures?

Exercise 8.7. Celeste and Dana are playing squash, and Dana is determined to win at least two games. Unfortunately, his chance of winning any one game is only $\frac{1}{4}$, and this chance remains constant however many games he plays against Celeste. The players agree to play 5 games and, if Dana has won at least two by then, play ceases. Otherwise, Dana persuades Celeste to play a further 5 games with him. What is the probability that a. only 5 games are played, and Dana wins at least two of them; a. that 10 games are played, and Dana wins at least two of them.

Exercise 8.8. A machine produces bolts, and for each bolt there is a probability p of it being defective, where results for different bolts are assumed independent. A large batch of the machine's production is inspected by a customer in order to determine whether the batch should be purchased. In the inspection, 10 bolts are selected at random and examined: if none is defective the batch is accepted, and if three or more is defective, the batch is rejected. If only 1 or 2 are defective, a further batch of 10 is selected. If in total between the two batches three or more are defective, the batch is rejected.

a. What is the probability that a decision is made at the first stage?
b. What is the probability that the batch is accepted if $p = 0.15$.

Exercise 8.9. Find the probability of getting a total of 7 at least once in three tosses of a pair of fair dice.

Exercise 8.10. Twenty air-conditioning units have been brought in for service. Twelve of the units have broken compressors and the other eight have broken fans. If seven of these units are randomly selected for repair, what is probability that exactly three have broken fans?

Exercise 8.11. The probability that a computer running a certain operating system crashes on any given day is 0.1.

a. Find the probability that the computer crashes for the first time on the twelfth day.
b. Find the probability that the third crash comes on day 30.

Exercise 8.12. There are 30 restaurants in a particular town. If you assume that 4 of them have health code violations, and a health inspector is going to visit 10 of them at random, what is the probability that:

a. Exactly two of the restaurants visited will have violations.
b. None of the restaurants visited will have violations?

Exercise 8.13. A traffic light at a certain intersection is green 0.5 of the time, yellow 0.1 of the time, and red the remainder of the time. A car approaches the intersection once per day, at a random time. Suppose that X represents the number of days until the car reaches three red lights.

a. Find $P(X = 7)$.
b. What is $E[X]$.
c. Suppose that the car only drives on weekdays, and the first day under consideration is a Saturday. What is $P(X \leq 18)$?

Exercise 8.14. A computer program has a bug that causes it to fail in one out of every thousand runs, on average. In order to find the bug the program is run, independently, until it has failed 5 times.

a. How many times will the program need to run, on average?
b. If it takes 30 seconds to run the program, what is the standard deviation for the amount of time that this debugging will take?

Exercise 8.15. Ten items are to be sampled from a lot of 60. If more than one is defective, the lot will be rejected. What is the probability that the lot will be rejected if:

a. there are 5 defective items.
b. there are 10 defective items.
c. there are 20 defective items.

Exercise 8.16. Suppose that X is a Poisson random variable with rate 4. Find

a. $P(X = 1)$
b. $P(X = 0)$
c. $P(X < 2)$
d. $P(X > 1)$
e. $E[X]$
f. $\text{var}(X)$.

Exercise 8.17. The number of flaws in a given area of aluminum foil happen uniformly at a rate of 3 per square metre. If X represents the number of flaws in a $1m^2$ random sample of foil, what is

a. $P(X = 5)$.
b. $P(X = 0)$.
c. $P(X < 2)$.
d. $P(X > 1)$.

Exercise 8.18. Negative reactions from a particular injection occur at a rate of 1 per 1000 people in the population. Suppose that 2000 people receive the injection.

a. What is the probability that exactly 3 individuals have a negative reaction?
b. What is the probability that more than 2 individuals have a negative reaction?

Exercise 8.19. A telephone exchange receives, on average, 5 calls per minute. Find the probability that

a. in a 1 minute period, no calls are received.
b. in a 2 minute period, fewer than 4 calls are received.
c. in 5 separate 1 minute periods there are exactly four in which 2 or more calls are received.

Exercise 8.20. The number of defective components produced in a process follows a Poisson distribution with a mean of 20. Every defective item has a probability of 0.6 of being repaired, independently of all others.

a. Find the probability that exactly 15 defective components are produced.
b. Given that exactly 15 defective components are produced, find the probability that exactly 10 can be repaired.
c. If N is the number of components produced, and X is the number which are repairable, then given the value of N, what is the pmf of X?

9 Continuous Random Variables

Our discussions of probability distributions and their summaries have focused entirely on discrete random variables. To recap, a discrete random variable is any random numeric quantity that can take on a countable number of values. Discrete random variables are defined in contrast to continuous random variables, which take on values over the span of intervals in uncountably large sets. Suppose that X can take any real number between 0 and 1. There is no way to enumerate the set of possible values for this random quantity, and so it must not be discrete.

Many quantities of interest are better treated as a continuous quantity rather than a discrete one even when this is not technically correct. For instance, time measured in seconds is often best thought of as continuous, even though any stop watch used to grab these measurements will have some limit to the precision with which they can measure. Similarly, lengths and heights will often be better treated as continuous quantities, even though any measuring device will necessarily have some minimal threshold after which it cannot discern distances. Thus, deciding whether a quantity is continuous or discrete is sometimes a judgment call. In general discrete quantities are harder to work with when the set of possibilities is very large. In these cases, not much is lost by treating the random variables as though they were continuous. This distinction is another area which requires the active development of intuition, but once present, the distinctions become second nature.[1]

9.1 Continuous Versus Discrete

Distinguishing whether a random quantity is continuous or discrete is crucial as, broadly speaking, the two types of quantities are treated differently. The same underlying ideas are present, but the distinctions between the two settings require some careful thought. The use of continuous random variables necessitates an understanding of introductory calculus. These course notes are designed to be understood without any experience in calculus, and as a result, we will not spend much time focusing on continuous random variables. However, continuous random variables are *the* dominant type of random variables outside of introductory courses. As a result, understanding the distinctions, and becoming familiar with how they are to be manipulated is an important skill.

[1] Recall Example 5.3 for practice in this.

The key difference between discrete and continuous random variables is that, for discrete random variables the behaviour is governed by assessing $P(X = x)$ for all possible values of x, while for continuous random variables $P(X = x) = 0$ for **every** value of x. This is likely a surprising statement, and as such it is worth reiterating. With discrete random variables we discussed how all of the probabilistic behaviour is governed by the probability mass function. This is defined as $p_X(x) = P(X = x)$. If X is a continuous random variable, we must have $P(X = x) = 0$. Correspondingly, continuous random variables do not have probability mass functions, and to understand the behaviour of these random variables we must us other quantities.

Singleton Probabilities with Continuous Random Variables

While the fact that $P(X = x) = 0$ may seem unintuitive at first glance, it is worth exploring this even further. The nature of a continuous random variable is such that there is no possible way to enumerate all of the values that are possible to be realized by the random quantity. Suppose that we take a set of countably many possible observations and gave each of these a probability of greater than 0 of occurring. Even if we take an infinite number of them, there will still be an uncountably infinite number of events in the sample space that we have not accounted for. We know that the total probability of the sample space must be 1, and so we must have the total probability of the first set of events being less than 1[2].

Now suppose we take another set of countably many events, again giving each of them a positive probability. Once more the sum of all of these probabilities must be less than 1, and specifically, the sum of both sets must also be less than 1. Even after these two sets, there are still uncountably infinite events to go and so we continue this process. Because we always need the total probability to be 1 once all events have been accounted for, and because we will always have uncountably infinite events remaining to account for, we can **never** have a positive probability assigned to each event in a set of events. Even if we made the probability of each of these sets very, very small[3] after some fixed number of countable events the probabilities would be greater than 1, which cannot happen. As such, each event itself must have 0 probability.

An alternative technique for understanding this intuition is to think about how unlikely it really would be to observe any specific value. Suppose that X takes values on the interval $[0, 1]$. Recall that, when we defined probabilities, we discussed them as being the long run proportion of time that an event occurs. Take some event, say $X = 0.5$. Suppose that we took repeated measurements of X which are independent and identically distributed. Now suppose that at some point we exactly do observe $X = 0.5$. Should we expect that this will ever happen again? The next time we get near 0.5, might we instead observe 0.51 or 0.49 or 0.5000000000000000001 or any of the other uncountably infinite values in the very near vicinity of 0.5? Each time that we make an observation the denominator of our proportion is growing, but if every value between 0 and 1 is truly possible, as time goes on the number of times that $X = 0.5$ must stay much, much smaller than the total

number of trials. If we continue this off to infinity, in the limit, the probability must become 0.
This conclusion leads to a few different points. First, impossibility is not the same as probability 0. Impossible events do have probability 0, but possible events may also have probability 0. Events which are outside of the sample space are impossible. Events inside the sample space, even probability zero events, remain possible. Second, we require alternative mathematical tools for discussing the probability of events in a continuous setting. Ideally this would be analogous to a probability mass function, but would somehow function in the case of continuity.

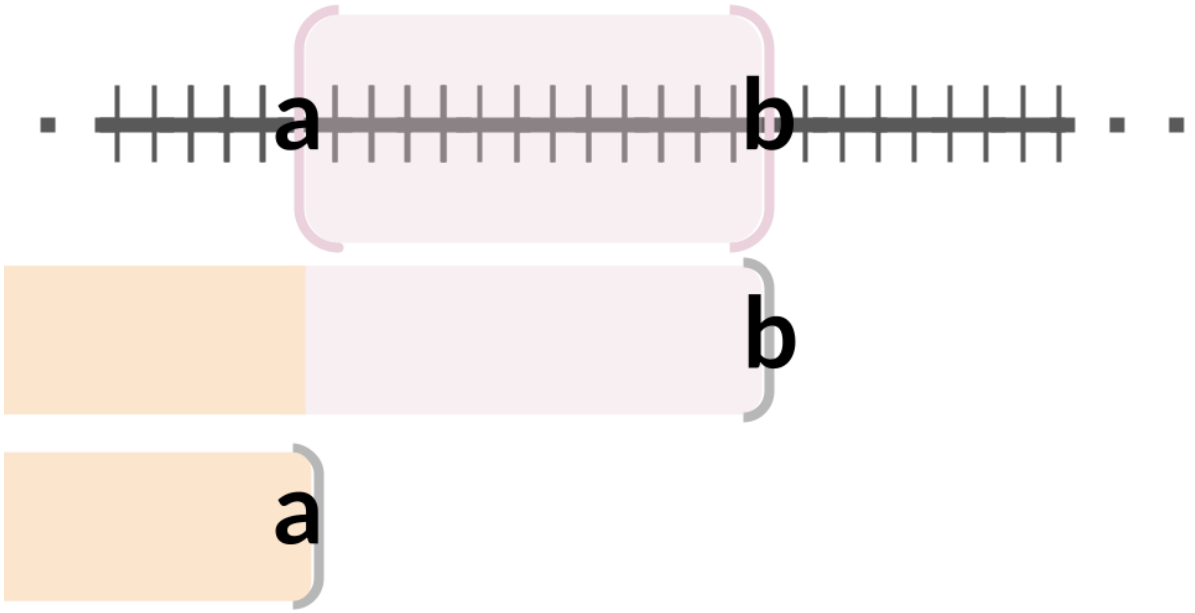
9.2 Cumulative Distribution Functions

To begin building to the continuous analogue of the probability mass function, we will start by focusing on events that are easier to define in the continuous case. Suppose that X is defined on some continuous interval. Instead of thinking of events relating to $X = x$, we instead turn our focus to events of the form $X \in (a, b)$ for some interval defined by the endpoints a and b. Now note that, relying only on our knowledge of probabilities relating to generic events, we can rewrite $P(X \in (a, b))$ slightly. Specifically,

$$\begin{aligned} P(X \in (a,b)) &= 1 - P(X \notin (a,b)) \\ &= 1 - P(\{X < a\} \cup \{X > b\}) \\ &= 1 - (P(X < a) + P(X > b)) \\ &= 1 - P(X > b) - P(X < a) \\ &= P(X < b) - P(X < a). \end{aligned}$$

[3] If it summed to 1 then this would not be a continuous random variable, since there would only be a countable number of possible events.
[3] Like, 1 in a million, or 1 in a billion, or 1 in the number of atoms in the universe.

Figure 9.1: A number line showcasing the interval (a, b) and demonstrating how you can take everything less than b and remove all that is less than a to be left with only (a, b).



In words we know that the probability that X falls into any particular interval is given by the probability that it is less than the upper bound of the interval minus the probability that it is less than the lower bound of the interval. Notice that $X < a$ is an event, and if we knew how to assign probabilities to $X < a$ for arbitrary a, then we could assign probabilities to any interval. Also note that, even in the continuous case, it make sense to talk of $P(X < a)$ for some value a. These intervals will contain an uncountably infinite number of events, and as such, can certainly occur with greater than 0 probability.

Consider an example with X defined on $[0, 1]$. In this case we know that $P(X < 1) = 1$. Note that we could have written $P(X \leq 1) = 1$, which may have been more obviously true. However, $P(X \leq 1) = P(\{X < 1\} \cup \{X = 1\}) = P(X < 1) + P(X = 1)$ and we know that $P(X = 1) = 0$. In the continuous case we do not need to worry whether we use $X \leq a$ or $X < a$, and we will interchange them throughout.

Example 9.1 (Charles and Sadie Wait for the Bus)**.** Charles and Sadie are visiting a large city and are getting around via public transit while there. They are not quite familiar with the bus schedules yet, but they know that the bus they are waiting for will show up sometime in the next 15 minutes. They realize that this is a continuous random quantity, and decide to pass the time by trying to reason about how long they will be waiting for the bus to arrive. To do so, they make the assumption that the bus is equally likely to show at any point during this interval.

a. What is the probability that Charles and Sadie wait 15 or fewer minutes for the bus?
b. What is the probability that Charles and Sadie wait 5 or fewer minutes for the bus?
c. What is the probability that Charles and Sadie wait exactly 7 minutes for the bus?
d. What is the probability that Charles and Sadie wait between 8 and 12 minutes for the bus?
e. What is the probability that Charles and Sadie have to wait longer than 13 minutes for the bus?

Solution

Let X represent the amount of time that they wait for the bus.

a. If we trust their judgment, then they have claimed that they will wait between 0 and 15 minutes for the bus. As a result, $P(X \leq 15) = 1$.
b. Here we can reason that if any time is equally likely for the bus to arrive, then the first 5 minutes must be as likely as the second interval of 5 minutes or the third. That is, in the first 5 minutes there is a total of $\frac{5}{15} = \frac{1}{3}$ of the possible arrival times taken up, and so we must conclude that $P(X \leq 5) = \frac{1}{3}$.
c. This would be $P(X = 7)$. X is a continuous quantity, and so $P(X = 7) = 0$.
d. We can reason through this in two different ways. First, we can use the property discussed above that $P(X \in (8, 12)) = P(X \leq 12) - P(X \leq 8)$, and then use the same procedure as in (b). That gives

$$P(X \in (8, 12)) = \frac{12}{15} - \frac{8}{15} = \frac{4}{15}.$$

Alternatively, because each interval of the same length must be equally likely, we can reason that this is just a length 4 interval. Thus, it must have the same likelihood as any other length 4 interval, given by $\frac{4}{15}$.
e. Note that, longer than 13 is the complement to less than 13. Thus,

$$P(X \geq 13) = 1 - P(X \leq 13) = 1 - \frac{13}{15} = \frac{2}{15}.$$

The centrality of events of the form $X \leq a$ prompts the definition of a mathematical function which we call the **cumulative distribution function**.

Definition 9.1 (Cumulative Distribution Function)**.** The cumulative distribution function of a random variable X, typically denoted as $F(x)$ or $F_X(x)$, is defined as the function that gives the probability that the random variable is less than or equal to some threshold. That is, $F_X(x) = P(X \leq x)$. We may also refer to the cumulative distribution function simply as the **distribution function**.

Once we have defined the distribution function for a random variable, using the above derivation we are able to determine the probability associated with any events based on intervals. It is worth noting that the cumulative distribution function can also be defined for discrete random variables. In the case of a discrete random variable, we would have

$$F_X(x) = \sum_{k \in \mathcal{X}; k \leq x} p_X(k).$$

Since it is simply the summation of the probability mass function it tends to be a less useful quantity.

Suppose that, for a random variable X, we know the cumulative distribution function. This knowledge permits the computation of *any* probability associated with the random variable. Consider some event defined in terms of X which we may wish to determine the probability of. We know that events are subsets of the sample space. Every one of these events can be written using our basic set operations (unions, intersections, and complements) applied to intervals of the form (a, b) and sets of the form $\{x\}$.[4] The axioms of probability allow us to compute probabilities across the set operations. Further, our knowledge of the cumulative distribution function, the conversion of $P(X \in (a, b))$ into $F_X(b) - F_X(a)$, and the fact that $P(X = x) = 0$ for all x gives all of the results we need to derive probabilities for these events.

Example 9.2 (Charles and Sadie's Light Bulbs)**.** Charles and Sadie decide that they need to replace some light bulbs. In their research online, a particular manufacturer of light bulbs, *Bayesian Brights*, lists the cumulative distribution function for the lifetime of their bulbs, in hours. The model Charles and Sadie are considering is said to have a lifetime (in hours) governed by

$$F_X(x) = 1 - \exp\left(-\frac{x}{10000}\right).$$

a. What is the probability that a purchased lightbulb lasts for less than 5000 hours?
b. What is the probability that a purchased lightbulb lasts for between 7500 and 12000 hours?
c. What is the probability that a purchased lightbulb lasts for more than 8000 hours?

Solution

a. Using the given cumulative distribution function this can be calculated as

$$P(X \leq 5000) = F(5000) = 1 - \exp\left(-\frac{5000}{10000}\right) = 1 - e^{-1/2} \approx 0.39346934.$$

[4] These are called singletons. We addressed these in terms of probability assignment using the language of singletons above. As a general rule, when working with continuous quantities, we simply ignore all of the singletons.

b. Here we get

$$P(X \in (7500, 12000)) = P(X \leq 12000) - P(X \leq 7500) = F(12000) - F(7500)$$
$$= \exp(-0.75) - \exp(-1.2) \approx 0.171172341.$$

c. We can convert this to be the complement event so that it is expressible in terms of the cumulative distribution function. That is, $P(X \geq 8000) = 1 - P(X \leq 8000)$, and so

$$P(X \geq 8000) = 1 - F(8000) = 1 - (1 - e^{-0.8}) = e^{-0.8} \approx 0.449328964.$$

9.3 The Probability Density Function

The distribution function will be the core object used to discuss the probabilistic behaviour of a continuous random variable. All of the behaviour of these random quantities will be described by the distribution function, and as such we will take the distribution function as a function which defines the distribution of a continuous random quantity. This is all that we need in order to analyze these random variables, however, it may be a little unsatisfying in contrast with the discrete case.

We had set out to find a quantity that paralleled the probability mass function. Instead, we concluded that the cumulative distribution function can be made to play the same role in terms of describing the behaviour of the random quantity. Still, it may be of interest for us to have a function which takes into account the *relative likelihood* of being near some value. Suppose, for instance, that for a random variable defined on $[0, 1]$ we wanted to know how likely it was to be in the vicinity of $X = 0.5$. We could take a small number, say $\delta = 0.01$ and calculate

$$P(X \in (0.5 - \delta, 0.5 + \delta)) = F(0.5 + \delta) - F(0.5 - \delta).$$

This is perfectly well defined based on our discussions to this point. Now, suppose that δ is small enough so that it is reasonable to assume that this probability is fairly evenly distributed throughout the interval. Then, if we wanted to assign a likelihood to each value, we could divide this total probability by the length of the interval, 2δ. As a result, in this case, the probability that X is nearly 0.5 will be approximately given by the expression

$$\frac{F(0.5 + \delta) - F(0.5 - \delta)}{2\delta}.$$

We had taken $\delta = 0.01$, but the same process could be applied for smaller and smaller δ, say 0.001 or 0.0001. Intuitively, as the size of this interval shrinks more and more we are getting a better and better estimate for the likelihood that the random variable is in the immediate vicinity of 0.5. Moreover, as δ gets smaller and smaller our assumption of a uniform

probability over the interval becomes more and more reasonable. Unfortunately, we cannot set $\delta = 0$, exactly. We can ask what happens *in the limit* as δ continues to get smaller and smaller. This question is in the purview of calculus, and can in fact be answered. While working out the answer is beyond the scope of the course, we will provide the result anyway.[5] The resulting function is called the **probability density function**, and is related to the cumulative distribution function through derivatives (and integrals).

Definition 9.2 (Probability Density Function)**.** A random variable X, with cumulative distribution function $F(x)$, is further characterized by its probability density function, denoted $f(x)$. The density function describes the relative likelihood of a random variable taking on values in a particular interval, and mirrors the behaviour of the probability mass function in the discrete case. Formally, the probability density function is equal to the derivative of the cumulative distribution function.

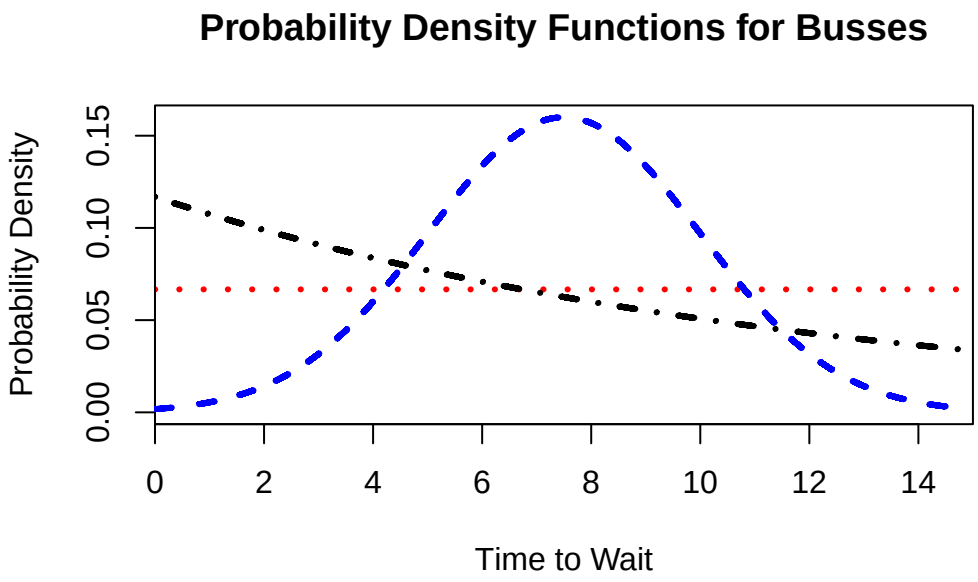
Roughly speaking, the density function evaluates how likely it is for a continuous quantity to be in a small neighbourhood of the given value. Critically, **probability density functions do not give probabilities directly**. In fact, probability density functions may give values that are greater than 1![6] Still, if we see the shape of the probability density function, we can state how likely it is to make observations near the results of interest. We will often graph the density functions. The high points of the graph indicate regions with more probability than the regions of the graph which are lower. Again, the specific probability of any event $X = x$ will always be 0, but some events fall in neighbourhoods which are more likely to observe than others.[7]

Example 9.3 (The Types of Busses in the City)**.** After returning from their trip to the big city, Charles and Sadie are talking to their friend Garth, with a lot more experience in the matter. Garth explains that there are actually several different types of busses, with different arrival schedules over the 15 minute interval. Knowing that Charles and Sadie have started to learn about probability density functions, Garth draws out the following sketch, giving the probability density function for three different busses.

[5] If you have taken calculus, you may recognize this process as feeling related to the first principle definition of a derivative. While it is often described in a slightly different way, this feeling of connection is intentional.

[6] This is not a rare occurrence either. For instance, if we consider a random variable that will take on values in the interval $[0, 0.5]$ with equal likelihood, the probability density function in this case will be $f(x) = 2$.

[7] Think about the fact that, while we may think of human adult heights as a continuous quantity, it is far more likely to observe a height around 5 feet 6 inches than around 9 feet 2 inches.



a. Which of the three buses is most likely to show up around the 1 minute mark?
b. Which of the three buses is most likely to show up around the 6 minute mark?
c. Which of the three buses is most likely to show up around the 14 minute mark?
d. Which bus is most likely to show up at 10 minutes exactly?
e. During which intervals would each of the buses be more likely than the other two to arrive?
f. Describe the behaviour of each of the three buses.

Solution

a. At $x = 1$, the black density is higher than the red density which is higher than the blue density. This gives the relative ranking that arrivals around 1 minute or more likely for the black bus than red or blue.
b. At $x = 6$, the blue density is higher than the red or black densities, which are roughly the same, and as such, it is the most likely bus to arrive in the vicinity of 6 minutes.
c. At 14, the red density is higher than the black density which is higher than the blue. This gives the relative ordering of likelihood of arrivals.
d. All buses have probability 0 of arriving at 10 minutes exactly since these are the densities for continuous random variables.
e. The black bus is most likely to arrive for times up until around 4 minutes, where the blue bus is then the most likely to arrive until around 11 minutes, where the red bus is most likely to arrive until the 15 minute mark.

f. The red bus has an equal probability of arriving across the full interval. The black bus is more likely to arrive near the start of the interval, slowly decaying in likelihood as time goes on, while remaining fairly likely at the 15 minute mark. The blue bus is very unlikely to take a very short period of time, or a very long period of time, and far more likely to sit in the middle of the possible interval.

9.4 Using Continuous Distributions

With the exception of the previously indicated differences, continuous and discrete random variables are treated similarly. The tools to analyze them differ[8], but the fundamentals remain the same. It is possible to compute expected values, medians, and modes, with roughly the same interpretations. It is possible to describe the range, interquartile range, and variance, with similar interpretations. The axioms of probability still underpin the manipulation and analysis of these random variables. The distinction is merely that in, place of elementary mathematics to complete the calculations, calculus is required.

Just as with discrete distributions, there are named continuous distributions. These are typically governed by either a density function or cumulative distribution function, alongside the expected value and variance. Just like the named discrete distributions, by matching the underlying scenario to the correct process we are able to avoid a lot of work in understanding the behaviour of the random quantities. Now, because calculus is not assumed knowledge, we will not work too widely with continuous random variables. We will introduce only two named continuous distributions: the uniform distribution[9] and the normal distribution.[10]

9.5 The Uniform Distribution

The uniform distribution[11] is parameterized over an interval specified as (a, b). On this interval, equal probability density is given to every event, which is to say that the density function is constant. Specifically, if $X \sim \text{Unif}(a, b)$ then

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{otherwise.} \end{cases}$$

[8] In the continuous case we cannot sum over the sample space, and so we must use techniques from calculus to mirror this process, for instance.

[9] Which we have actually already seen.

[10] Which is far and away the most important distribution (discrete or continuous) in all of probability and statistics.

[11] Sometimes called the continuous uniform distribution to distinguish it from its discrete counterpart.

From the density function we can find that

$$F(x) = \frac{x-a}{b-a}$$

for $x \in (a, b)$, with $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x > b$. Moreover, we have $E[X] = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a)^2}{12}$.

Example 9.4 (Characterizing the Wait Time)**.** Still considering the wait time for buses in the city, Sadie points out that the bus they were waiting for follows a Unif$(0, 15)$ distribution. They can use their newfound wisdom to make deeper conclusions about the process![12]

a. How long should they expect to wait for the bus?
b. What is the variance for the amount of time that they will be waiting?

Solution

Here, we know that $X \sim \text{Unif}(0, 15)$.

a. $E[X]$ is $\dfrac{a+b}{2} = \dfrac{0+15}{2} = 7.5$, so they should expect to wait 7.5 minutes.
b. The variance is $\text{var}(X) = \dfrac{(15-0)^2}{12} = 18.75$.

The uniform distribution is analogous to the discrete uniform. Any time there is an interval of possible outcomes which are all equally likely, the uniform distribution is the distribution to use. Compared with other distributions it is also fairly straightforward to work with, which makes it a useful demonstration of the concepts relating the continuous probability calculations.[13]

9.6 The Normal Distribution

The normal distribution, also sometimes referred to as the Gaussian distribution, is a named continuous distribution function defined on the complete real line. The distribution is far and away the most prominently used distribution in all of probability and statistics. In fact, most people have heard of normal distributions even if they are not aware of this fact. Any time that there is a discussion of a bell curve, for instance, this is in reference to the normal distribution. Normally distributed quantities arise all over the place from measurements of heights, grades, or reaction times through to levels of job satisfaction, reading ability, or blood pressure. There

[12] In Example 9.1 we implicitly used the cumulative distribution function of the uniform. It is worth revisiting these probabilities *knowing* that this is the case.
[13] For instance, in Example 8.6, we implicitly used a continuous uniform distribution to give the probability that Charles hits the dartboard. This type of reasoning is very prevalent.

is a tremendous number of normally distributed phenomena naturally occurring in the world, which renders the normal distribution deeply important across a wide range of domains.

Perhaps more important than the places where the normal distribution arises in nature are the places where it arises mathematically. At the end of this course we will see a result, *the central limit theorem*, which is one of the core results in all of statistics. Much of the statistical theory that drives scientific inquiry sits atop the central limit theorem. And at the core of the central limit theorem is the normal distribution. It is virtually impossible to overstate the importance of the normal distribution, and as a result, it is worthy of investigation.

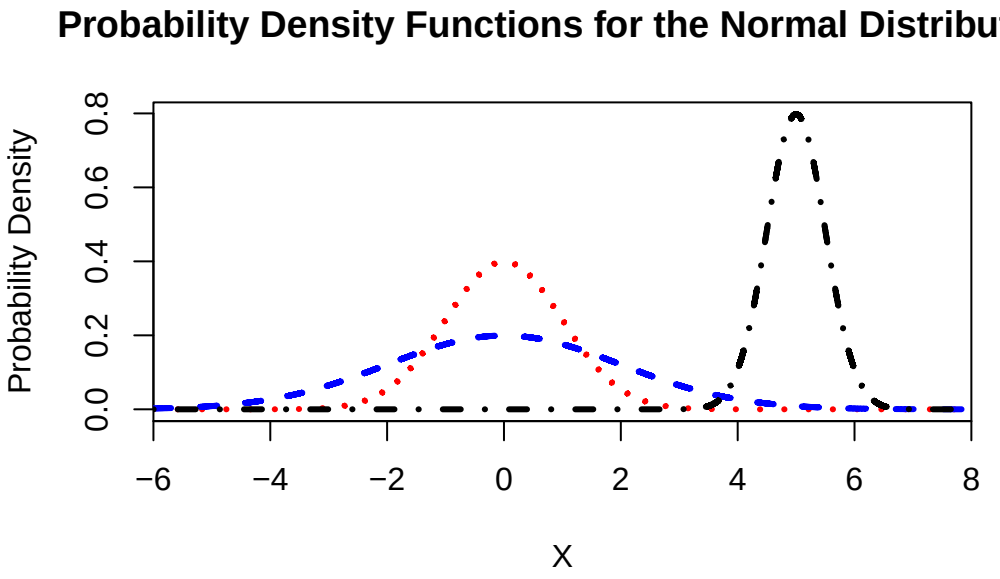
9.6.1 The Specification of the Distribution

A normal distribution is parameterized by two parameters: the mean, μ, and the variance σ^2. We write $X \sim N(\mu, \sigma^2)$. These parameters directly correspond to the relevant quantities such that $E[X] = \mu$ and $\text{var}(X) = \sigma^2$. The density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This can be quite unwieldy to work with, however, when it is plotted we see that the normal distribution takes on a bell curve which is centered at μ.

Figure 9.2: Three normal distribution densities plotted with mean 0 (red and blue), and mean 5 (black). The red density has variance 1, the blue density has variance 4, and the black density has variance 0.25.



9.6.2 The Standard Normal Distribution

Normally distributed random variables are particularly well-behaved. One way in which this is true is that if you multiply a normally distributed random variable by a constant, it will remain normally distributed, and if you add a constant to a normally distributed random variable, it will remain normally distributed. Consider, for $X \sim N(\mu, \sigma^2)$, taking the $X - \mu$. From our discussions of expected values we know that $E[X-\mu] = E[X]-\mu = 0$. Furthermore, adding or subtracting a constant will not change the variance. Thus, $X - \mu \sim N(0, \sigma^2)$.

Now, consider dividing this by σ, or equivalently, multiply by $\frac{1}{\sigma}$. The expected value of the new quantity will be $\frac{1}{\sigma} \times 0 = 0$, and, from our discussions regarding the variance of linear transformations, the variance of the new quantity will be $\frac{1}{\sigma^2} \times \sigma^2 = 1$. Taken together then, if $X \sim N(\mu, \sigma^2)$,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

This holds true for *any* normal distribution with *any* mean or variance values. This straightforward transformation allows us to discuss normal distributions in terms of $N(0, 1)$. We call this

the **standard normal distribution**, and will typically use Z to denote a random variable from the standard normal distribution.

Definition 9.3 (Standard Normal Distribution)**.** The standard normal distribution is the version of the normal distribution with mean 0 and variance 1. We say that Z follows a standard normal, and write $Z \sim N(0,1)$, if the density of Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

We denote the density of Z as $\varphi(z)$, and the cumulative distribution function of Z as $\Phi(z)$. The cumulative distribution function, $\Phi(z)$, does not have a nice form to be written down, however, it is a commonly applied enough function that many computing languages have implemented it, including of course R.[14]

The utility in this process of converting normally distributed random variables to be standard normal random variables, a process known as **standardization**, is demonstrated by realizing that events can be converted using the same transformations. Specifically, suppose we have $X \sim N(\mu, \sigma^2)$, and we want to find $P(X \leq x)$. Note that, $X \leq x$ must also mean that

$$\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma},$$

through an application of the same transformation to both sides. But we *know* that the left hand side of this inequality is exactly Z, a standard normal random variable with cumulative distribution function $\Phi(z)$. Thus,

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Using this trick of standardization any normal probability can be converted into a probability regarding the standard normal.

Example 9.5 (Charles and Sadie Explore House Plants)**.** Charles and Sadie learn that the heights of fully grown house plants are often normally distributed. They are exploring which plants will work best in their new apartment, but are finding it difficult to reason about them in comparison to one another. Some plants end up being taller on average with a lot more variability, while others may be shorter but far more certain. Sadie realizes that the cumulative distribution function of a standard normal, $\Phi(z)$, can be evaluated using their smart phones. As a result, they get to work expressing different probabilities in terms of $\Phi(z)$.

They are comparing four different plant species. These species have heights according to the following random variables.

[14] See Section 9.6.3 for more details on this.

i. Plant A has height X, which follows a normal distribution with mean 90 and variance 100.
ii. Plant B has height Y, which follows a $N(110, 400)$ distribution.
iii. Plant C has height V, which follows a normal distribution centered on 70 with standard deviation 7.
iv. Plant D has height W, which follows a normal distribution with $\mu = 85$ and $\sigma^2 = 81$.

Express each of the following probabilities in terms of $\Phi(z)$, the cumulative distribution function of a standard normal random variable.

a. One spot the plants cannot be too tall. What is the probability that they can fit plant A into a spot which cannot accommodate plants taller than 80cm?
b. A second spot would be strange to have too short of a plant in. If they require the plant to be at least 125cm, what is the probability they can use plant B?
c. A third spot can accommodate plants that are either under 60cm if they use a stand, or else over 90cm. How likely is it that plant C will work in this spot?
d. A fourth spot requires a plant that is somewhere between 80cm and 90cm. What is the probability that plant D will work?

Solution

a. Here we wish to standard a $N(90, 100)$ random variable. To do this we note that

$$\frac{X - 90}{\sqrt{100}} = \frac{X - 90}{10} \sim N(0, 1).$$

Thus, taking $Z \sim N(0, 1)$, we know that

$$P(X \leq 80) = P\left(\frac{X - 90}{10} \leq \frac{80 - 90}{10}\right) = P(Z \leq -1) = \Phi(-1).$$

b. Here we wish to standard a $N(110, 400)$ random variable. To do this we note that

$$\frac{Y - 110}{\sqrt{400}} = \frac{Y - 110}{20} \sim N(0, 1).$$

Thus, taking $Z \sim N(0, 1)$, we know that

$$P(Y \geq 125) = P\left(\frac{Y - 110}{20} \geq \frac{125 - 110}{20}\right) = P(Z \geq \frac{3}{4}) = 1 - \Phi(0.75).$$

c. Here we wish to standard a $N(70, 49)$ random variable. To do this we note that

$$\frac{V - 70}{7} \sim N(0, 1).$$

Thus, taking $Z \sim N(0,1)$, we know that

$$P(V \leq 60) = P\left(\frac{V-70}{7} \leq \frac{60-70}{7}\right) = P(Z \leq -\frac{10}{7}) = \Phi(-\frac{10}{7}).$$

Moreover, we know that

$$P(V \geq 90) = P\left(\frac{V-70}{7} \geq \frac{90-70}{7}\right) = P(Z \geq \frac{20}{7}) = 1 - \Phi(\frac{20}{7}).$$

Thus, the probability this will be acceptable will be $1 - \Phi(\frac{20}{7}) + \Phi(-\frac{10}{7})$.

d. Here we wish to standard a $N(85, 81)$ random variable. To do this we note that

$$\frac{W-85}{\sqrt{81}} = \frac{W-85}{9} \sim N(0,1).$$

Thus, taking $Z \sim N(0,1)$, we know that

$$\begin{aligned} P(80 \leq W \leq 90) &= P\left(\frac{80-85}{9} \leq \frac{W-85}{9} \leq \frac{90-85}{9}\right) \\ &= P(Z \leq \frac{5}{9}) - P(Z \leq -\frac{5}{9}) = \Phi(\frac{5}{9}) - \Phi(-\frac{5}{9}). \end{aligned}$$

As a result, combining our knowledge of continuous random variables, with the process of standardization we are able to calculate normal probabilities for any events relating to normally distributed random quantities. Moreover, since the shape of the normal distribution is so predictable, it is often easy to draw out the density function, and indicate on this graphic the probabilities of interest, which in turn helps with the required probability calculations. Calculating probabilities from normal distributions remains a central component of working with statistics and probabilities beyond this course. Developing the skills and intuition at this point, through repeated practice is a key step in successfully navigating statistics here and beyond.

When you have access to a computer, and your interest is in calculating a normal probability, as described above, there is not typically a need for standardization. However, it remains an important skill for several reasons. First, by always working with the same normal distribution, you will develop a much more refined intuition for the likelihoods of different events. It goes beyond working with the same family of distributions, you get very used to working with exactly the same distribution. Second, you will likely become quite familiar with certain key **critical values** of the standard normal distribution. These values arise frequently, and allow you to quickly approximate the likelihood of different events. Finally, as we begin to move away from studying probability and into studying statistics, the standard normal will feature prominently there.

9.6.3 Normal Probability Calculations in R

Just as with the named discrete distributions, R provides functions for calculating probabilities related to normal random variables. The relevant functions are the `dnorm` and `pnorm` function, for the density function and the cumulative distribution function, respectively. These functions can take arguments for the mean and standard deviation[15] of the normal. If not provided, it will default to the standard normal.

```
# N(0, 1)
dnorm(0.25) # f(0.25)
pnorm(0.25) # P(X <= 0.25)

# N(5, 4)
dnorm(4.9, mean = 5, sd = sqrt(4)) # f(4.9)
pnorm(4.9, mean = 5, sd = sqrt(4)) # P(X <= 4.9)

# N(0, 1) - Explicitly
dnorm(1, mean = 0, sd = 1) # f(1)
pnorm(1, mean = 0, sd = 1) # P(X <= 1)
## [1] 0.3866681
## [1] 0.5987063
## [1] 0.199222
## [1] 0.4800612
## [1] 0.2419707
## [1] 0.8413447
```

9.6.4 The Empirical Rule and Chebyshev's Inequality

Another way in which the normal distribution is well behaved is summarized in the **emprical rule**. The shape of the distribution is such that, no matter the specific mean or variance, all members of the family remain quite similar. This enables the derivation of an easy, approximate result, to help intuitively gauge the probabilities of normal events.

The Empirical Rule

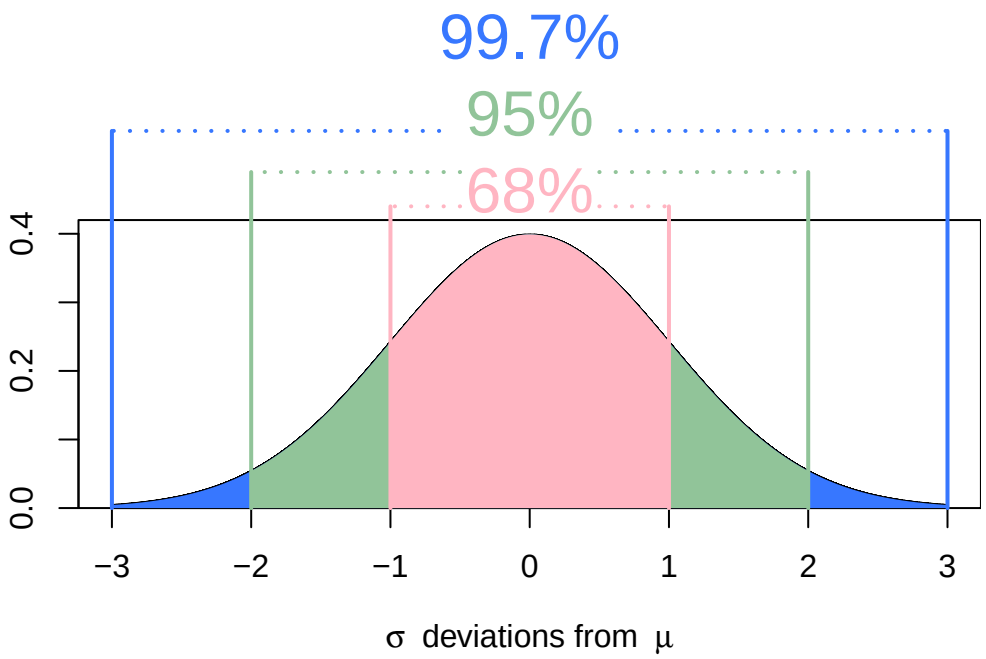
The empirical rule is a mathematical result regarding the probability of a normally distributed random variable. If X has a normal distribution with mean μ and variance σ^2, then:

1. The probability of observing a value within σ of the mean is approximately 0.68;
2. The probability of observing a value within 2σ of the mean is approximately 0.95;

[15] Recall, the standard deviation is the square root of the variance.

and

3. The probability of observing a value within 3σ of the mean is approximately 0.997.



In words, the empirical states that almost all of the observations from a normal distribution will fall in the interval $\mu \pm 3\sigma$. In mathematical terms, the empirical rule is summarized as

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 0.68 \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 0.95 \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 0.997. \end{aligned}$$

With the standard normal we can replace μ with 0, and σ with 1 to get a version which is slightly more concise to state. It is then possible to combine these different intervals by recognizing the symmetry in the normal distribution. That is, $P(\mu \leq X \leq \mu + \sigma) \approx \dfrac{0.68}{2} = 0.34$.

Example 9.6 (Charles and Sadie Can Never Have Enough House Plants)**.** Their standardization efforts paid off, and Charles and Sadie found some plants that should fit the spaces that they need once they've grown up. Because of how well the plants worked out, they wanted to buy some more. They end up back at the store, and they know that they need to get a plant that will grow to be between 70cm and 106cm. They have several options for plants again:

i. A plant with heights X according to $N(88, 36)$.
ii. A plant with heights Y according to $N(124, 324)$.

iii. A plant with heights W according to $N(82, 144)$.

Unfortunately, Charles and Sadie forget their phones at home and so they cannot make direct calculations using $\Phi(z)$. Can you help them, without using a normal probability calculator, determine which plant has the highest probability of being acceptable?

Solution

Without access to $\Phi(z)$ we can instead make use of the empirical rule. Consider that for X we have $\mu_X = 88$ and $\sigma = \sqrt{36} = 6$.

1. If we take the interval $[70, 106]$ then we can consider the terms
$$\frac{70-88}{6} = \frac{-18}{6} = -3 \quad \text{and} \quad \frac{106-88}{6} = 3.$$
Thus, $P(70 \leq X \leq 106) = P(\mu_X - 3\sigma_X \leq X \leq \mu_X + 3\sigma_X)$. Using the empirical rule we know that this probability is approximately 0.997.
2. We can approach the same tactics for Y. This time, we get
$$\frac{70-124}{\sqrt{324}} = -3 \quad \text{and} \quad \frac{106-124}{18} = -1$$
. Thus, $P(70 \leq Y \leq 106) = P(\mu_Y - 3\sigma_Y \leq Y \leq \mu_Y - \sigma_Y)$. Note that from μ to $\mu - 3\sigma$, according the empirical rule, there is $\dfrac{0.997}{2} = 0.4985$ probability. Similarly, according to the empirical rule there is $\dfrac{0.68}{2} = 0.34$ probability of being between μ and $\mu - \sigma$. Thus, if we take $0.4985 - 0.34 = 0.1585$, this gives the probability of being between $\mu - 3\sigma$ and $\mu - \sigma$, as required.
3. As above,
$$\frac{70-82}{\sqrt{144}} = -1 \quad \text{and} \quad \frac{106-82}{12} = 2$$
. As a result, $P(70 \leq W \leq 106) = P(\mu_W - \sigma_W \leq W \leq \mu_W + 2\sigma_W)$. Note that as in (2) there is 0.34 probability of being between μ and $\mu - \sigma$. To be between μ and $\mu + 2\sigma$ there will be probability $\dfrac{0.95}{2} = 0.475$. As a result, the total probability here will be $0.34 + 0.475 = 0.815$.

As a result, the first plant has a 0.997 chance to fit, the second a 0.1585, and the third a 0.815 (approximately).

The empirical rule is not exact, and when computing probabilities with access to statistical software it is likely of limited direct utility. However, it is another tool to leverage to continue refining your intuition for the behaviour of random quantities. It is also a good "check" to have, giving an immediate sense of the likelihood of different events. If you compute an answer

which seems to contradict the empirical rule, take a second look. If you have someone tell you that they have observed events which are out of line with the empirical rule, be skeptical.

The empirical rule is a useful result to aid in building intuition regarding the normal distribution. However, when quantities are not normally distributed, it does not apply. A related, though somewhat weaker result is Chebyshev's Inequality. This will hold for *any* distribution, and can be seen as a useful extension to the empirical rule.

Chebyshev's Inequality

Chebyshev's Inequality provides probabilistic bounds on the likelihood of deviating from the mean for any random variable. In words, there is a probability of 0.75 or more of observing an observation within two standard deviations, and a probability of at least 0.8889 of observing a value within three standard deviations of the mean. Formally, for any $k > 0$,

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

Here k can be any real number which is greater than 0. If $k \leq 1$, this result is uninteresting since the bound simply is 0. However, taking $k = 2$ gives the 0.75 lower bound outlined above, which is a more useful result. Additionally, there is no requirement for k to be an integer here, and so, for instance, the probability of observing a value within $\mu \pm \sqrt{2}\sigma$ is at least 0.5, for all distributions.

Example 9.7 (Fitting in the Strange Plants)**.** Charles and Sadie selected a plant to fit in the spot requiring one between 70 and 106cm. However, as they are checking out at the store they see a plant that they like much better. They are conflicted since they do not know whether this plant will satisfy normality assumptions in its height. The worker tells them that on average the plant grows to be 88cm, and figures that the variance will be approximately 51.84. Charles and Sadie want to be at least 90% certain that the plant will fit. If they are, they will buy it!

a. If they assume that the plants heights are normally distributed, should they buy the plant?
b. If they do not assume that the plants heights are normally distributed, should they buy the plant?

Solution

a. Note that in this case, we have a mean of 88 and a standard deviation of 7.2. As a result, we have that 70 is $\mu - 2.5\sigma$ and 106 is $\mu + 2.5\sigma$. The empirical rule does not tell us directly how to address probabilities in this case, however, we know that

   $$P(X \in [\mu - 2\sigma, \mu + 2\sigma]) \leq P(X \in [\mu - 2.5\sigma, \mu + 2.5\sigma]) \leq P(X \in [\mu - 3\sigma, \mu + 3\sigma]).$$

   According to the empirical rule then we can say that, if this plant follows a normal

distribution, the probability it is accessible will be between 0.95 and 0.997. If the normal distribution is a reasonable assumption, then they should buy the plant.

b. If the distribution is non-normal, then we can instead apply Chebyshev's inequality. Here we have $k = 2.5$, as solved for in (a). As a result,

$$P(\mu - 2.5\sigma \leq X \leq \mu + 2.5\sigma) \geq 1 - \frac{1}{(2.5)^2} = 0.84.$$

As a result, if they want to be 90% certain, without knowing more about the distribution they should *not* purchase the plant.

9.7 Closure of the Normal Distribution

We have seen a certain type of *closure property* for the normal distribution when we discussed standardization. That is, adding and multiplying by constants does not change the distribution when working with normally distributed quantities. This is an interesting property which does not hold for most distributions, and makes normally distributed random variables quite nice to work with. The normal distribution has an additional type of closure property which is frequently used.

Suppose that X and Y are independent, with $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$. In this setting,

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

That is to say, the addition of two independent normally distributed random variables will also be normally distributed. This extends beyond two in the natural way, simply by applying and reapplying the rule (as many times as is required).

Closure of the Normal Distribution

Suppose that $X_1, \ldots, X_n$ are all independent, with each $X_i \sim N(\mu_i, \sigma_i^2)$. Then the summation

$$\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2),$$

where $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$. Notably, if $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then the summation will be $N(n\mu, n\sigma^2)$.

If instead of considering the summation, we consider the average of n independent and identically distributed $N(\mu, \sigma^2)$ variables, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}).$$

This follows from an application of our standard expectation and variance transformation rules. This type of result is central to the practice of statistics, and this closure under addition further aids in the utility of the normal distribution.

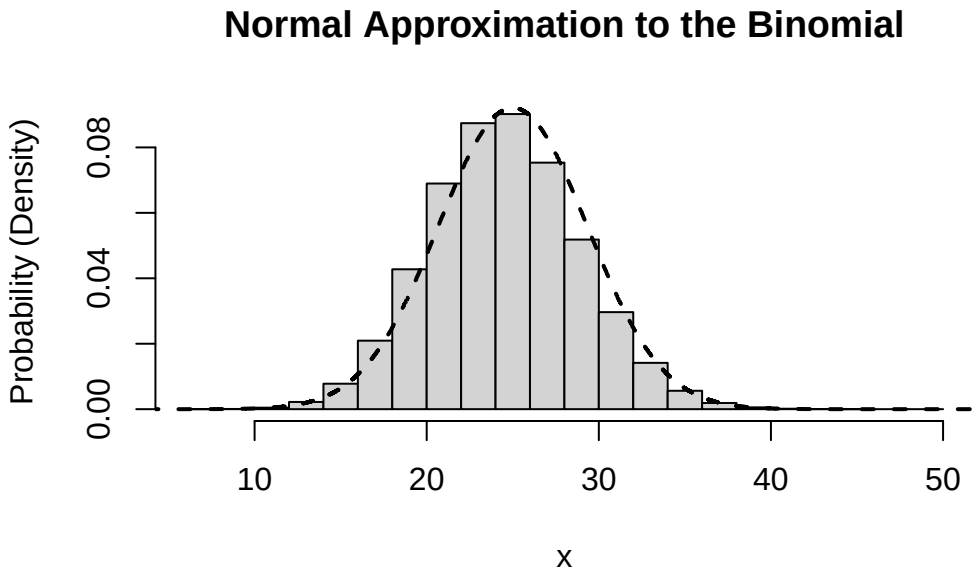
9.8 Approximations Using the Normal Distribution

A final utility to the normal distribution is in its ability to approximate other distributions. While several of these approximations exist, we will focus on the normal approximation to the binomial as an illustrative example. Historically, these approximations were critical for computing probabilities by hand in a timely fashion. Owing to the widespread use of statistical software, these use cases are more and more limited. However, there are two major advantages to learning these approximations. First, with an approximation it becomes easier to leverage the intuition you will build regarding the normal distribution in order to better understand the behaviour of other random quantities. Second, the normal approximation has the same *flavour* as many results in statistics, and so it presents an additional path to familiarity with these types of findings.

Suppose that $X \sim \text{Bin}(n, p)$. Through knowledge of the binomial distribution, we know that $E[X] = np$ and $\text{var}(X) = np(1-p)$. If n is sufficiently large then it is possible to approximate a binomial distribution using a normal distribution with the corresponding mean and variance. That is, for n large enough, we can take $X \sim \text{Bin}(n, p)$ to have approximately the same distribution as

$$W \sim N(np, np(1-p)).$$

Figure 9.3: A plot showing the probability mass function of a $\text{Bin}(100, 0.25)$ distribution, with a $N(25, 18.75)$ density overlaid, demonstrating the utility of the approximation.



One consideration that we need to make when applying this approximation has to do with the fact that the normal distribution is continuous while the binomial distribution is discrete. As a result, the normal distribution can take on any value on the real line, where the binomial is limited to the integers. A question that we must answer is what to do with the non-integer valued numbers. The natural solution is to rely on rounding. That is, for any value between $[1.5, 2.5)$ we would round to the nearest integer, which is 2.

Definition 9.4 (Continuity Correction)**.** The continuity correction is a technique for adjusting the probabilities computed using a continuous approximation to a discrete random variable. The correction relies on rounding non-integer values, which may be observed with regards to the continuous random variable, to the corresponding integer values for the discrete random variable that is being approximated.

This natural solution is in fact a fairly useful technique, and it is the one that we will make use of in the normal approximation. While rounding is quite natural, the process for leveraging this idea in probability approximation is somewhat backwards. That is, we typically will need to go from probabilities relating to X and transform those into probabilities relating to W. So, for instance, if we wish to know $P(X \leq 2)$, then we need to be able to make this a statement regarding the random variable W. In order to do this we need to ask "what is the largest value for W that would get rounded to 2?" The answer is 2.5 and so $P(X \leq 2) \approx P(W \leq 2.5)$.

A similar adjustment would be required if we instead wanted $P(X \geq 5)$. Here we would ask "what is the smallest value for W which would get rounded to 5?" and note that the answer is 4.5. Thus, $P(X \geq 5) \approx P(W \geq 4.5)$. Once we have expressed the probability of interest in terms of the normal random variable, we can use the standard techniques previously outlined to compute the relevant probabilities. It is very important to note that these two results are of the form $X \geq x$ and $X \leq x'$. If we instead had considered $X > x$ or $X < x'$, we would need to take an additional step.

For continuous random variables whether $X \geq x$ or $X > x$ is considered makes no difference. However, for discrete random variables this is not the case. As a result we should first convert the event the an equivalent event which contains the equality sign within the inequality, and then apply the continuity correction. That is, if we want $P(X > 3)$ first note that for $X > 3$ to hold, we could equivalently write this as $X \geq 4$. Alternatively, if the event of interest is $X < 8$, this is the same as $X \leq 7$.

Example 9.8 (Charles and Sadie Payouts over a Year)**.** Charles and Sadie are back sitting at the coffee shop, reflecting on all of the probability that they have learned since their games began. They realize that each time they play the game to see who will pay, that is a Bernoulli trial with 0.5 probability. As a result, over the course of the year, if they go 200 times to get coffee, the number of times each of them will have to pay is governed by a $\text{Bin}(200, 0.5)$ random variable. Charles realizes that this is infeasible to work with a binomial distribution, and so seeks another way.[16] Sadie suggest that they could use the normal approximation to the binomial.

a. What is the approximate probability that Charles pays more than 115 times in a year, expressed in terms of Φ.
b. What is the approximate probability that Sadie will pay between 86 and 107 times? Explain how this can be approximated numerically.
c. Give an upper and lower bound that both Charles and Sadie can be 95% sure they will pay between (that is, two numbers such that the probability they pay at least the lower and at most the upper bound is 0.95).

Solution

For these questions we will take $X \sim \text{Bin}(200, 0.5)$, and therefore, $X \tilde{\sim} W \sim N(100, 50)$. Thus, to calculate these probabilities, we can use normal approximations, making sure to apply the continuity corrections.

a. We want $P(X > 115)$. This is the same as asking $P(X \geq 116)$, and with the

[16] Note that, for instance, $\binom{200}{75} = 168849997346404286704489530268603459022868706883102845056$. This is 168 septendecillion. This is 3.2 million times more than the number of possible arrangements of a chess board. This is a silly large number.

continuity correction, $P(X \geq 116) \approx P(W \geq 115.5)$. Thus, we take

$$P(W \geq 115.5) = P\left(Z \geq \frac{115.5 - 100}{\sqrt{50}}\right) = 1 - \Phi\left(2.192\right).$$

This could be worked out using a normal probability calculator.[17]

b. We want

$$P(86 < X < 107) = P(87 \leq X \leq 106 \approx P(87.5 \leq W \leq 106.5).$$

Note that if we consider

$$\frac{106.5 - 100}{\sqrt{50}} = 0.919 \quad \text{and} \quad \frac{87.5 - 100}{\sqrt{50}} = -1.838.$$

If we wanted to get a rough approximation of this, we could say that, from the empirical rule, the probability that

$$\begin{aligned} P(\mu_W - 1.838\sigma_W \leq W \leq \mu_W + 0.919\sigma_W) &\leq P(\mu_W - 2\sigma_W \leq W \leq \mu_W + \sigma_W) \\ &= \frac{0.68}{2} + \frac{0.95}{2} = 0.815. \end{aligned}$$

Note that this gives an upper bound on the probability. We could also get a lower bound on the probability by moving in the other directions,

$$P(\mu_W - 1.838\sigma_W \leq W \leq \mu_W + 0.919\sigma_W) \leq P(\mu_W - 2\sigma_W \leq W \leq \mu_W) = \frac{0.95}{2} = 0.475.$$

This gives a fairly wide range, but we can be relatively confident it will be a lot closer to 0.815 than to 0.475, since the values are a lot close to -2 and 1 than -1 and 0.[18]

c. Note that, if we use a normal approximation and the empirical rule, we know that there is approximately a 0.95 probability that a random variable falls within 2σ of the mean. The mean of W is 100 and $\sigma = \sqrt{50}$, so we can take the interval $[85.8578, 114.1421]$. Note that if we widen this slightly the probability will increase slightly, so that the probability $W \in [85.5, 114.5]$ must be greater than 0.95. The quantity $P(85.5 \leq W \leq 114.5)$ corresponds to $P(86 \leq X \leq 114)$, and so we can say that Charles and Sadie can each be roughly 95% confident that they will need to pay between 86 and 114 times.[19]

When it is not necessary, it rarely makes sense to use an approximation. There will be cases

[19] Doing so results in 0.0141898. Note, if we had used a binomial probability calculator instead, we would have gotten 0.0140627.

[19] In fact, the actual probability here is 0.7824647.

[19] The actual binomial probabilities here are 0.9519985, while the normal approximation is 0.959695.

where the approximation is directly useful, and in those moments it is great to be able to use it. This example of using the normal distribution to approximate a discrete random variable serves as a nice bridge from the study of probability to the study of statistics. In statistics we take a different view of the types of problems we have been considering to date, and we require the tools of probability that have been brought forth. As a result, a deep comfort with manipulating probability expressions is required to build a strong foundation while studying statistics.

Exercises

Exercise 9.1. A boy is trying to climb a slippery pole and finds that he can climb to a height of at least 1.850 m once in 5 attempts, and to a height of at least 1.700 m nine times out of ten. Assuming that the heights he reaches form a normal distribution:

a. What is the mean and standard deviation of the distribution?
b. If he climbs the pole 1000 times, what height can he expect to exceed once? Express your answer in terms of $\Phi(z)$.

Exercise 9.2. A machine produces widgets of which an average of 10% are defective.

a. Find an approximate value for the probability that a random sample of 500 of these articles contains more than 25 which are defective.
b. What, approximately, is the probability that the sample contains fewer than 60 defectives?

Exercise 9.3. The mean inside diameter of a sample of 200 washers produced by a machine is 0.502 inches with a standard deviation of 0.005 inches. The purpose for which these washers are intended allows for a maximum tolerance in the diameter of 0.496 to 0.508. If we assume that the washer diameters are normally distributed, what is the probability that a washer is defective?

Exercise 9.4. The wingspans of the females of a certain species of bird of prey form a normal distribution with mean 168.75cm and a standard deviation of 6.5cm. The wingspans of males of the species are normally distributed with mean 162.5 and standard deviation of 6. What is the probability if, with a male and female selected at random, the male has the longer wingspan?

Exercise 9.5.

a. Find the probability of getting between 3 and 6 heads inclusive in 10 tosses of a fair coin.
b. Approximate the same probability using the normal approximation. How close is the approximation?

Exercise 9.6. Suppose that the cumulative distribution function of T is

$$f(t) = 1 - e^{-0.45t}.$$

a. Find $P(T > 3)$.
b. Find the median of T.

Exercise 9.7. Suppose that for a random variable, X,

$$F(x) = \frac{x(x^2 + 9x + 27)}{(x + 3)^3}.$$

a. What is the probability X falls between 1 and 3?
b. What is the median of X?
c. What is $\zeta(0.3)$ for X?

Exercise 9.8. Resistors labeled 100Ω have true resistances that uniformly fall between 80Ω and 120Ω. Let X be the mass of a randomly chosen resistor.

a. What is the probability that a resistor has resistance equal to 90Ω.
b. What proportion of resistors have resistance less than 90Ω?
c. Find the mean resistance.
d. Find the variance of the resistances.
e. Find the probability that the resistance is less than kΩ, for arbitrary k.
f. Find the median resistance.

Exercise 9.9. Suppose that a random variable X is defined on $[1, \infty)$. We know that $E[X] = 5$ and $\text{var}(X) = 3$.

a. Give a bound on $P(X \geq 10)$.
b. Find a value, a, such that $P(X a).

Exercise 9.10. Suppose that X is drawn from a Unif$(-8, 2)$ distribution.

a. What is $P(-6 \leq X \leq 0)$?
b. Approximate this probability using Chebyshev's Inequality. How close are the two results?

Exercise 9.11. Suppose that lifespans of tortoises are normally distributed with a mean of 100 and variance of 81.

a. What is the probability that a tortoise lives between 91 and 118 years?
b. What is the probability that a tortoise lives less than 100 years?
c. What is the probability that a tortoise lives less than 127 years?

d. What is the probability that a tortoise lives longer than 109 years?

Exercise 9.12. The heights of students in a class follow a normal distribution with a mean of 65 inches and a standard deviation of 4 inches.

a. Express the probability that a students is between 57 and 73 inches tall in terms of $\Phi(z)$.
b. Estimate the probability from (a).

Exercise 9.13. A factory produces light bulbs with a mean lifespan of 1000 hours and a standard deviation of 50 hours. Suppose the lifespan of the bulbs follows a normal distribution.

a. What percentage of bulbs can be expected to last between 950 and 1100 hours? Express the probability in terms of $\Phi(z)$.
b. Estimate the probability from (a).

Exercise 9.14. A farmer grows apples with a mean weight of 150 grams and a standard deviation of 20 grams. Suppose the weights follow a normal distribution.

a. What percentage of apples weigh between 110 and 130 grams? Express the probability in terms of $\Phi(z)$.
b. Estimate the probability from (a).

Exercise 9.15. A survey indicates that the average monthly income for employees in a company is \$3000 with a standard deviation of \$500. Suppose the incomes follow a normal distribution.

a. What percentage of employees earn between \$3000 and \$3500 per month? Express the probability in terms of $\Phi(z)$.
b. Estimate the probability from (a).

Part II

Part 2: Statistics

10 Introduction to Statistics

10.1 From Probability to Statistics

Until this point we have focused on the study of probability. At its core, probability is a subject which seeks to quantify the uncertainty present in statistical experiments. In the study of probability we begin by first making assumptions about the state of the world[1] and from there we draw conclusions about what must be true about the state of uncertainty in the world. In this regard, probability answers questions of the form "if this is true about the world, what should we see?" For instance, if a fair coin is tossed 100 times, what is the likelihood that more than half of the tosses come up heads? We have made the assumption that we have a fair coin, tossed independently, and we wish to quantify our degree of uncertainty about this scenario.

This is not the only way that we could frame problems related to uncertainty. For instance, what if we asked "given the results from 100 tosses of this coin, do we believe that the coin is fair?" This inversion of the previous question starts from the observation of information from an experiment and asks questions about the underlying mechanisms that generated these data.[2] This type of question is addressed by the field of **statistics**.

Definition 10.1 (Statistics (Field of Study))**.** Statistics is the discipline in which data are collected, analyzed, and presented with the goal of understanding the mechanisms through which those data were generated.

In Probability we make assumptions about the world and calculate probabilities. These probabilities describe what we should expect to see if we were to observe the processes as they were assumed to exist. In Statistics, we collect information from statistical experiments, and use these data to infer what conditions were likely to have given rise to our observations. We are back solving for what set of assumptions are most plausible, given the observations. Uncertainty remains at the core of statistics. We will rarely be able to know *for certain* what different assumptions gave rise to the data that we observe, but instead look to clarify and quantify the ever-present uncertainty. Probability remains central to the study of statistics.

[1] For instance, we assume that particular probability mass functions hold, that certain distributions are present, that random quantities are independent

[2] Note, the word *data* is a plural noun in english. That is, we say "The observed data are …" rather than "The observed data is …" Some statisticians care deeply about correcting this misconception, and forget how weird those types of sentences to non-statisticians. I promise it will eventually sound more familiar!

Specifically, probability is the core tool for quantifying the ever-present uncertainty. Probability statements are the language of Statistics. As a result, the study of Statistics is largely the study of how we can take the set of tools that have been developed throughout the first part of this course, and apply them in the reverse direction. Statistics gives us the tools that we need to make sense of the world around us. Statistics serves as a process for evaluating the quality of evidence and drawing conclusions from it. Statistics is the necessary area of study to draw informed conclusions from the information that we collect. Ultimately, it is Statistics which powers quantitative decision-making. Virtually every avenue of the modern world demands that we make decisions on the basis of incomplete or imperfect information, and through Statistics we can ensure that these decisions are as informed as possible.

10.2 Background and Data

It is important to formalize some of the terminology upon which we will rely. A key challenge in formalizing these ideas is that, for many of these central concepts, we have an intuitive or colloquial sense of the idea. Just as with Probability, a large part of our goal during the early phases of learning Statistics revolves around connecting formalized ideas to intuitive concepts that we are familiar with from other contexts.

Definition 10.2 (Data)**.** Facts, figures, observations, or recordings in virtually any form (images, sounds, text, measurements) which are gathered and processed to form and communicate conclusions.

Data sit at the center of Statistics as the prime objects of study. We are concerned with how we can take data and draw valid conclusions. This may be by ensuring that the data are collected in a way which is suitable to draw conclusions, or by finding ways to graphically display the information within collected data, or by drawing inferences about the world using the data on hand. Data are the prime focus of Statistics. The data themselves are not particularly descriptive or actionable. Instead, the data are transformed into useful information through statistical techniques. We will broadly refer to any such process as a statistical analysis.

The goal of a statistical analysis can be placed into one of four categories. These categories define the four purposes of statistics.

1. **Descriptive statistics**: Descriptive statistics focuses on organizing and summarizing information. With descriptive statistics we seek to **describe** the current state of the world which lead to the data we have collected.
2. **Inferential statistics**: Inferential statistics provides methods for drawing conclusions and quantifying the uncertainty surrounding these conclusions, with regards to a population or process response for the data collected. With inferential statistics we seek to **infer** the underlying truth about a population or process of interest.

3. **Predictive statistics**: Predictive statistics provides methods for making predictions regarding the future behaviour of a process or population based on past observations from that population or process. With predictive statistics we seek to **predict** what is to come in the future.
4. **Prescriptive statistics**: Prescriptive statistics provides methods for suggesting interventions into a population or process according to its likely impact on a chosen criteria. With prescriptive statistics we seek to **prescribe** interventions based on what is likely to happen.

Example 10.1 (Charles and Sadie Categorized Questions)**.** Back out for coffee after a relaxing break, Charles and Sadie turn their attention to thinking about the possible use cases for Statistics. They begin to play a game, trying to identify which of the four major categories would be most appropriate to address their various questions of interest. For each of the following, identify whether the problem is best approached through descriptive, inferential, predictive, or prescriptive statistical techniques.

a. Charles wonders how many people, on average, visit the coffee shop each day.
b. Sadie wonders if the type of music playing in the shop impacts what purchases customers make.
c. Charles wants to determine how many chocolate chip cookies the coffee shop should prepare for Saturday morning.
d. Sadie wonders what the most common drink add-in is.
e. Charles wants to know how much the coffee shop should sell their coffees for, if they are trying to maximize income.
f. Sadie wants to understand how many people have signed up for the loyalty program.
g. Charles wonders if there is a meaningful difference between people's orders who sign up for the loyalty program and those who don't.
h. Sadie, in turn, questions how much the loyalty program is likely to grow over the next month.
i. Charles wants to understand how the reward tiers can be changed to grow the royalty program faster.

Solution

a. This is **descriptive**. This is an attempt to describe the state of the world as it exists.
b. This is **inferential**. This is an attempt to draw conclusions regarding the true state of the world.
c. This is **predictive**. This is an attempt to understand the future behaviour of uncertain quantities.
d. This is **descriptive**. This is an attempt to describe the state of the world as it exists.
e. This is **prescriptive**. This is an attempt to suggest an intervention to achieve a

desired outcome.

f. This is **descriptive**. This is an attempt to describe the state of the world as it exists.
g. This is **inferential**. This is an attempt to draw conclusions regarding the true state of the world.
h. This is **predictive**. This is an attempt to understand the future behaviour of uncertain quantities.
i. This is **prescriptive**. This is an attempt to suggest an intervention to achieve a desired outcome.

Each of the various roles that statistics can play is defined in terms of **populations**[3]. We understand this at an intuitive level, and this intuition is strong place to begin to formalize Statistics.

Definition 10.3 (Population)**.** The collection of all individuals or items that are under consideration in a study or experiment.

In many settings, the population is a well-defined, concrete idea. We may think of all individuals who attend a particular university, all birds of a species living in a particular park, all cards of a particular model made last year at a given factory. In each of these cases we can envision taking all members of the population[4] and placing them in one location. If we were able to do this, any questions we had about the population could be directly answered. This is not typically possible in these cases owing to practical considerations regarding the resources that would be required.[5] In many other settings, we cannot even imagine grouping the entire population of interest together, since the population is less concretely defined.

Consider, for instance, investigating the quality of vaccines that are produced at a particular facility. This facility will continue producing vaccines indefinitely into the future, and we may wish to know about the set of these future items. Similarly, we may wish to understand the impact of a particular teaching style on children's ability to learn math skills. In this case, we are not concerned with one particular school or one particular school board or one particular set of students. Rather we want to know how do children **in general** respond to this teaching intervention. In cases like these the population of interest is less concrete and more **conceptual**. It is not a specific well-defined group of individuals or items, and it may be possibly infinite. Instead of being able to collect all of the items of the population together we are only able to assess any individual or item and answer "is this a member of the described population?". We refer to these as **conceptual populations**.

[3] or *processes*

[4] Be that individuals or objects.

[5] It is not impossible to do so. For instance, governments often run national censuses, which are a full survey of every member of the population of a country. These are incredibly large undertakings, however, and are not feasible in many settings.

Definition 10.4 (Conceptual Population)**.** A set of individuals, items, or observations which are hypothetical in the sense that they do not tangibly exist as a concrete group, but instead share a common feature which defines the population. The units in the conceptual population are linked through the circumstances that they arise under resulting from conditions which are equivalent in some way. Sometimes conceptual populations are called **hypothetical populations**.

The utility of a conceptual population is that it allows us to unify the framework of Statistics whether we are studying groups of people or objects that really do exist in front of us, or those which we can describe but not collect. Even something as well-defined as the population of a country, for instance, is a population which may be conceptual in many regards. There are constantly new individuals being born in the country, those who are dying, those moving to or away from it. Still, none of us are confused about what we mean by the "population of a country". Likewise, conceptual populations in statistics are well-defined, even if they remain intangible.

Example 10.2 (Charles and Sadie Identify Populations)**.** Charles and Sadie had such fun identifying the uses for statistics during their last conversation, today at coffee they decide to identify populations of interest. They open up the local paper to the science section, and begin to read the headlines. For each headline, indicate the population of interest and specify whether or not this is a conceptual population.

a. "Study Finds Link Between Coffee Consumption and Productivity in Office Workers"
b. "Research Shows Decline in Pollinator Populations Across Agricultural Regions"
c. "Poll Indicates Attitudes Toward Healthcare Reform Among Registered Voters"
d. "Research Reveals Impact of Air Pollution on Respiratory Health Among Children in New York City in 2023"
e. "Survey Explores Relationship Between Social Support and Mental Health Among LGBTQ+ Youth"
f. "Poll Indicates Satisfaction with Public Transportation Among Commuters in Metropolitan Areas in Canada"

Solution

a. The population of interest here is simply "office workers". This is a conceptual population as we can easily tell whether or not someone belongs to the population (ask whether they work in an office), but we cannot easily describe the complete group of individuals.
b. The population of interest here is pollinators in agricultural regions. This is a conceptual population as it will be constantly shifting and changing. We are able to describe whether a particular animal is a pollinator living in an agricultural region, but we are not able to enumerate through the animals which would satisfy this.

c. The population of interest here is registered voters (wherever the study is run). This is not a conceptual population as a voter registry contains a list of all of the people within this population. It may be the case that this will change over time, but we can concretely determine the population at this point in time.
d. The population is children in New York City in 2023. This is not a conceptual population as, while there is not any practical way of gathering all children living in New York City in 2023 together in one place, this is a fully describable population that (given enough resources) could be gathered together.
e. The population of interest here is LGBTQ+ youth. This is a conceptual population, as we are able to assess membership to the population, but not fully enumerate the members.
f. The population here is Canadian metropolitan-area commuters. This is a conceptual population.

Ultimately, our goal with statistics is to understand a population. However, as a general rule, we are unable to directly observe the entirety of the population. While it is typically infeasible to observe the entire population, we are often able to observe some units from the population. These units, when collected together, are referred to as a **sample**.

Definition 10.5 (Sample)**.** A sample is a subset of a population which is observed, and as a result, information regarding these units is obtainable.

Thus, taken together we are interested in a particular population. We are typically unable to observe our population in full, and instead content ourselves with the capacity to view a subset of this population, which is referred to as a sample. Generally speaking, we are interested in some numeric quantities which describe the population. Perhaps we wish to know the average height of students in a school, or the total number of calls that are made at a company over a period of time, or the maximum litter size for a breed of house cats, or the proportion of defective units produced during a manufacturing run. In each of these situations, the question of interest relates to a quantity describing the population. If we were able to view the entirety of the population, we could simply compute the value of quantity. We refer to such quantities as **parameters**.

Definition 10.6 (Parameter)**.** A parameter is a numeric quantity of interest which is defined with regards to a population. A parameter captures the behaviour of the population. Typically the value of parameters will be unknown and unknowable.

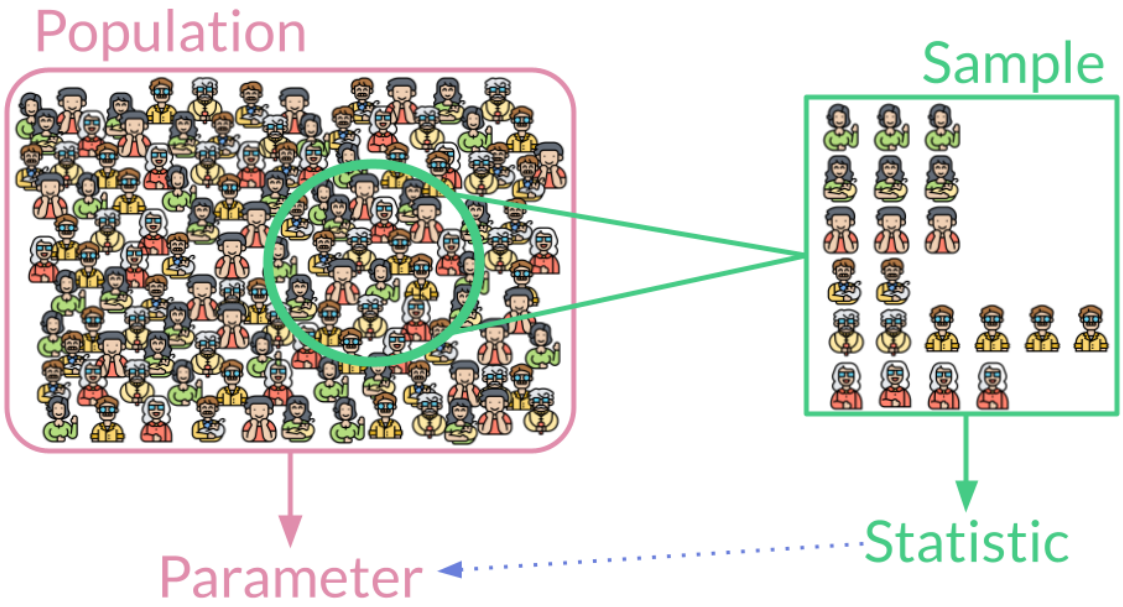
The fact that parameters are generally unknowable is the central tension at the heart of many statistical problems. To resolve this tension we turn our focus towards quantities which can be computed, namely those which are derived from samples that we have taken. These quantities are aptly named **statistics**.

Definition 10.7 (Statistics (Quantities))**.** A statistic is a numeric quantity of interest that is computed on a sample. Any quantity which is calculated based on observed data from a sample is a statistic.

In this regard, Statistics as a subject is the study of statistics (as quantities). For instance, if we take the average height of a group of students, or the number of calls made by a selection of employees during a period of time, or the maximum size that a particular cat breeder has for a litter, or the proportion of defective products from a random selection of items sampled from a manufacturing run, each of these are statistics. Note that to differentiate a statistic from a parameter we are functionally differentiating between whether the quantity was computed on a sample or a population. There is no difference in the quantity itself, it is what the quantity is computed with respect to.

Thus, with these definitions we are able to concretely outline the process of statistics. We have interest in a particular population, conceptual or otherwise. Specifically, we have questions which are answered by parameters defined for this population. These parameters are unknowable since it is infeasible to observe all members of the population, and so instead we turn to taking observations for subsets of the population. These subsets are called samples, and samples are observable. Once observed, we are able to compute quantities of interest on the samples, referred to as statistics. It is our hope that somehow these statistics will be representative of the underlying parameters of interest, thus allowing us to answer the questions about the populations using information from the sample.

Figure 10.1: On overview of the process of statistics. Visualized is the complete population, which is unable to be observed in full. There exist a parameter (or parameters) of interest that are computable from this population. We take a sample (subset) of this population, and can more accurately observe this sample, allowing us to compute the value of a statistic. This statistic is meant to correspond to the parameter of interest.



Example 10.3 (Experiments in the Coffee Shop)**.** Charles and Sadie, fully bought into the process of statistics, decide to put their new knowledge to the test. To do so they wish to determine how the process of statistics would apply to the world around them, in the coffee shop. For each of the following scenarios indicate what the population of interest would be, whether it is conceptual or concrete, identify the parameter of interest, a possible sample of size 4, and the relevant statistic.

a. To understand the daily traffic in the coffee shop, Charles counts the number of individuals who pass into the store in a particular hour.
b. To better understand the profitability of the store, Sadie collects the receipt totals for each of the customers arriving at the coffee shop.
c. To understand how the coffee shop has integrated into the community, both Charles and Sadie monitor the proportion of customers who are students at the local school, each day.

Solution

a. The population of interest is the customers of the coffee shop. This is a conceptual population. The parameter of interest is the average number of customers per day arriving at the coffee shop. Notably, Charles is measuring the average number per hour rather than per day, but this could be accommodated for during the analysis. Samples of size 4 would observe watching four separate hours, and counting the number of customers in the store each hour: one possibility would be $\{5, 35, 3, 22\}$. The statistic calculated is likely the average per hour (or a scaled version to convert it to the day). Based on the sample that is 16.25.
b. The population of interest is the purchases made at the coffee shop. This is a conceptual population. The parameter of interest is the average size of a purchase. Samples of size 4 would include any four observation of totals that could be observed at the store, perhaps $\{\$1.20, \$23.18, \$14.21, \$5.87\}$. The statistic of interest would be the average size of the transaction. Based on the sample that is \$11.115.
c. Here, the population of interest is the customers of the coffee shop. This is a conceptual population. The parameter of interest is the proportion of customers who are students at the local school. A sample of size 4 would be any four proportions observed from the day, perhaps $\{0.2, 0.1, 0.08, 0.25\}$. The statistic of interest is the average proportion, based on the sample that is 0.1825.

With this process outlined, we can revisit the roles that statistics will serve. Ultimately, our goal is to effectively use collected data[6] to discern and communicate information. We look to do this by:

1. **describing** the collected data, conveying the information that we have gathered;
2. **inferring** conclusions about our population parameters from our sample statistics;
3. **predicting** out-of-sample observations, based on the sampled ones; or
4. **prescribing** interventions for the population to influence a parameter.

The value in each of these applications stems from the capacity that we have to connect the sample to the population. As a result, much of our statistical focus centers on finding ways to ensure that our conclusions drawn from our sample are reflective of the overall population.

To understand how this is possible, consider an experiment which seeks to determine whether a particular coin is fair. Suppose that we toss it 100 times, and see 54 heads. Is this a fair coin? While we cannot be entirely sure, this seems to be more-or-less in line with the number of heads that we would expect to see if the coin were fair. There is uncertainty present, but we can be more certain that this is a fair coin compared to our beliefs prior to running the experiment. Now imagine that instead of seeing 54 heads on 100 tosses, we had seen 94. Immediately we should be skeptical that this coin is fair. It is perfectly possible that we see 94 heads on 100

[6] The sample.

tosses of a fair coin[7], but it is not likely. It does not seem to be what we would expect to observe, and as a result, we are right to be skeptical of this.

This intuitive connection between what we can say about the population and the sample relies upon the sample being representative of the population. We would not take flips from another coin to be evidence of whether our coin is biased. We would not consider the sample to be particularly representative if instead of writing down every result, we ignored every time that more than one tail came up in a row. The intuitions we have about the relationship between our sample and our population rely on the assumption that the sample represents the population in some meaningful sense. Choosing a representative sample is, as a result, an important aspect of the statistical process.

10.3 Sampling

Sampling is an area of study within statistics which focuses on the process through which units are recruited from the population into our sample. If the units we observe are biased in some way, or are inappropriately recruited, then it is immediately clear that conclusions drawn about the sample will not be transportable to the population. Imagine, for instance, that we are interested in the height of students at a university. If our sample contains only members of the basketball and volleyball teams, this will not be reflective of the heights of most students at the school. Our sample is unrepresentative. With sampling our goal is to select the sample in such a way so as to make guarantees about the information we learn from it, quantifying our uncertainty, and being relatively confident in our conclusions.

When we deal with populations which are conceptual in nature it may make less sense to think of sampling directly. For instance, it seems strange to think of our previous example of repeatedly rolling a die in the same manner that we think of recruiting students and measuring their heights. In the case where we wish to learn about a process rather than a concrete population, we will often frame the generation of our sample not through the language of *sampling* but rather through the language of **experimental design**. The design of experiments refers to the same set of factors that are considered for sampling: how can we ensure that the units we observe will be representative of the underlying population or process. However, with experimental design it is largely the case that we are in direct control of the types of factors that may lead our process to being unrepresentative. It is up to us to ensure that the units that are being generated, through the conceptual population, are representative of the units we are interested in based on the research question.

Our focus in this class will be on investigating several of the techniques used in sampling and experimental design to ensure that the samples we use are related to the population of interest in predictable ways. It will always be possible that, even following best practices and being

[7]The probability of this is 0.000000000000000000000940363535334879653069384024393902003768896946667155134491 62, which is very, very small, but not zero.

very careful with our implementation, we end up choosing a sample or having experimental units which are nonrepresentative. This is uncertainty which is unavoidable in any scientific inquiry. Our goal then is to understand this uncertainty, to quantify it, and to ensure that we are able to understand precisely how big of a risk is this lack of representation, and how will that change the results we can report. We begin by describing techniques which can be used to ensure that samples are good representations for the populations of interest. In the next section, we will turn to the same types of considerations for experimental design.

10.3.1 Simple Random Sampling

When considering the process of data collection via sampling, the primary decision to make is on the **sampling design**. The sampling design refers to the strategy that is employed to decide which members of the population will comprise the sample. Some sampling designs will not result in valid or representative samples. The most straightforward sampling design, which, if applied correctly, will produce valid samples is simple random sampling.

Definition 10.8 (Simple Random Sampling)**.** Simple random sampling is a sampling procedure in which each possible sample of a given size is equally likely to be obtained.

Simple random sampling produces a **simple random sample**. Simple random sampling is far and away the most important sampling scheme. It is an effective way of drawing a representative sample itself, it is intuitive, and it forms the basis of many other, more complex sampling schemes. Generally, we can think of simple random sampling as **sampling without replacement**.[8] Suppose we take our population to correspond to items in an urn, each labelled with the corresponding unit. Then a simple random sample is typically formed by selecting n items from the urn without replacement. Those which are selected form the sample. If desired, for any reason at all, it is possible to form a simple random sample **with replacement**, where in this setting the balls would be placed back into the urn after each selection. Whether the sample is to be formed with or without replacement, the same general procedure will be followed. Each member of the population will get assigned a numeric label (from 1 through to N, the population size) and then software is used to select a subset of n of the labels.

Example 10.4 (Rating Coffee Orders)**.** Charles and Sadie are still attending the coffee shop, and Charles is still working through the 960 different orders that are available (recall Example 3.5). Sadie, with a stronger grasp on statistics now, decides to try to understand the general quality of orders at the coffee shop. To do so, instead of ordering every possible meal (as Charles is doing), Sadie considers a simple random sample of possible orders.

a. Suppose that Sadie wants to understand the quality based on the next twenty visits to the coffee shop. Describe the procedure for forming a simple random sample.

[8] This, as we will remember from our previous chapters, makes the hypergeometric distribution deeply connected to simple random sampling.

b. What is the probability that Sadie's current order will be one of the orders included in the simple random sample?

Solution

a. Sadie desires a simple random sample of size $n = 20$ from the $N = 960$ different orders. To do so, suppose that we order each of the orders from $1, 2, \ldots, 960$ (perhaps ordered based on the order in which Charles will try them). Then Sadie will randomly selected, without replacement, 20 numbers from these values. These orders will then be taken over the next twenty visits, with Sadie recording whatever information is relevant for each of them. For instance, Sadie may end up select the orders corresponding to:

```
##  [1] 834 499 529 376 910 931 549 620 623 169 419 754 296  80 276 342 918 855 265
## [20]  87
```

b. Sadie's current order is 1 of the possible options of the 960. While we can work this out from first principles, we can also take this to be a hypergeometric random variable with $n = 20$, $N = 960$, and $M = 1$. Then we want $P(X = 1)$. For this, using the probability mass function of a hypergeometric random variable, we get

$$P(X = 1) = \frac{\binom{1}{1}\binom{959}{19}}{\binom{960}{20}} = \frac{1}{48}.$$

Note this is $\frac{20}{960} = \frac{n}{N}$. In general, this will be true for **any** simple random sample.

There is an appeal in the simplicity of simple random sampling. Moreover, it is quite clear how, as long as enough units are sampled, simple random sampling will result in a sample which is representative of the overall population. Despite these benefits, there are some drawbacks that are not easily overcome in the simple random sampling paradigm. For instance, if you imagine a situation in which your sample is spread over a large geographic region, it is unlikely to be practical to form a simple random sample. Additionally, if you do not have a list of all population members, a simple random sample cannot be formed as described.[9] Another practical concern involves sampling in this regard when units have a natural ordering. Suppose that you are looking to test the impact of a new cancer therapy, and wish to form a sample of current cancer patients who will receive the experimental treatment to see if it improves over the current practice. If you form through simple random sampling it is possible that you will have only patients who are newly diagnosed or else only patients who have had their diagnosis for a long time. Neither situation is a particularly effective method for testing the therapy,

[9]There are ways of doing this, but they are more complex and fall beyond the scope of this course.

and it becomes a large practice issue where you likely want to ensure that you have both sets of individuals represented in the sample.

To overcome these issues with simple random sampling, alternative sampling designs have been proposed. These alternatives can lead to more convenience in the sampling, and perhaps yield more accurate results than a simple random sample can. It is important to note that these alternative designs are only more effective when the design itself is taken into account when analyzing or describing the data.

10.3.2 Systematic Random Sampling

One alternative design to simple random sampling, which is closely related, is known as systematic random sampling.

Definition 10.9 (Systematic Random Sampling)**.** In systematic random sampling the sample is selected by choosing a random starting point from the list of members of the population, and then sampling every kth member until the desired sample size is reached.

Systematic sampling forms a sample that looks like a simple random sample, but it is more straightforward to implement. If you want a sample of size 50 from a population of size 500, then by selecting every 10th member of the population, you will achieve the sample you desire. You want to be able to pick any individual from the population, and so you should randomly select the starting point before picking every 10th member. Selecting every $10th$ individual is more straightforward administratively than generating random numbers and sampling those indices, particularly when there is a natural ordering of the individuals. However, there are some implementation decisions which need to be made, notably: what should k be and who should be the first individual included? It is common to have the process of systematic random sampling described as follows.

1. Divide the population size, N, by the desired sample size, n, and round the result down to the nearest whole number. This will be k.
2. Select a number, m, randomly between 1 and k. This will be the starting point.
3. Include in the sample m, $m + k$, $m + 2k$, and so forth until the last unit of the sample.

This will generally form a usable sample, if its shortcomings are properly accounted for. However, it is not without its shortcomings as a procedure.

To understand why this can lead to issues suppose that we have $N = 7$, and want to form a sample of size $n = 3$. Using this procedure we get $k = 2$, as the result of rounding down $\frac{7}{3} = 2.33\dot{3}$. Next, we select either 1 or 2 as our starting point. If we select 1 then we end up including $\{1, 3, 5\}$ and if we select 2 then we get $\{2, 4, 6\}$. Note that in we will never select item number 7, which means that there is no chance it is represented in our sample. This is a

problem. There are plenty of ways to resolve this concern, some of which lead to other issues themselves.

One small modification that can made is to select m between 1 and $N-(n-1)k$.[10] This will ensure that it is always possible to select up to the last unit. In our example with $N = 7$ and $n = 3$, the starting point is selected from $1, 2, 3$ giving in addition to the two possibilities outlined above, $\{3, 5, 7\}$ as a third option. This alleviates the issues of not including some members of the population in any possible sample. When this technique is used, however, it is worth noting that some elements become more likely to be included than the others.[11] This can be accounted for during analysis, but it needs to be completely understood to do so.

Example 10.5 (Randomly Sampling Customer Experience)**.** Sadie, content with the results of the simple random sampling of possible meals, decides to try to understand the overall customer satisfaction of individuals coming into the coffee shop. Charles suggests that a systematic sample may be in order, and they set out planning this.

Suppose that Charles and Sadie expect there to be 98 customers arriving in a day, and they wish to sample 10 of them.

a. Describe the process of forming a systematic sample from this population, including the specific values for the choices that are made.
b. What is a risk of this sampling design?

Solution

a. Here we have $N = 98$ and want $n = 10$, thus we take $k = \lfloor \frac{98}{10} \rfloor = 9$. We want to select the starting value between 1 and $98 - 9 \times (10 - 1) = 98 - 81 = 17$. Thus there are a total of 17 different samples we could wind up with. These are summarized as follows:

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1   10   19   28   37   46   55   64   73    82
##  [2,]    2   11   20   29   38   47   56   65   74    83
##  [3,]    3   12   21   30   39   48   57   66   75    84
##  [4,]    4   13   22   31   40   49   58   67   76    85
##  [5,]    5   14   23   32   41   50   59   68   77    86
##  [6,]    6   15   24   33   42   51   60   69   78    87
##  [7,]    7   16   25   34   43   52   61   70   79    88
##  [8,]    8   17   26   35   44   53   62   71   80    89
##  [9,]    9   18   27   36   45   54   63   72   81    90
## [10,]   10   19   28   37   46   55   64   73   82    91
```

[10] Instead of selecting m between 1 and k.
[11] For instance, 3 and 5 both show up twice, while every other element shows up only once.

```
## [11,]   11   20   29   38   47   56   65   74   83   92
## [12,]   12   21   30   39   48   57   66   75   84   93
## [13,]   13   22   31   40   49   58   67   76   85   94
## [14,]   14   23   32   41   50   59   68   77   86   95
## [15,]   15   24   33   42   51   60   69   78   87   96
## [16,]   16   25   34   43   52   61   70   79   88   97
## [17,]   17   26   35   44   53   62   71   80   89   98
```

b. This sampling design has several risks. The first is that there is no guarantee that $N = 98$ is the correct population size. As a result, this could end up not producing $n = 10$ or having $n = 10$ before the end of the day. A second possible issue is if there are temporal patterns to customers arriving, such that groups of customers are missed with high probability. Cyclical patterns do not lend themselves well to systematic sampling, and it is plausible that customer arrivals to a coffee shop may exhibit some cyclical tendencies.

10.3.3 Cluster Sampling

While systematic sampling can be a more straightforward method for implementing a sample that looks like a simple random sample, it is generally not going to alleviate concerns with (for instance) geographic separation. In this case the issue with either of the two aforementioned sampling schemes is that it they each would take substantial resources to send researchers to the area where the units to be sampled are. A remedy to this is to turn to **cluster sampling**.

Definition 10.10 (Cluster Sampling)**.** In cluster sampling individuals are grouped together into *clusters*. The clusters are then sampled at random, according to a simple random sampling scheme. Any selected cluster is then sampled in full.

For instance, you may define clusters based on the geographic region that is occupied. This way you can ensure, for instance, that you only visit a set number of geographic regions, while still sampling enough individuals to collect useful data. A key criteria for cluster sampling to be valid is that each cluster should represent the overall population well. This can be an issue where members of a cluster are often more similar to one another than to members of other clusters. As a result, you can end up with an unrepresentative sample owing to the clustered nature of the sampling. In order to form a cluster sample, the procedure is essentially equivalent to simple random sampling. First, the population is divided into groups (clusters) which are labelled from 1 through to the number of clusters that there are. Then, the clusters are randomly sampled, according to a simple random sample. Finally, all members of the selected clusters are included into the overall sample.

Example 10.6 (Charles and Sadie Sample their City)**.** Charles and Sadie are reflecting upon the chocolate bars that they sold for charity in the past. They want to understand the feelings that members of their city have towards charitable giving. They feel that sampling 300 homes is a useful number, but given that they will be going door-to-door, they want to do this in a clustered pattern. The city is made up of 947 blocks, each with 20 homes on it.

a. Describe how Charles and Sadie could use clustered sampling to form the sample that they desire.
b. What issues may arise using clustered sampling in this way?

Solution

a. Here we have $N = 18940$ and we want a sample of $n = 300$. In order to achieve $n = 300$, we require selecting 15 clusters, since each cluster has 20 homes in it. Thus, Charles and Sadie should label the blocks 1 through 947, and then select 15 of these at random. Once the 15 blocks are selected, all 20 homes on each should be visited. One possible sample would be

```
##  [1] 834 499 529 376 910 931 549 620 623 169 419 754 296  80 276
```

b. A major issue with this type of clustering is that homes that are on the same block are likely to be fairly homogenous. That is, city blocks will often be divided by socioeconomic status, education, and other demographic factors. As a result, selecting a full cluster is not likely to give a representative cross-section of the complete city, and instead, is likely to be skewed based on which clusters happen to be included. It seems very likely that with a survey investigating thoughts towards charitable giving, socioeconomic factors will play a role.

The major concern with cluster sampling is that, if the clusters are grouped together based on relevant information, the sample becomes predictably unrepresentative of the overall population. Because the clusters are often naturally formed based on a relevant factor (like geographic location), if this factor influences the topic that is being studied, the clustering design will influence the validity of the results. Still, cluster sampling, when done correctly, alleviates many of the difficulties with practically implementing a simple random sample.

Remark (Systematic Sampling as Cluster Sampling). Mathematically, systematic sampling can be seen as a particular form of cluster sampling. To see this note that, once k and m are defined, the set of individuals who are included in any given sample are completely defined and grouped together. As a result, you could preform these groupings of individuals, and treat those as clusters together. Then, instead of sampling individuals, you are sampling a cluster of individuals.

The key difference between cluster sampling and systematic sampling is that the clusters in systematic sampling are not typically naturally defined. There is not normally going to be a

clear separation forming the groups in this way, and as a result, it may not be easier to run a geographically isolated systematic sample than it is to run a geographically isolated simple random sample. However, the understanding of equivalence mathematically is useful when data from these samples are to be analyzed as the tools from cluster random sampling can be put to use to validly analyze systematic data.

10.3.4 Stratified Random Sampling

The key issue with cluster sampling is that natural clusters of individuals typically exhibit self-similarity. This is an issue when your sample is formed via complete clusters, however, it can be turned into a benefit to ensure a greater reliability of the sample itself. Exploiting this self-similarity leads to a sampling technique known as stratified sampling.

Definition 10.11 (Stratified Sampling)**.** In stratified sampling the population is divided into *sub-populations* known as strata. These strata should be comprised of groups of similar individuals. A simple random sample is formed within each strata, and each of the simple random samples are combined together to form the overall sample.

The major benefits of stratified sampling are two-fold. First, you will typically have more precision in your conclusions being drawn than from other sampling schemes. The reason being that individuals within a strata will be similar to one another, and so the variability that will arise based on which individuals are included is smaller than in other sampling schemes. Second, you are able to split your data into the different sub-populations describing results and conclusions for each group of individuals. This allows us to make conclusions both at the population level as a whole, but also at the sub-population level, which is often of direct interest.

To implement stratified sampling, you need to decide how many individuals will be sampled within each strata. A simple but effective way of doing this is to use **proportional allocation**. With proportional allocation the strata that are larger will have more members sampled than the strata which are smaller, which is a natural decision to make. To implement stratified random sampling with proportional allocation: 1. Divide the population into strata, based on a relevant and natural dividing criteria. 2. Within each strata, conduct a simple random sample of size n_j, where n_j is given by $n \times \frac{N_j}{N}$, where n is the desired sample size, N_j is the size of the jth strata, and N is the size of the population as a whole. This should be rounded to the nearest whole number. 3. Form the sample by including all members from each of the sampled strata.

Example 10.7 (Charles and Sadie Push for Transportation Infrastructure)**.** Charles and Sadie have faced some push-back on their attempts to increase access to well-funded public transportation options within their city. To understand better where the resistance is coming from, they decide to conduct a survey of households in the town. There are 18940 total homes

in the city, and Charles and Sadie figure that perhaps household income levels will influence opinions on investments in public infrastructure. Charles and Sadie categorize 940 of these households as high income, 10000 as middle income, and the remaining 8000 as low income. Suppose that they want a sample of size approximately 75 from this population.

a. Describe how a stratified sample can be formed in this setting.
b. What potential drawbacks are there from using stratified sampling here?
c. What other factors may have been useful to segment the population into natural strata?

Solution

a. This population is made up of three strata, with $N_1 = 940$, $N_2 = 10000$, and $N_3 = 8000$. Using proportional allocation this will result in a sample from the first of size $n_1 = \frac{940}{18940} \times 75 = 3.7 \approx 4$. For the second we get $n_2 = \frac{10000}{18940} \times 75 = 39.6 \approx 40$. For the third we get $n_3 = \frac{8000}{18940} \times 75 = 31.6 \approx 32$. Taken together this gives 76 as the sample size, which is close to desired. Then, three separate simple random samples are to be formed. The first is a sample of 4 homes randomly selected from the numbers, $\{1, \dots, 940\}$, the second is 40 homes randomly selected from $\{1, \dots, 10000\}$ middle income homes, and the remaining 32 are selected from $\{1, \dots, 8000\}$ from the low income households. These three groups are combined together, and interviewed to form the overall sample.
b. The major drawback is that this is likely to result in a costly sampling scheme. Specifically, as contrasted with the cluster sampling in Example 10.6, there is no guarantee that Charles and Sadie would not be required to travel across many of the blocks in their city. It is possible that to visit the 76 households, they have to visit 76 different blocks, which greatly increases the cost of administering the sample as described.
c. It may have also been useful to stratify the population, for instance, based on vehicle ownership, based on commuter status, or based on political leanings. Each of these is likely to be relevant to subgroup analyses, while also reducing the variability within the strata themselves.

While stratified samples are often the most efficient samples, statistically, they do not alleviate many of the practical concerns regarding simple random samples. In fact, in some cases, these issues may even be exacerbated. If, for instance, strata are formed on the basis of geographic location, then forming a stratified sample will guarantee that it is required to visit each geographic location. The sampling scheme which is selected should be concordant with the goals of the analysis as well as the restrictions and constraints that are at play.

10.3.5 Multistage Sampling

Sometimes the nature of the population, question of interest, or constraints at play render any single sampling design ineffective to construct a useful sample. In these cases multistage sampling can be an effective way of tailoring the sampling design to the specific requirements.

Definition 10.12 (Multistage Sampling)**.** In multistage sampling one or more of the discussed sampling techniques are combined into a multistage procedure to effectively and efficiently target the population of interest. Multistage sampling may combine simple random sampling, systematic random sampling, cluster sampling, or stratified sampling in different sequences and orders to achieve a custom-specified sampling scheme.

For example, household surveys are commonly run. To do so, a researcher may randomly sample cities in the area of interest. Within those cities, they may stratify based on region in the city, and within each region cluster based on the blocks. Then, a simple random sample within the blocks is taken, and those households are selected. This sampling scheme can be understood in relation to each of the component sampling schemes, and creates a flexible way of constructing a sampling scheme which meets the needs of the situation.

10.4 Experimental Design

While sampling is often a useful framing for collecting data, there are times where our goal is not to simply understand the way that a population is, but rather to understand the impact of particular intervention. Consider, for instance, studies which look at the efficacy of medical treatments, the utility of new fertilizers on agricultural yield, or the impact of political interventions on climate change. In each of these cases we are not interested in the current state of a population, but rather how some action influences the state of a population. For this, we turn to the process of experiment design. In experimental design we seek to understand how **experimental units** have their **response variables** impacted, based on the **levels** of particular **factors**, or **treatments**. In plain language our goal is to understand how specific interventions influence some trait in a population. To proceed we formally define each of these concepts.

Definition 10.13 (Experimental units)**.** Individuals or items upon which the experiment is being performed. If the experimental units are humans, we will often call them subjects or patients, in place of units.

Definition 10.14 (Response Variable)**.** The characteristic or trait of the experimental unit that is measured or observed. The experiments purpose is to understand how a response variable reacts to a particular intervention.

Definition 10.15 (Factor)**.** A variable whose effect on the response variable is of interest. The factor is the variable which is being controlled or manipulated within the experiment, and is the *cause* of the change in response variables.

Definition 10.16 (Levels)**.** The possible values of a factor. We are often comparing two or more levels of the factor, determining how specific levels impact the response variable.

Definition 10.17 (Treatment)**.** A treatment refers to the complete experimental condition. That is, the set of all levels across all factors that are assigned to an experimental unit. In one-factor experiments, this is the levels of the single factor. In multifactor experiments, a treatment is a combination of the levels of the factors.

In an experiment, experimental units are observed after having been given the treatment. The response variable is measured, and compared across the different levels of the factors (or across the different treatments) to determine how the various treatments impact the outcome. In a well-structured experiment it is possible to conclude that the treatment *causes* a particular impact on the response variable, so long as the analysis takes into account the limitations of the experiment that was run.

Example 10.8 (Sadie's House Plant Growth)**.** Sadie has become very interested in understand the conditions under which house plants will thrive. Through some informal experimentation Sadie believes that the frequency of watering, type of fertilizer, and amount of sunlight are all important factors in the plant growth. Sadie is predominantly interested in determining the impact on the height of the plants.

a. Indicate the experimental units, response variable, factor(s), possible level(s), and treatment(s) in this proposed experiment.
b. Indicate further examples of response variables and factors that may be relevant to Sadie's question of interest.

Solution

a. In this experiment, the units are the houseplants that Sadie is watching. The response variable will be the height of the plants (measured in, for instance, centimeters). The factors under consideration are the frequency of watering, type of fertilizer, and the amount of sunlight. Some possible levels for frequency of watering may be weekly, versus biweekly, versus monthly. Some possible levels for type of fertilizers may be natural compost, synthetic compost, or no fertilizer. Some possible levels for the amount of sunlight may be direct, indirect, or no sunlight. The treatments will be comprised of the combinations of the different levels for the three outlined factors (for instance, weekly watering, with natural compost, and direct sunlight; weekly watering, with natural compost, and indirect sunlight; etc.). If these levels are considered then there will be a total of $3 \times 3 \times 3 = 27$ total

treatments under consideration.

b. Sadie may instead wish to measure something to do leaf size, or stem width, or number of flowering plants. Each of these will produce different results, and will be useful for slightly different sets of considerations that Sadie may have. For other factors, Sadie may consider the pot size, or perhaps exposure to other stimulus.[12]

10.4.1 The Principles of Experimental Design

Just as there were many different schemes for constructing a sample, there are also many different experimental designs. We will explore two common designs, but it is useful to first understand the guiding philosophy of experimentation in Statistics. There are three key factors that ensure that data collected from an experiment are useful for drawing scientific conclusions.

First, is the idea of statistical control. It is not enough to know whether or not a particular treatment was followed by a positive response in the response variable. Instead, we require there to be some point of comparison. We call this point of comparison a statistical **control**. The idea is that we should always be comparing two or more treatments, even when our interest is in one particular treatment. If the treatment of interest is truly beneficial, we should be able to see this by comparison to other treatments. This way, we are able to ensure that the changes in the response variable that we see are related to our intervention, rather than by random chance. Sometimes we wish to see whether a particular treatment is effective, and there does not exist another treatment option that is a plausible candidate. In these cases, we often will take a treatment option to be "nothing at all", which is to say no direct intervention on the given factor.[13] In this way our control can still be used to see if our active intervention improves over doing nothing, when that is the alternative.

Beyond statistical controls, experiments rely on **randomization** and **replication**. With randomization the idea is that the treatment that each experimental unit gets should be be randomly selected, without consideration of the specific unit. This way unintentional selection bias can be avoided within the groups. For instance, in a medical study, if you end up giving the experimental treatment to patients who are otherwise healthier, then you may expect that the experimental treatment will produce better results not because it was more effective, but because the patients who received it are the ones who are expected to have better outcomes irrespective of treatments. In addition to randomization, replication serves a central role in experimentation. The experiment should be conducted on a sufficient number of experimental units to ensure that random noise does not cloud conclusions. In particular, the sufficient sample size will ensure that the groups created via randomization will truly resemble each other, and the more units in the study, the better able you are to discern differences which exist

[12] For instance, some people claim that music will help with plant growth.

[13] In medical studies this takes the form of a placebo: a treatment which looks like a medical treatment but has no active ingredients.

between treatments. These principles can be put to work across various different experimental designs, depending on the specific experimental setup and constraints that are at play.

10.4.2 Completely Randomized Design

The key question when defining an experimental design is how do the experimental units get assigned to the various treatment options. The most obvious choice is to randomly assign each unit to one of the treatment options, ignoring any underlying factors about the units. This is considered a **completely randomized design**.

Definition 10.18 (Completely Randomized Design)**.** A completely randomized design is one in which the experimental units are randomly divided into groups, one for each treatment in the experiment. The treatments are then assigned to each of the groups, randomly selected for each.

Typically we will consider an equal assignment of numbers of experimental units to each treatment option, though this is not strictly required. In the completely randomized design, we ignore anything that we may know about the experimental units beforehand.

Example 10.9 (Sadies Houseplants: Completely Randomized Design)**.** Sadie is going forth with experimentation to understand how different factors impact the height of house plants. Sadie decides to test treatments comprised of combinations of three factors: watering frequency (low frequency high volume versus high frequency low volume), fertilizer use (store bought fertilizer, versus homemade compost, versus no fertilizer), and sun exposure (direct sun exposure, versus indirect sun exposure, versus artificial light exposure). There are a total of 72 house plants, and Sadie wishes to use a completely randomized design.

Describe how this experiment can proceed as outlined by Sadie.

Solution

There are a total of $2 \times 3 \times 3 = 18$ different treatment options. This means that for each treatment option, $\frac{72}{18} = 4$ houseplants should be assigned to it. We can form these 18 groups by sampling **without replacement** from the numbers 1 through 72, 4 at a time. Each number is assigned to one of the houseplants, and that forms the groups. For instance, the 18 groups may be:

```
##       [,1] [,2] [,3] [,4]
##  [1,]   66   48   47   38
##  [2,]   17   34    1   15
##  [3,]   14    8   68   71
##  [4,]   35   26   62   42
##  [5,]   37   11   61    2
```

```
##  [6,]   72    4   19   67
##  [7,]   41   64   18   25
##  [8,]   69   39   10   70
##  [9,]   50   53   30   33
## [10,]   40   29   31   24
## [11,]   16   49   32   36
## [12,]   20   44   55   51
## [13,]   22   65   59   54
## [14,]   60    6   13   45
## [15,]   23    3   52   46
## [16,]    9    5   43   57
## [17,]   58   28   63    7
## [18,]   56   27   21   12
```

Then, each of the groups gets assigned one of the treatments (low frequency/store bought/direct exposure; low frequency/store bought/indirect; low frequency/store bought/artificial; etc.). The treatments are given to the units, and then the heights are measured and recorded alongside the treatment that was assigned.

There are at least two shortcomings of completely randomized designs which we may wish to overcome. The first is that, in a completely randomized design, we may not be able to understand the impact on subpopulations of interest. Because randomization occurs without consideration of any other factor it is also not possible to directly analyze how treatment may impact the outcome variable segmented by these factors. While this is often not the primary question of interest, it will often be the case that having answers to these types of questions is desirable. Second, a completely randomized design may be less efficient at capturing the true effect of treatment, when treatment is mediated by other factors. If some groups of experimental units respond more favourably[14] than others, ensuring that treatment allocation is split within these groups will lead to more precise estimates of the true treatment effect. As a result, we will often turn to more involved experimental designs to allocate treatment options.

10.4.3 Randomized Block Design

When we wished to exploit the structure of a population in sampling, making use of systematic differences, we divided the population into groups called strata on the basis of these traits. We can do the same thing with our experimental units forming **blocks** of experimental units. This blocking procedure gives rise to the *randomized block design*, an alternative to a completely randomized design.

[14] Or less favourably…

Definition 10.19 (Randomized Block Design)**.** A randomized block design assigns treatments randomly to all units within a block of experimental units. That is, the experimental units are separated into various blocks, and then within each block a completely randomized procedure is used.

With the use of the block design, you are able to assess not only is there an overall impact of treatment on the response variable, but also is this impact of treatment impacted[15] through the blocking factor(s). This may be of scientific interest directly, and it also may help to ensure that random noise does not erode the ability to discern the true impact of treatment on the outcome. Typically, blocking factors will be natural factors which are suspected, or known, to influence the outcome, but which are of secondary interest to the experimenter.

Example 10.10 (Sadie's House Plants: Randomized Block Design)**.** Of Sadie's 72 plants, 18 of them are species of trees, 18 of them are species of vines or other crawlers, and the remaining 36 are flowering plants. If Sadie decides to test treatments comprised of combinations of three factors, watering frequency (low frequency high volume versus high frequency low volume), fertilizer use (store bought fertilizer, versus homemade compost, versus no fertilizer), and sun exposure (direct sun exposure, versus indirect sun exposure, versus artificial light exposure), how can a randomized block design to perform this experiment?

Solution

Here, the type of plant is the natural blocking factor. There are 18 different treatment options. This means that for each type of plant we perform a completely randomized design with 18 treatments. To do so, we can assign each of the trees to a single treatment option (there are 18 trees and 18 treatments, so this is a matter of simply randomizing the order of treatments and assigning one to each), we can assign each of the vines to a single treatment option (same as with the trees). For the flowering plants, there are 36 of them and 18 treatments, so each treatment should have two different plants. If we label the flowering plants 1 through 36, then we can simply draw **without replacement** from the numbers 1 through 36, 2 at a time. For instance, the 18 groups may be:

```
##       [,1] [,2]
##  [1,]    2   11
##  [2,]   17    4
##  [3,]   14   10
##  [4,]   36   21
##  [5,]   15    1
##  [6,]    9   29
##  [7,]    3   12
##  [8,]   18    7
##  [9,]    8   26
```

[15] Or "mediated".

```
## [10,]   16   35
## [11,]   20   24
## [12,]   22    5
## [13,]   25   30
## [14,]   23   19
## [15,]   31   32
## [16,]   27    6
## [17,]   33   34
## [18,]   28   13
```

Then, each of the groups, within each of the blocks gets assigned one of the treatments (low frequency/store bought/direct exposure; low frequency/store bought/indirect; low frequency/store bought/artificial; etc.). The treatments are given to the units, and then the heights are measured and recorded alongside the block and the treatment that was assigned.

10.5 Data Description and Organization

Whether data are collected via sampling or experimentation, it is important to ensure that statistical principles are followed so that the observed data are representative of the population of interest and useful for accomplishing the goals of the statistical analysis. There is a substantial amount of statistical work which goes into ensuring that data collection is valid. Once valid data have been collected, we must *do* something with them. As previously discussed, there are typically four use cases for data. In this course, our focus is on description and inference. Before we can use the data to describe patterns or conduct inference, we must first develop a shared language around what data *are*. Some of this has been informally introduced throughout our discussions thus far, however, the formal definition remains important for ensuring the foundation for statistical analyses.

Definition 10.20 (Variable)**.** A characteristic or trait that can vary from one observation to the next is called a variable. Variables are the relevant pieces of information that are recorded in our data. We may have one or more variable recorded for each individual unit in our data.

Definition 10.21 (Observation)**.** An observation is an individual piece of data. Our data are comprised of multiple observations across the various units in our sample (or on our experimental units).

Generally speaking, we make observations of variables, and together this forms our data. We use the data to answer the questions of interest or conduct our analyses. Every variable can be categorized as either a **qualitative** or **quantitative** variable. Qualitative variables are

the non-numeric variables we observe, following categories or other less structured formats. Quantitative are numeric variables.

Definition 10.22 (Qualitative Variable)**.** Any variable which is not numerical, such as those which fit into categories, are referred to as qualitative variables.

Definition 10.23 (Quantitative Variable)**.** A quantitative variable is any variable which is described numerically.

Quantitative variables can be either discrete or continuous. We saw this distinction when working with random variables, and the distinction is equivalent in the case of variables in a collected dataset as well. A variable is considered discrete if it can take on a (countable) number of listable values. A variable is considered continuous if it can take on any value from a defined range of values. Often times, just as with random variables, we make the distinction based on how we wish to think about the variables, rather than based on the theoretical underlying truth.[16]

Definition 10.24 (Discrete Variable)**.** A quantitative variable which can take on values from a listable set. There are either finitely many options for the variable, or else a countably infinite number.

Definition 10.25 (Continuous Variable)**.** A quantitative variable that can take on an uncountably infinite number of values is called continuous. Continuous variables can theoretically take on any value over a range of values, with the possibilities unable to be enumerated.

Example 10.11 (Charles and Sadie Categorize Variables)**.** Charles and Sadie realize that oftentimes there are many different ways of measuring qualities or traits that are of interest in a study. Upon realizing this, they begin to discuss a number of topics, considering how they may be measured.

For each of the following traits, discuss different options for variables that could represent the quantity discussed. For each, include possibilities which are qualitative and those which are quantitative, and specify whether the quantitative are discrete or continuous.

a. Charles suggests that there are many ways of thinking about attained education.
b. Sadie, still thinking of plants, realizes that there may be many ways of thinking about the size of different plants.
c. Based on an overheard conversation, Charles wonders how socioeconomic status may be measured.
d. Sadie, after enjoying a snack at the coffee shop, thinks about how we might measure the quality of food.

[16] As a result, times, heights, or volumes will often be considered continuous even if theoretically they are discretized.

Solution

a. Attained education may be a qualitative variable if it is described via categories (for instance, high school, college diploma, undergraduate degree, etc.) It could almost be made to be quantitative if, for instance, you measured the total number of years that someone had formal education for. This would be a discrete quantitative variable.
b. The size of different plants could be qualitative by using subjective sizes (for instance, small, medium, and large). It is possible to measure these quantitatively as well, for instance by considering the height or volume of the plant. Likely height and volume both are best considered continuous values, but they may be discretized in certain datasets.
c. Socioeconomic status can be a qualitative variable if, for instance, it is categorized on the basis of high/medium/low. To form a quantitative variable here you may consider household income, or wealth. Depending on how income or wealth are measured it may be discrete (for instance, counting the number of thousands of dollars) or continuous (if you listed exact dollar figures). This is an example of a variable which is *always* discrete, technically, but may be better treated as continuous.
d. Food quality could be categorically rated (bad/average/good), or on a similar graded scale (S/A/B/C/D). Alternatively, the food could be subjectively graded numerically, giving for instance a star rating (out of 5) or a rating out of 100. Likely these ratings would be discrete, but would be possible to come up with a continuous rating scale if that were desired.

10.6 From Data to Insight

Whether data are collected via sampling or via experiments, the data themselves are not particularly useful for insight. If you are presented with a large dataset, it will likely not be possible to directly interpret the data, or communicate a message. Instead, we need to take the data as input and convert them to more useful products. The remainder of this course will focus on ways of doing this within statistics.

We will focus both on how to summarize and communicate data that have been collected, and then how to begin gaining insight from these data. These are the first two roles of statistics, as introduced before: description and inference. All of the roles that statistics plays build from the idea that we have been able to collect data which are somehow relevant and representative of the underlying population of interest. We investigate the data, describe what has been observed in the sample, or attempt to conduct inference not because we are interested in the data themselves, but because we hope that the data will be reflective of the population of interest. We are not primarily interested in the statistics that we calculate, but on what these statistics say about the parameters of interest.

It is important, as we begin to explore how we can use data directly, to keep in mind that the entire statistical enterprise relies on having high quality data available. This relies on having measured the factors that we care about, in ways that are meaningful. It relies on representative samples and well-designed experiments. Without adhering to the principles discussed throughout this chapter, statistics cannot proceed in a way which addresses our goals. High quality data is not a substitute for statistical analysis, but it is a prerequisite for it.

Exercises

Exercise 10.1. For each of the following scenarios, identify the population of interest and the sample. Is the described sampling effective for studying the population of interest?

a. A television network wants to gauge the viewership of a new reality TV show. They collect ratings data from 1,000 households in a specific region.
b. Researchers are studying the effects of a new exercise program on heart health. They recruit 50 adults from a local fitness center.
c. A civil engineering firm is studying the strength of concrete in a particular city by taking core samples from randomly selected buildings.
d. A hospital is conducting a patient satisfaction survey to improve its services. They send questionnaires to 300 patients who recently stayed at the hospital.
e. A polling agency conducts a survey to predict the outcome of an upcoming election. They interview 2,000 registered voters across the country.
f. Researchers are investigating the prevalence of a specific genetic mutation associated with a rare disease in a certain region. They collect DNA samples from 500 individuals living in that area.
g. An aerospace company is conducting a study to determine the reliability of a new aircraft engine. They collect performance data from 20 engines produced in their factory.
h. A public health agency is conducting a study to assess the vaccination coverage for a new vaccine in a specific age group. They survey 1,000 children aged 5 to 12 in a given city.
i. A software development company is conducting a usability test on a new mobile app. They invite 50 users from their customer database to participate and provide feedback.
j. A car manufacturer wants to test the fuel efficiency of its new electric vehicle model by collecting data from 100 vehicles sold in the United States.

Exercise 10.2. Define the population for each of the following scenarios.

a. A shipment of bolts is received from a vendor. To check whether the shipment is acceptable with regard to shear strength, an engineer reaches into the container and selects 10 bolts, one by one, to test.
b. The resistance of a certain resistor is measured five times with the same ohmmeter.

c. A graduate student majoring in environmental science is part of a study team that is assessing the risk posed to human health of a certain contaminant present in tap water in their town. Part of the assessment process involves estimating the amount of time that people who live in that town are in contact with tap water. The student recruits residents of the town to keep diaries for a month, detailing day-by-day the amount of time they were in contact with tap water.
d. Eight welds are made with the same process, and the strength of each is measured.
e. A quality engineer needs to estimate the percentage of parts manufactured on a certain day that are defective. At 2:30 in the afternoon he samples the last 100 parts to be manufactured.

Exercise 10.3. For each of the following descriptions, indicate whether the described quantity is a parameter or a statistic. Why?

a. The average height of all adults in the world.
b. The average income of a sample of 100 households in a city.
c. The proportion of students who passed a standardized test in a specific school.
d. The median age of all people living in a country.
e. The unemployment rate for the entire country.
f. The mean score of your friends on a recent quiz.
g. The percentage of cars in a state that are electric vehicles.
h. The percentage of cars in a parking lot that are electric vehicles.
i. The variation of a species' weight in a national park.
j. The percentage of customers who prefer brand A over brand B in a nationwide survey.
k. The average number of students per classroom in a school district.
l. The proportion of people who voted for a specific candidate in a county.
m. The range of salaries in a particular company.
n. The range of salaries in a particular department at a company.
o. The percentage of defective products in an entire manufacturing plant.
p. The percentage of defective products in a batch from a manufacturing plant.
q. The most common age of all employees in a corporation.
r. The percentage of students who participate in extracurricular activities in a specific school district.
s. The spread in test scores for all students in a university.
t. Is the percentage of households with pets in a country.

Exercise 10.4. For each of the following variables, identify whether it is a categorical or numeric variable. For the numeric variables, identify whether it is discrete or continuous. Explain.

a. The colour of cars in a parking lot.
b. The number of books a person has read.
c. The contents of a shopping cart at the grocery story.

d. Temperature measured in degrees Celsius.
e. Self-identified gender.
f. Height in centimeters.
g. The genre of a film.
h. Household income.
i. Socioeconomic status.
j. Age for study participants.

Exercise 10.5. For each of the following questions, identify whether this would be addressed through descriptive, inferential, predictive, or prescriptive statistics. Why?

a. What is the average income of employees in our company?
b. What are next years sales trend for our product going to be?
c. How many students passed the math exam last year?
d. Is there a difference in test scores between students who received extra tutoring and those who did not?
e. What percentage of customers are likely to purchase our new product based on market research?
f. Can we recommend changes to the manufacturing process to reduce defects in our products?
g. How confident are we that the recent increase in website traffic will lead to higher sales?
h. What are the key factors influencing employee job satisfaction in our organization?
i. Should we implement a new marketing strategy based on customer feedback and buying patterns?
j. What is the margin of error for the survey results on public opinion regarding a proposed policy change?

Exercise 10.6. For each of the following scenarios, indicate whether the described sampling method is simple random sampling, systematic random sampling, cluster random sampling, stratified random sampling, or multistage sampling. For each, indicate whether it is an appropriate choice, and what the drawbacks of using it may be.

a. To understand commuting habits in a city, researchers divide the population into income brackets and then randomly select individuals from each bracket.
b. A university wants to study the satisfaction level of its students. They randomly select several classrooms and survey all the students present in those classrooms.
c. A market researcher selects every 10th person who enters a shopping mall to participate in a survey about consumer preferences.
d. A pharmaceutical company wants to test a new drug. They randomly select five hospitals and then sample all the patients being treated for the targeted condition in those hospitals.

e. A city council wants to assess the opinion of its citizens on a proposed policy change. They divide the city into neighborhoods and randomly select a few neighborhoods, from which they randomly select households to survey.
f. A company wants to know the opinion of its employees about a new HR policy. They randomly select employees from the company list to survey about the policy.
g. A magazine wants to survey its readers about their preferences for future content. They randomly select subscribers from each geographical region where the magazine is distributed.
h. A political pollster wants to gauge voter preferences in a country. They divide the country into regions and randomly select a few regions to conduct interviews with all a random selection of eligible voters.
i. A company wants to study the spending habits of different age groups. They divide the population into age brackets and then randomly select individuals from each bracket.
j. A bank wants to assess customer satisfaction with its services. They decide to survey every 15th customer that enters a particular branch location on a particular day.

Exercise 10.7. For each of the following scenarios, indicate what sampling scheme may be appropriate, and discuss the rationale (weighing the benefits and costs) for implementing it.

a. A city council wants to assess the opinions of citizens on a new recycling program.
b. A university conducts research on student satisfaction with campus facilities.
c. A political campaign team aims to understand voter preferences in a state election
d. A national magazine conducts a survey to determine readers' preferences for content.
e. A health organization plans to conduct a study on the prevalence of a rare disease in the country.
f. A wildlife conservation organization wants to study the population dynamics of a rare species in a national park.
g. A sports league conducts a survey to gather feedback from fans about their experiences attending games.
h. A government agency wants to assess the impact of a policy change on small businesses across the country.
i. A tourism board wants to gather feedback from tourists visiting a popular destination
j. A financial institution wants to analyze customer satisfaction with online banking services
k. A climate research institute plans to study the effects of climate change on ecosystems in a specific region.

Exercise 10.8. A software company wants to gather feedback from its users about a new feature in their application. They have a database of 5,000 active users and plan to randomly select 200 of them for the feedback survey.

a. Describe how a simple random sample can be implemented in this situation.
b. What concerns may the company have with using the simple random sample in this setting?

c. What are the advantages to using simple random sampling for this particular use case?
d. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.9. Conservationists want to assess the population of a specific wildlife species in a forest reserve. They are unable to conduct a census of the reserve, and instead rely on having divided the reserve into several regions, from which they plan to do a simple random sample and then count the prevalence of the species in each of the sampled regions. Suppose that they divide the reserve into 500 regions, and have the capacity to sample 20 of them.

a. Describe how a simple random sample can be implemented in this situation.
b. What concerns may the conservationists have with using the simple random sample in this setting?
c. What are the advantages to using simple random sampling for this particular use case?
d. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.10. A manufacturing company produces thousands of identical products daily and wants to ensure quality control by inspecting a sample of the products. The company wants to select a sample of products for inspection without inspecting every single item, as it would be time-consuming and costly. They decide that of the $20,000$ products produced per week, they want to sample and test 100 of them.

a. Describe how the company can use systematic random sampling to approach this problem.
b. What concerns may the company have with using systematic random sampling in this setting?
c. What are the advantages to systematic random sampling for this particular use case?
d. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.11. A research team wants to study the air quality in a city over a period of one month. They plan to collect hour-long air samples hourly at various locations throughout the city over the course of a month. They have the resources to collect 50 samples at each location.

a. Describe how systematic random sampling can be implemented in this situation.
b. What concerns may the there be with using systematic random sampling in this setting?
c. Suppose instead the researchers only had resources to collect 20 samples per location. How might this cause further issues?
d. What are the advantages to systematic random sampling for this particular use case?
e. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.12. A marketing agency wants to conduct a survey to assess consumer preferences for a new product in a large city. The city can be divided into roughly 50 neighbourhoods which are dispersed throughout the city. On average, each neighbourhood has approximately 100 houses. In total, the agency has the resources to visit approximately 500 houses.

a. Describe how cluster random sampling can be implemented in this situation.
b. What concerns may the agency have with using cluster random sampling in this setting?
c. Is the agency guaranteed to have a sample size of 500? Why or why not?
d. What are the advantages to using cluster random sampling for this particular use case?
e. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.13. A school district wants to assess the effectiveness of a new teaching method across multiple schools. There are 50 schools in the district in total, and the new teaching method will be best served by being implemented throughout the entirety of the school at once. As a result, all teachers and students at the selected schools will be interviewed to assess its efficacy. Resources exist to trial this at 5 schools.

a. Describe how cluster random sampling can be implemented in this situation.
b. What will the sample size be in this situation? Explain.
c. What concerns may the agency have with using cluster random sampling in this setting?
d. What are the advantages to using cluster random sampling for this particular use case?
e. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.14. A health department wants to assess the prevalence of a disease in a city with a diverse population. They plan to divide the population into age groups (children, adults, seniors) and randomly select individuals from each group for testing. In this city, there are 8000 children, 15000 adults, and 6000 seniors.

a. Describe how stratified random sampling, with proportional allocation, can be implemented in this situation.
b. What concerns may the health department have with using stratified random sampling in this setting?
c. What are the advantages to using stratified random sampling for this particular use case?
d. Would any of the other sampling schemes be useful for this particular problem? Explain.

Exercise 10.15. A research team wants to study the academic performance of students in a school district with schools of varying socioeconomic status. They plan to divide students into socioeconomic strata (low, medium, high) and randomly select students from each stratum for analysis. There are approximately 3000, 4000, and 1000 students living in low, medium, and high socioeconomic areas in the city.

a. Describe how stratified random sampling can be implemented in this situation.
b. What concerns may the research team have with using stratified random sampling in this setting?
c. What are the advantages to using stratified random sampling for this particular use case?
d. Would any of the other sampling schemes be useful for this particular problem? Explain.
e. How might the researchers decide to stratify this population differently? How might this choice inform the conclusions that are drawn from the analysis?

Exercise 10.16. For each of the following techniques for generating a sample, describe the issues which may arise from the sample, and suggest alternative schemes which may alleviate those concerns.

a. A marketing team wants to gather feedback on customer preferences at a local convenience store. They approach customers who happen to be in the store at the time of the survey and ask them to participate.
b. A student organization conducts a survey on campus cafeteria food quality by approaching students dining in the cafeteria during lunch hours and asking them to participate.
c. A company conducts a poll on social media platforms to gather opinions on a new product feature. They post the poll on their social media pages and invite followers to participate.
d. A nonprofit organization conducts a survey on housing insecurity by approaching individuals accessing community support programs, and asking them to participate.
e. A company sends out an online survey to its email subscribers to gather feedback on a recent product launch. They invite subscribers to participate by clicking on a link in the email.

Exercise 10.17. For each of the following experiments, identify the experimental units, response variable, factor(s), possible level(s), and treatment(s).

a. A study is conducted to investigate the effect of different fertilizer types on the growth of tomato plants.
b. Researchers want to test the effectiveness of three different teaching methods on students' math test scores.
c. A pharmaceutical company wants to evaluate the impact of two dosage levels of a new medication, coupled with differing levels of physical activity regimens and dietary changes, on patients' blood pressure.
d. An environmental organization plans to assess the effect of varying light conditions on the growth of algae in aquatic ecosystems.
e. A psychology research team aims to study the influence of different music genres and modes of listening (through speakers or headphones) on individuals' mood and productivity levels.
f. A sports scientist wants to investigate the impact of hydration levels on athletes' performance during endurance exercises.
g. A food manufacturer conducts a study to determine the effect of preparation methods and cooking temperatures on the nutritional content of vegetables.
h. Researchers plan to investigate the effect of different sleep durations (6 hours, 8 hours, and 10 hours) on cognitive function and alertness levels.
i. A wildlife biologist wants to assess the impact of habitat types (forest, grassland, and wetland) and prevalence of predators on bird species diversity.
j. A social scientist aims to study the effect of social media usage (low, moderate, and high) on individuals' self-esteem levels.

k. A transportation agency aims to assess the impact of traffic congestion levels (low, moderate, and high) on air pollution levels in urban areas.

Exercise 10.18.

a. Describe what statistical controls are and the importance of statistical controls in experimental design.
b. Describe what randomization is and the importance of randomization in experimental design.
c. Describe what replication is and the importance of replication in experimental design.

Exercise 10.19. For each of the following experimental designs, identify an issue with the proposed design, and indicate a possible remedy for it.

a. A pharmaceutical company tests the efficacy of a new drug by administering it to a group of patients with a specific
b. A study aims to investigate the effect of exercise on blood pressure by comparing the blood pressure before-and-after the exercise program for a group of individuals.
c. A psychologist conducts an experiment to examine the impact of music genre on stress levels by randomly assigning participants to either listen to music several different genres of music or to sit quietly for 10 minutes, then measures their stress levels. There are 8 participants in the study, and 7 genres of music being compared.
d. A nutritionist wants to evaluate the effect of a new diet plan on cholesterol levels by recruiting volunteers and asking them to follow the diet plan for one month, then measuring their cholesterol levels at the end of the month.
e. An agriculture researcher tests the effectiveness of a new pesticide by spraying it on one field of crops and leaving another field untreated, then comparing the yield of the two fields.
f. A medical researcher considers the efficacy of a new surgical intervention to a previous pharmaceutical treatment. The surgical intervention is performed on a set of individuals who are deemed healthy enough to undergo surgery, and the pharmaceutical treatment is used for the others.

Exercise 10.20. Suppose you want to investigate the effect of different types of exercise on heart rate. The response variable is heart rate, and the factors are the types of exercise. The possible treatment levels include running, jogging, cycling, and swimming. How would you implement a completely randomized design for this experiment?

Exercise 10.21. You are interested in examining the influence of different teaching methods on student test scores. The response variable is test score improvement, and the factors are treatment is teaching method. The teaching method is defined based on course structure (lecture-based teaching, problem-based learning, and group discussions), the amount of contact time (low, medium, or high), and the format of lesson delivery (slides, whiteboard, or video delivery). How would you implement a completely randomized design for this study?

Exercise 10.22. Suppose you want to investigate the effect of various study methods on exam performance. The response variable is exam score, and the factors are study methods. However, you suspect that students' initial knowledge levels might affect the results. How would you implement a randomized block design for this experiment?

Exercise 10.23. You aim to study the impact of different fertilizer types on crop yield. The response variable is crop yield, and the factors are fertilizer types. However, you suspect that soil quality might affect the results. How can you design a randomized block experiment to account for this?

Exercise 10.24. You are interested in examining the influence of various medication dosages on patient recovery time. The response variable is recovery time, and the factors are medication dosages. However, you suspect that age might influence patients' responses to the medications. How would you implement a randomized block design for this study?

Exercise 10.25. For each of the following variables, indicate whether it is a qualitative or quantitative variable. If it is quantitative, indicate whether it is discrete or continuous. If the variable can fit into multiple categories, explain.

a. Age
b. Gender
c. Height
d. Favorite colour
e. Income
f. Marital Status
g. Temperature
h. Education Level
i. Number of Siblings
j. Blood Type
k. Shoe Size
l. Political Affiliation
m. Grade Point Average (GPA)
n. Type of Car Owned
o. Number of Pets
p. Number of Facebook Friends
q. Opinion on a Political Issue
r. Annual Rainfall
s. Type of Smartphone Owned
t. Level of Agreement on a Survey Question

11 An Introduction to Univariate Descriptive Statistics

11.1 The Purpose of Descriptive Statistics

Suppose that you have collected a dataset, either through a sample of individuals or else through an experiment. You have likely done this with the understanding that these data provide useful insight into the world around you, that they will better inform decisions, or elucidate the truth about questions of interest. However, the data themselves provide very little information directly. Looking through a spreadsheet of numeric values is not a sound way to gather useful insight from data. Instead, we need to rely on pictorial and summary statistics, which take the data and describe or summarize the useful features that we are likely to care about.

We may use tables, charts, graphs, or other numerical summaries. The idea is that we want to use these tools to describe the **distribution** of the dataset. Recall from our study of probability that the distribution of a random variable is the probabilistic behaviour of the random quantity. The distribution of the dataset similarly refers to what values the data take on, and with what frequency those values occur.

We will continue with the same type of notation that has been used throughout our study of probability. A variable, when it has not yet been observed, will be represented by a capital letter, say X. This notation will emphasize the fact that, until we have our sample, X can be thought of as unknown and random. Once we make observations for a variable, we denote these observations as x. So, if we for instance, observe a sample of 100 individuals, we may observe $x_1, x_2, \ldots, x_{100}$ for these individuals.

In general, we will use x_i to represent the ith observation of X, in a sample of size n. We may have multiple different variables that are observed for each individual. In this case, we may use $Y = y_1, \ldots, y_n$ or $Z = z_1, \ldots, z_n$. Generally, the ordering of the data will be arbitrary, which is to say there is no meaningful difference between individual $i = 1$ and $i = 10$, except that $i = 1$ happens to written first in the data. With this notation, we are able to begin considering how to display data for effective summarization. When describing or summarizing data that have been collected, we will consider numeric summaries, tabular summaries, as well as graphical summaries. For numeric and tabular summaries we are thinking about condensing the data that have been collected into key representations of these values. The information will often be presented in the form of data itself – which is to say, a table of summary numbers – however,

it is done so in a way to highlight the key features of the larger dataset. The other alternative is to use graphical displays of information.

The Connection between Data and Random Variables

We previously thought of the collection and generation of data through both sampling and experimental design. The idea was that there was a population or a random process which we wanted to understand, but which could not be fully observed. Instead, we rely on being able to observe partial information from this, through our collected data. When we begin to discuss descriptive statistics, we will be considering the description of the data we have collected themselves – the experimental results or sample. The utility of describing these data stems primarily from the fact that, if we have followed best statistical practices while collecting these data, we *should* find that they are representative of the complete population. We will rarely be able to say that they are *perfect* representations for the population, but we are able to use our insight from them to draw conclusions about the population.

Because of this we constantly need to keep two quantities in our minds: the population and the sample. From our perspective, the sample is data, they are pieces of information that we have actually observed. The population, however, is random. It is unobservable, except through the sample or experiment. If we envision a numeric variable of interest, for instance, then once we have collected our sample we will have specific numeric values for this variable for those people in our sample. Which values we observe depends on which members of the population we happened to sample, and if we were to take another random sample, we would get different realizations for this. As a result, we can view sampling and observing as performing a **statistical experiment**. That is, there is some random quantity of interest, X, which we only see a value for (x) once we take our sample. At that point, we know that $X = x$.

This is *precisely* the notion of random variables introduced earlier on. That is, we view the quantities that we are going to measure as being random in the population, and our observations of them are **realizations of the random variable**. If we knew the distribution of these quantities then we could make different statements regarding the likelihood of observing various events. As a general rule, however, we do not know the exactly population distribution, and instead are trying to make inference about it through the use of the observed values. The notation introduced above emphasizes this point. We use lower case x to represent the values since lowercase letters represent observations that have actually been made. When discussing the unrealized values in the population, these can be expressed as capital letters.

For instance, suppose we conduct a survey of heights in some population. We might say: consider the height of an individual in the population to be a random quantity, denoted by X. If we want to understand $E[X]$ we can do so by first drawing a sample of n individuals from the population. If we draw these individuals independently, then we can think of each individual's height as a random variable, $X_1, \ldots, X_n$ independent

and identically distributed from the distribution of heights in the population. Once we actually select and sample the values we observe $x_1, \ldots, x_n$, where we are saying that $X_1 = x_1$, $X_2 = x_2$, and so on through to $X_n = x_n$. At this point, we have seen actual realizations for X_i, and we can use these realizations to try to draw conclusions about $E[X]$. If we were to repeat this process many times over, each time we did it, we would see different values for X_i, depending on who is included in our sample. Thus, X_i is a random variable.

11.1.1 The Utility of Data Visualizations

Graphical displays of information, or graphs, use visual representations of qualitative or quantitative data in order to provide an overview of key features of the distribution. If used well, this can allow for an efficient display of dense information in a manner which is easily interpretable. The type of informational display used depends, primarily, on two factors: what feature of the distribution are you trying to emphasize, and what type of data are you working with. Broadly, the types of graphics for qualitative data will differ from those for quantitative data. As multiple variables are collected, and the relationships between these different variables becomes the most interesting component of the distribution, we may combine both qualitative and quantitative variables together into a single display.

Historically, there has been a set of graphics which are considered standard, and which would be taught in an introductory course. We will understand the construction and utility of several, common visualizations. However, the landscape around data visualization has rapidly evolved in recent years. Aided by powerful and comparatively straightforward computer programs, far more creativity and artistry has been injected into the world of data visualizations. There are plenty of electronic visualizations which are interactive, there are people effectively using video or audio mediums to add to the display, and the constraints of "standard practice" have largely been overcome. This advancement in technology is not a universal gain, as with every possibility of doing something novel and effective with this technology there are at least an equal number of ways to do something which obscures the truth. Still, data visualization has emerged as a field in its own regard, one which combines statistics, design, and artistry together to great effect.[1]

Because of this, we will not cover the entire suite of historical figures. Graphs such as the stem-and-leaf plot or dot plot, while not entirely without purpose, were created prior to the advent of modern computer graphics. This enabled individuals to construct plots by hand, or with primitive early computers, and these were useful for those settings. The utility of by-hand plot construction is greatly diminished, and the advancement of graphics engines has rendered

[1] To get a glimpse into the world of graphical displays of information, used both well and poorly, it is worth a scroll through of two subreddits: /r/dataisbeautiful and /r/dataisugly. While I do not universally agree with the categorization of these, a lot of the posts at least demonstrate the ways in which modern technology has expanded the potential for creativity.

many of these plots essentially out-of-use. Rather than spend time learning or constructing these, we will instead focus on plots which remain in frequent use.

The Historic Utility of Data Visualizations

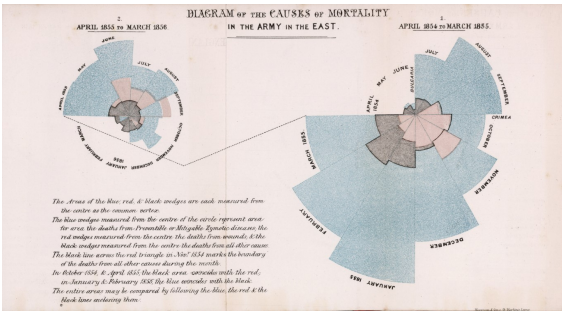
Throughout history there have been many prominent illustrations of the utility of data visualizations. Two prominent examples that come to mind are Florence Nightingale, with her work on causes of death during the Crimean War[2] and John Snow, with his work mapping an 1854 cholera outbreak in London.[3]

Florence Nightingale, a British nurse who worked during the Crimean war, recognized that little was being done to prevent illness transmission throughout the military. She also recognized the importance of conveying the information in ways that would be properly seen and processed by those in positions of decision-making authority. Noting that there was a tendency to overlook tables of figures and data, she instead proposed several graphical displays of the information which would more convincingly illustrate the problem. This served as an important step in the use of data visualizations to tell compelling stories with data.

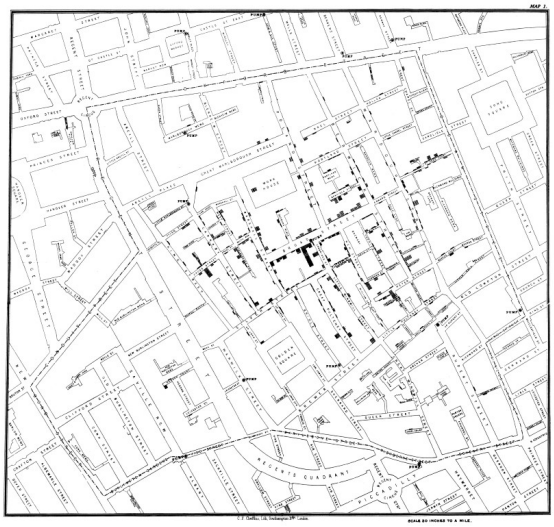
John Snow was an English physician who was an early proponent of the germ-theory of disease. This was proposed as an alternative to the then dominant *miasma theory*, which more or less attributed disease to "bad air". After a cholera outbreak in London in 1854, John Snow mapped out the houses in the are which had individuals sick with cholera. Using this graphic, it became evident that the cases of cholera were clustering around a particular water pump on broad street, lending strong evidence to the thought that this was the cause of transmission. This stands as a foundational study in epidemiology which informs practice to this day.[4]

Figure 11.1: Famous historical graphics produced by Florence Nightingale and John Snow which served great utility in improving health at their time, and stand as an important recognition of the power of visualizing data in a way that renders the message easily interpretable.

(a) One of Florence Nightingale's data visualizations demonstrating death of soldiers during the Crimean War. The blue area in these plots demonstrates the rate of deaths of preventible illness over the course of the war, the red/pink areas represent deaths from wounds, and the black areas represent all other causes This clearly demonstrates the outsized impact that the policies had on the health of soldiers.



(b) The map produced by John Snow illustrating the cholera outbreak. Infected areas are illustrated by the dark shaded rectangles, which appear to cluster around the pump, indicated near the center of the map on broad street. As households move further and further from this central location, the number of cases evidently drops off, providing a strong indication that this may be the cause.



Outside of graphical summaries, we will also consider numeric summaries. These summaries are typically useful for describing particular features of the data which may be of direct interest. These types of summaries are analogous to the summaries we saw for random variables where we condensed probability mass functions into measures of location and measures of variability. When we did this, we lost much of the nuance of the probabilistic behaviour, however, it became far easier to have a general sense of how a random variable will behave. The same concerns will exist in summarizing data. The more we summarize, the more information we will lose, however, the more we will be able to fully appreciate the numeric summaries that we do have. Descriptive statistics is often about balancing these competing interests.

[4]Read more about this in this Scientific American article.

[4]Read more about this in this brief description.

[4]There is some indication that, while the identification of the water pump lead to it being shutoff by the city, the cases of cholera were already reducing by this point. It is thus unclear whether the intervention was timely enough to be effective. However, the impact of the finding looms large to this day.

As previously mentioned, the tools that we will use to summarize data will depend primarily on the type of data that we have. The summaries available to summarize the behaviour of a qualitative variable differ from those available for quantitative variables. We will begin with a discussion of summarizing qualitative variables.

11.2 Descriptive Statistics for Qualitative Variables

Qualitative variables are those which are not numeric. As a discipline descending from math, statistics centers the ability to quantify information in a large number of its techniques. This presents a challenge with the tools that we have to summarize qualitative data. While the modern tendency to fuse graphics and art has enabled graphical displays of qualitative information, at its core, the process of descriptive statistics for qualitative data relies on first translating the qualitative information into numeric information. While the exact procedure for doing this will depend on the exact data in question[5] the most common method for extracting numeric representations from qualitative data is through the use of a **frequency distribution**.

Definition 11.1 (Frequency Distribution)**.** The frequency distribution summarizes the distinct values that a variable can take on, along with the number of observations that equalled each value. The frequency distribution can be thought of as the distribution of drawing a single observation from the sample at random.

The frequency distribution is a useful and intuitive way of summarizing a qualitative variable, numerically. In order to find the frequency distribution, the categories of the variable are listed through, and then the number of observations in each category are tallied up. This can be reported in tabular form, similar to contingency tables[6] or graphically through the use of bar plots.[7] When expressed in tabular form, it can be useful to work out the **relative frequencies** in addition to (or in place of) the counts, giving the proportion of observations for each category. This gives added context to the raw numbers themselves.

Example 11.1 (Charles and Sadie Count Coffee Orders)**.** Sitting in the coffee shop, Charles and Sadie begin to wonder how common the various different coffee orders are. They decide to

[5]For instance, if you are looking at measurements of some quantity over time, time can be used as a "numeric quantity". If the data are geographic in nature, perhaps the locations of different events, then you can use geographic location (latitude and longitude) as numeric data, to place them on a map. These two ideas have been combined to create some very effective and compelling data visualizations. Notably, *1945-1998* is a work of art by Isao Hashimoto, which shows a time lapse video of every nuclear detonation between 1945 and 1998. The video, available on YouTube can be quite affecting. It is very heavy, and while I think a phenomenal representation of the way in which data can be conveyed effectively, please only watch if you are in a position to do so.

[6]Recall from Section 4.5 a contingency table summarizes a frequency distribution in two or more variables.

[7]You can use other types of plots as well, such as pie charts. Most statisticians will be adamant in their disavowal of pie charts because they are typically pretty bad at doing what they set out to do.

categorize each order into one of the following categories, based on what was ordered: coffee only, coffee with food, coffee and non-coffee drinks, coffee with food and non-coffee drinks, food only, food with non-coffee drinks, and non-coffee drinks only. Over the course of an hour observing they collect the following data.

- Coffee + Food + Drink
- Coffee
- Coffee
- Coffee + Drink
- Coffee

- Coffee + Food
- Coffee
- Coffee + Drink
- Drink
- Food

- Coffee
- Food + Drink
- Food + Drink
- Coffee
- Food

Based on these data:

a. Write down the tabular frequency distribution for these data.
b. Write down the relative frequencies for these data.
c. Which order was observed the most? The least?
d. What is the most common order in the population?

Solution

a. We will complete (a) and (b) together in a single table. First note that there are 6 realizations of coffee, 1 of drink, 2 of food, 2 of coffee with drink, 1 of coffee with food, 1 of coffee with food and drink, and 2 of food and drink.
b. This leads to a total of 15 orders in the hour, and so taken together we can write down the following frequency distribution.

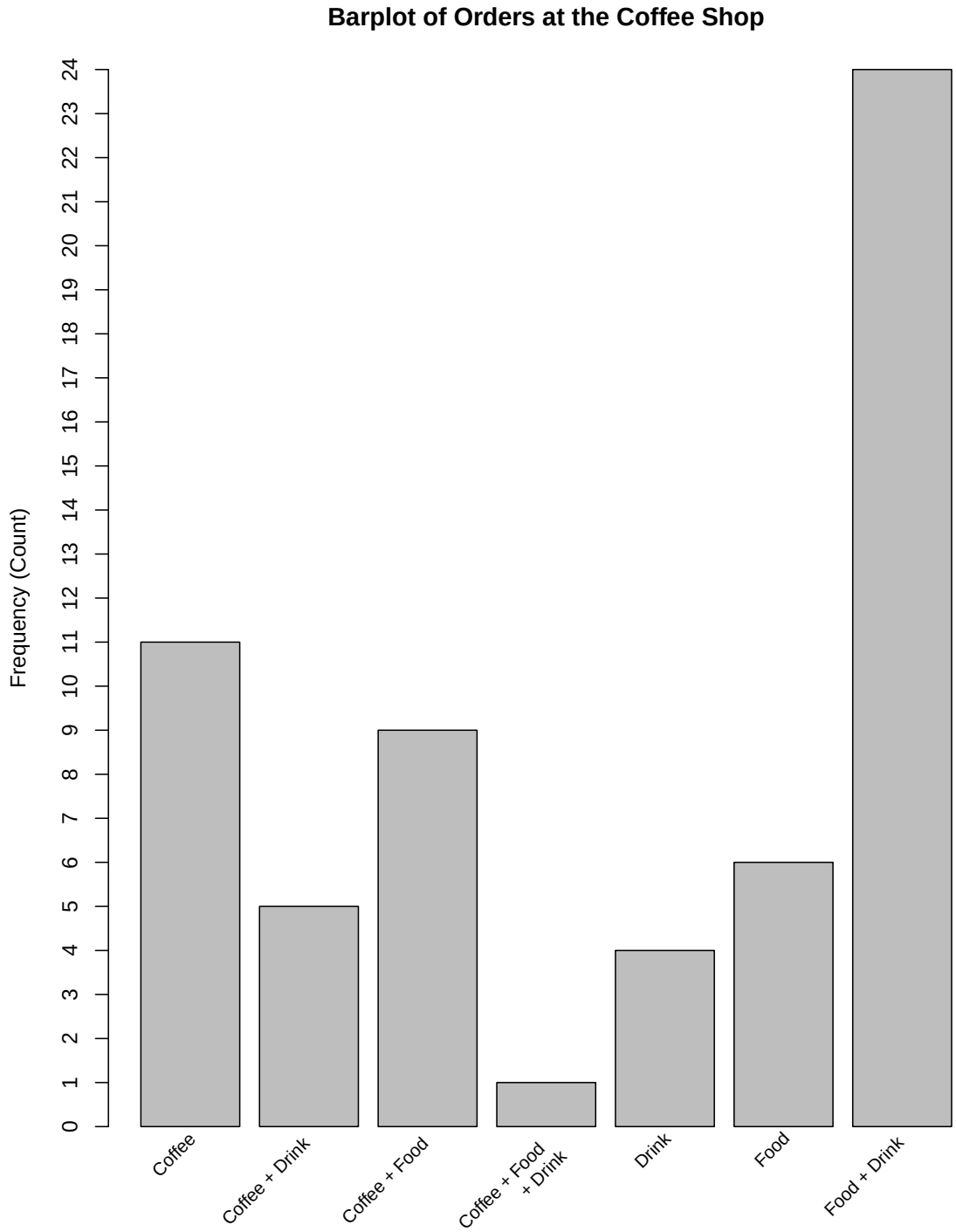
Order	Frequency (Count)	Relative Frequency
Coffee	6	$6/15 = 0.4$
Drink	1	$1/15 = 0.06666$
Food	2	$2/15 = 0.13333$

Coffee + Drink	2	$2/15 = 0.13333$
Coffee + Food	1	$1/15 = 0.06666$
Coffee + Food + Drink	1	$1/15 = 0.06666$
Food + Drink	2	$2/15 = 0.13333$

c. The most frequent order was coffee alone. This was observed by 6 customers. The least frequent orders were drinks alone, coffee with food, and coffee with food and another drink. These were observed 1 time each.
d. These data are from a sample, and by all accounts, not even a random sample. It is important to always keep in mind that there is a difference between population parameters and sample statistics. It is possible that coffee is the most common order in the morning, and that food is far more common later on throughout the day: if the hour watched was in the morning, that could explain this pattern at present. We are only able to **describe** what we observed, rather than **infer** about the population, based on this summary.

To express the frequency distribution in graphical form, we typically will make use of a bar plot. A bar plot is a graphic which along one axis (typically the x-axis, though horizontal plots exist) the distinct values of a qualitative variable are listed. Then, along the other axis, the frequencies of those are listed. The values are displayed based on rectangles with equal width for each category, and with a height that goes out to the value that of the observed variable. Then, to read the bar plot, we observe which rectangles are taller (corresponding to more prevalent values in the sample) or smaller (corresponding to values which were more rare in the sample). We can compare across categories, or even back solve for the entire frequency distribution.

Example 11.2 (Charles and Sadie Count Coffee Orders, Representatively)**.** With the understanding of the flawed methodology that they exhibited on their first attempt, Charles and Sadie decide to perform a random sample to collect data on the different coffee orders made at the local coffee shop. To this end, they randomly select different days of the week, different hours of the day, and then they observe all of the orders that come in over that time. Sadie produces the following bar plot based on their collected data.



Based on this plot, answer the following questions.

a. Which is the most common order in the sample, and what is the frequency with which it is ordered?
b. If there are a total of 60 orders, what is the relative frequency of the least common order?
c. How many orders had any coffee drink in the order?
d. Describe the overall frequency distribution.

Solution

a. In this sample, there are 24 orders of Food + Drink, which is the most common order.
b. The least common order is a coffee + food + drink, which occurred $\frac{1}{60} = 0.01666$ proportion of the time in the sample.
c. In total there were $11 + 5 + 9 + 1 = 26$ orders with coffee in them. This accounts for $\frac{26}{60} = 0.433333$ of all orders.
d. The most common order was food and a non-coffee drink, which was more than twice as common as the next most frequent order of just a coffee. It was very unlikely for people to get all three categories of item. People got coffee with food more frequently than food alone, drinks alone, or coffee with drinks.

Whenever we present a descriptive statistic, be that a numerical summary or a graphical summary, it is always worth asking the question: "what are we trying to highlight?" In the case of frequency distributions, we are typically thinking about highlighting the total counts and cross category comparisons of the different values. Often times these are comparisons that we wish to make and using a bar chart for this is quite effective. However, depending on what we are trying to communicate, there may be alternative choices that make sense to make. It is always important to ensure that your visualization or summary is informed by the goal of your presentation, rather than by outside guidance. Descriptive statistics is fundamentally a field predicated upon communication. With that said, any time that we are presented with an observed qualitative variable, the frequency distribution completely contains all of the information in the data. It may not always be the most useful presentation of the data, however, it is a way of summarizing everything that we know about that variable alone. As a result, deep comfort with frequency distributions will be instrumental to effective communication and description of qualitative data.

Example 11.3 (Charles finds Palmer's Penguins)**.** While daydreaming one day, Charles imagines the chance to work at the Palmer Station in Antarctica, researching penguins. The day dreams lead to a rich imagination, envisioning all species of penguins across the various islands. As the day dreams wind-on, Charles begins to count the penguins, leading to the following observations.

	Biscoe	Dream	Torgersen
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

Using these data[8] answer the following.

a. Write down the complete frequency distribution for penguin species.
b. Write down the complete frequency distribution for the inhabited island.
c. Describe or sketch the bar plots for each of the relevant frequency distributions.

Solution

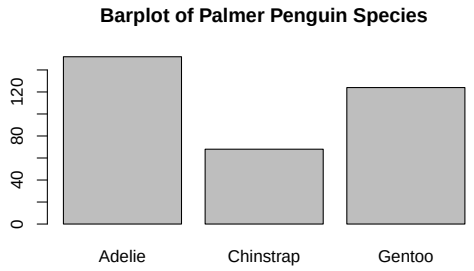
a. To get the distribution for species, we add up along each row to get the total observed data into a frequency table. The relative frequency is divided by 344.

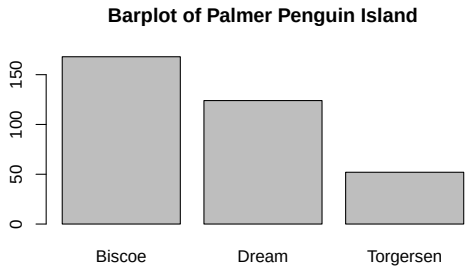
	Frequency	Relative Frequency
Adelie	152	0.4418605
Chinstrap	68	0.1976744
Gentoo	124	0.3604651

b. To get the distribution for locations, we add up along each column to get the total observed data into a frequency table.

	Frequency	Relative Frequency
Biscoe	168	0.4883721
Dream	124	0.3604651
Torgersen	52	0.1511628

c. The following represent the two bar plots for each distribution.





[8]Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package Version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi:10.5281/zenodo.3960218.

11.3 Descriptive Statistics for Quantitative Variables

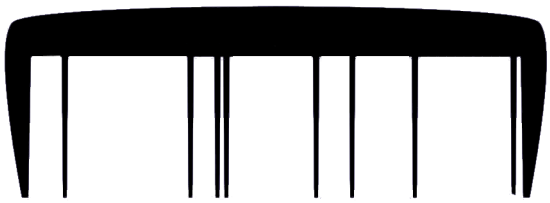
Where our approach for qualitative data was to first summarize the data numerically, and then analyze, with quantitative data the first step is unnecessary. When our data are numeric to begin, we can work directly with them in order to begin to summarize the behaviour of the observed variables. Despite this change in process, the frequency distribution remains an impactful concept in summarizing and describing data which have been observed.

11.3.1 The Frequency Distribution for Quantitative Variables

If the quantitative variables that have been observed are discrete, the frequency distribution can proceed in an exactly equivalent way as in the qualitative case. If, however, we have quantitative variables our frequency distribution needs to be adjusted. The issue is that, if a variable of interest is continuous, we do not expect to ever observe the same value more than once. This renders the frequency distribution to look something like a broken comb[9], rather than having any interesting features. To avoid this happening, we consider the process of **binning** quantitative variables, where values are placed into **bins** or **classes**, consisting of intervals, in order to better understand the structure of the frequency distribution.

Definition 11.2 (Data Binning)**.** Data binning is a (pre-processing) step of a data analysis in which quantitative variables (typically continuous ones) are placed into **bins** or **classes** based on their underlying value. If a quantitative variable x takes values in the interval $[a, b]$, then the interval $[a, b]$ is divided into several sub-intervals, say $[a, p_1], [p_1, p_2], \ldots, [p_{k-1}, b]$. Then, each observed value for x, x_i is placed into its corresponding bin, before the data are analyzed.

As a general rule, bins should be selected either based on some subject-matter justification (such that they are meaningful to the underlying data), or else to accurately balance the trade-offs of smoothness and accuracy in the frequency distribution. That is, we want to select enough bins so that the true behaviour of the data are correctly represented, while not selecting so many that noise and variability are the primary conclusions to be drawn from the summaries. Plenty of methods to select the number of bins have been proposed, and in most software packages for devising frequency distributions various techniques will be implemented. It is worth ensuring that the technique selected for any particular use case accurately summarizes and describes the available data. Generally, $10 - 30$ bins will likely suffice, though fewer or more may be necessary in certain situations.



9

The only hard-and-fast rules of binning is that, first, bins should[10] be of equal width. That is, if you take $[0, 1)$ to be the first bin in your data, then every bin should be of length 1. Second, bins should[11] span the complete range of your observed data. If you have points ranging from 0 to 1000, every value between 0 and 1000 should be contained in some bin. Once binned, quantitative frequency distributions can be described in exactly the same manner that qualitative were.

Example 11.4 (Charles and Sadie Count Coffee Order Items)**.** After their success in understanding the makeup of different coffee orders, Charles and Sadie set their sights on understanding the quantity of items ordered by customers at the coffee shop. The observe customers for an hour and consider the total number of items each customer orders. The following observations are made.

- 3
- 1
- 4

- 2
- 1
- 1

- 1
- 6
- 2

- 3
- 1

- 2
- 1

- 3
- 2

Based on these data, answer the following questions.

a. Write down the frequency distribution for the number of items on each order. Include the relative frequency for each observation.
b. Is data binning required for this frequency distribution? Describe.

[10] Almost always.
[11] Again, almost always.

Solution

a. Here the relevant categories are $\{1, 2, 3, 4, 5, 6\}$. We get

Order Size	Frequency (Count)	Relative Frequency
1	6	$6/15 = 0.4$
2	4	$4/15 = 0.266666$
3	3	$3/15 = 0.333333$
4	1	$1/15 = 0.066666$
5	0	$0/15 = 0$
6	1	$1/15 = 0.066666$

b. No. These data are discrete, and given that there are only 6 total categories there would be no particular utility to binning here. If the data were continuous, or were discrete with sufficiently many categories so as to be better treated as continuous than discrete, then binning would be pertinent.

Example 11.5 (Charles and Sadie Count Coffee Order Values)**.** As a final way of understanding the distribution of different coffee orders, Charles and Sadie decide to observe the total cost of orders for various customers coming through the store. The following observations are made over the course of an hour.

- \$2.25
- \$2.20
- \$5.13

- \$1.30
- \$2.02
- \$4.91

- \$1.64
- \$3.49
- \$0.98

- \$2.97
- \$3.84

- \$5.30
- \$2.53

- \$2.45
- \$4.66

Based on these data, answer the following questions.

a. Describe the considerations that should be made for bin sizes. Would a bin size of \$0.10 be reasonable? What about one that is \$3.00?
b. Suppose that a bin size of \$0.50 is used, starting at 0.50. Write down the frequency distribution.

Solution

a. The smallest observed value is 0.98 and the largest observed value is 5.30. That means that our bins should encapsulate both of these end points, and be evenly spaced throughout. Because there are only 15 data points, we likely want fewer bins rather than more, to ensure that our bins are not predominantly empty or with single items. If we use 0.10, we would require 43 bins at least to include all of the data. This would guarantee that most bins were empty, and is far too small of a divide to be useful. If we used \$3.00, we would span the full range in 2 to 3 bins. This is likely not particularly informative either, this time giving too little of a breakdown of the various values.
b. The following is the relevant frequency distribution.

Bin	Frequency (Count)	Relative Frequency
$[0.50, 1.00)$	1	$1/15 = 0.066666666$
$[1.00, 1.50)$	1	$1/15 = 0.066666666$
$[1.50, 2.00)$	1	$1/15 = 0.066666666$
$[2.00, 2.50)$	4	$4/15 = 0.266666666$
$[2.50, 3.00)$	2	$2/15 = 0.133333333$
$[3.00, 3.50)$	1	$1/15 = 0.066666666$
$[3.50, 4.00)$	1	$1/15 = 0.066666666$
$[4.00, 4.50)$	0	$0/15 = 0$
$[4.50, 5.00)$	2	$2/15 = 0.0.1333333$
$[5.00, 5.50)$	2	$2/15 = 0.0.1333333$

While the tabular representation of the frequency distribution for quantitative variables is a relevant summary, and one which serves a key role, we will see that there are far more ways of summarizing the behaviour of quantitative variables. Before that, however, it is worth determining how to graphically represent a quantitative frequency distribution, through the use of histograms.

11.3.2 Using Histograms for Visualizing Quantitative Frequency Distributions

If we expand the idea of a barplot to quantitative variables, we get the **histogram**. A histogram is primarily useful for displaying the distribution of a single quantitative variable. To do so, the horizontal (x-axis) represents the value of the variable of interest, and the corresponding vertical (y-axis) represents the frequency with which that value occurs in the data. That is, higher points correspond to more frequently occurring values, and lower points correspond to less frequently occurring values.

If the data are binned, then the histogram displays counts within the bins rather than at the values themselves. Just as with a barplot, the graphic proceeds by drawing a rectangle, with a height equal to the frequency, and a width equal to the length of the interval. The larger the rectangle, the more points that were observed in that range.

Sometimes, instead of having the y-axis measure the frequency, we may take it to measure the **density** of falling in that range. The density is given by the probability that a value in that range is observed, divided by the width of the range. For instance, if 10 of 50 observations fell between 2 and 4, then the height of the rectangle using the density representation would be $\frac{10/50}{4-2} = 0.1$. So long as every bin has an equal width, the same relative heights will occur whether using the frequency or density versions.

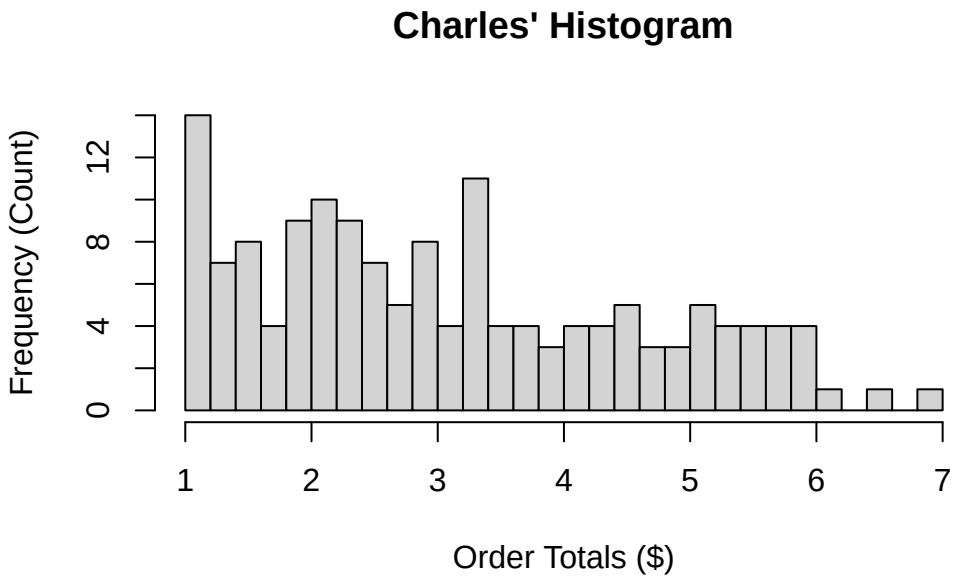
A key difference between histograms and barplots is that, since the data in a histogram are numeric, we typically consider the x-axis to be continuous. This means that the bins of the histograms expand along the complete axis, and adjacent bins will touch one another. In a barplot there is separation between these categories since there are no values between the two of them.

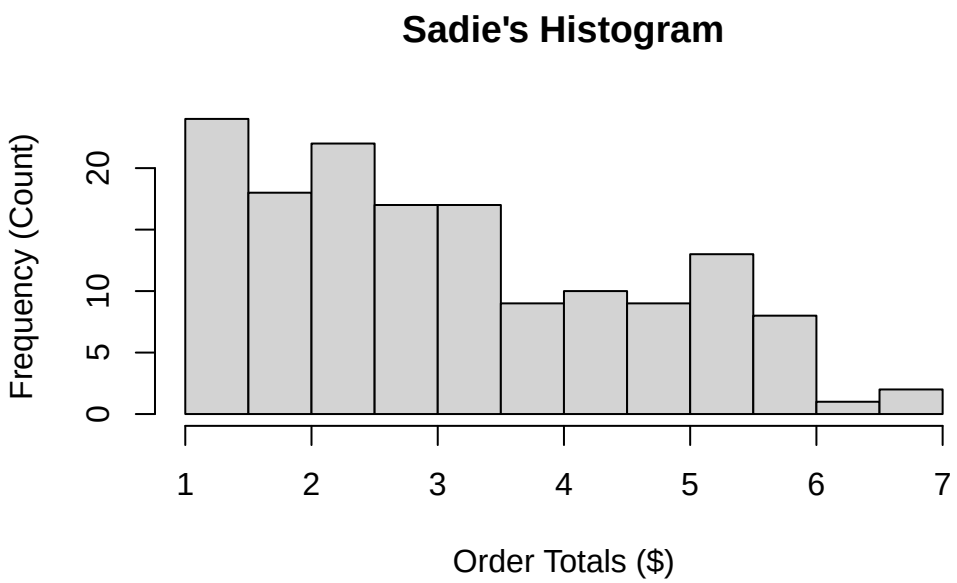
Example 11.6 (Charles and Sadie Plot the Coffee Orders)**.** Charles and Sadie realize that to make sure that they have a full understanding of the total spend that customers have at the coffee shop they should likely collect more data, and data which are spread out randomly over times of the day and days of the week. As a result, they conduct another random survey. Once collected, both Charles and Sadie produce histograms for the totals, as seen here.

a. What is the approximate bin width used by Charles? By Sadie?
b. Describe the frequency distribution as depicted by both histograms.
c. Does one histogram do a better job than the other at representing these data? Explain.

Solution

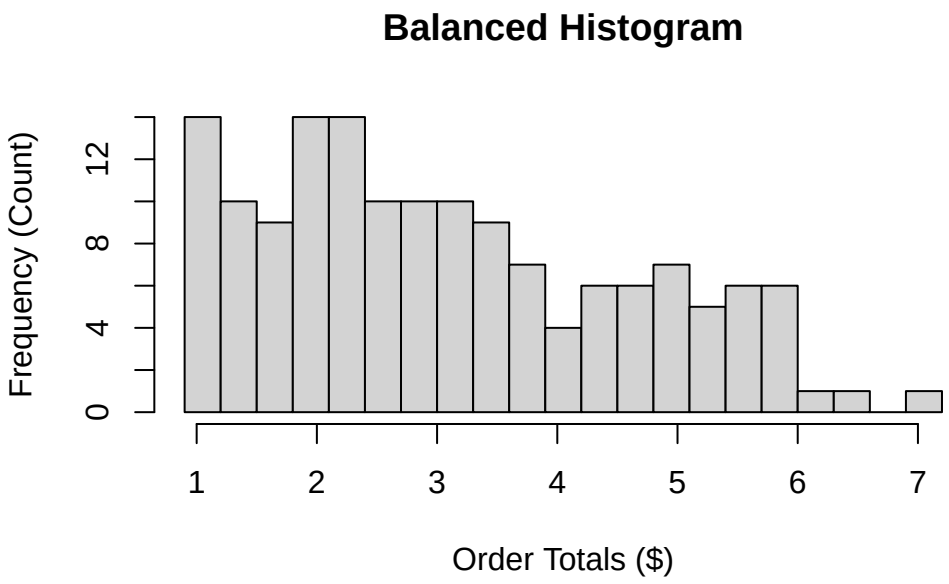
a. We can see that in every dollar along the x-axis, Sadie has two histogram rectangles. This suggests a bin width of approximately 0.50. For Charles, there are 5 per dollar, and as a result, the bin width will be approximately 0.20.
b. Both plots demonstrate that the most frequent order totals are comparatively low, with values around $1.00. With the histogram provided by Sadie, the order totals


Charles' Histogram
Frequency (Count)
0
4
8
12
1
2
3
4
5
6
7
Order Totals ($)


Sadie's Histogram
Frequency (Count)
0
5
10
20
1
2
3
4
5
6
7
Order Totals ($)

are fairly uniform between 1 and 3.50, with a small spike around 2.00. Following that, the order totals are fairly uniform between 3.50 and 6, before falling again above 6.00. The patterns in the histogram from Charles are similar, but with slightly more information. While there is a fairly uniform distribution of observations beyond 3.40, and then a dip beyond 6.00, for the smaller values there is an oscillating pattern. They spike around 1.00, 2.00, 3.00, and 3.40, with the other values being appreciably lower. Still, the lower values are definitely higher on average than the higher values, with roughly equivalent breakpoints as was seen with Sadie's.

c. The preferred histogram here will likely depend on the use case for the data. The data that Charles is demonstrating provides more specific information, however, it is possible that this added information is more noise than useful. It would be interesting to look at, for instance, the prices of various items at the store to see if there was a reason for the peaks: otherwise, it could be seen to be random variation that is not particularly noisy. Sadie's pattern, by contrast, is far more explicable, but it gains this by smoothing over a lot of the fine details within the graphic. Perhaps a graph that balances these two would be more suitable.



A histogram is a useful graphical display since it succinctly summarizes the entire distribution of a particular variable. You can easily see the range of the data, the points which came up frequently in your observations, those which were rare, and how this behaviour is expressed

throughout. It will allow you to readily view points that appear to not fit the trend of the rest of your data, and to investigate a single variable at a glance.

When constructing a histogram, the primary decision that needs to be made is how many bins you should use, or equivalently, how large your bins should be. As you have more and more observations you can typically get away with using smaller bin sizes as, even at the smaller sizes, you likely still have observations that fall into the given intervals. Just as with the discussion on data binning, software that implement histogram construction often provide several techniques for choosing a bin size, or the number of bins, in order to best summarize the data. It is worth considering these for the problem at hand, and ensuring that the choice that is made illustrates the data faithfully.

11.3.3 Characteristics of the Frequency Distribution

While we will typically focus on graphically displaying the distribution of a dataset, it is useful to consider what it is specifically that we are trying to display, and what are the properties of a dataset which are of interest to us? We are primarily concerned with three properties of a distribution: the location, the spread, and the skewness. We have seen all three of these concepts when discussing random variables, but their importance becomes central when summarizing data. More concretely, when describing data we want to make sure to describe the **shape** of the distribution, the **centre** of the distribution, and the **spread** of a distribution. Each of these three concepts have different measures or components which, when taken together, serve as a more complete description of the frequency distribution.

Definition 11.3 (Shape (of a Distribution))**.** The shape of the data distribution refers to the general pattern of points that are observed in the specific dataset. Typically, the shape of a distribution is decomposed into the **modality** and **skewness** of the distribution.

The modality refers to the number of peaks that are visible in the distribution: points that are higher than the surrounding points. A distribution is unimodal with one peak, bimodal with two peaks, or multimodal with more than two.

The skewness corresponds to how symmetric (or not) a distribution is. If a distribution can be mirrored around its center, with the same behaviour above and below the central point, we describe it as symmetric. If a distribution has differing tail-behaviour, extending out in a direction, we say that it is skewed. A distribution that has a long-tail to the right is called **right skewed** or **positive skewed**, where a distribution that has a long-tail to the left is called **left skewed** or **negative skewed**.

To describe the shape of a frequency distribution, we describe the modality and the skewness of it. These two features combine to give a good sense of the general picture of the distribution, such that someone should be able to sketch a reasonable approximation to the frequency

distribution from the description. However, there are many distributions with the same modality and skewness, which are otherwise quite distinct. To understand these differences, it is useful to turn towards measures of location, or central tendency, and measures of spread, or variability.

Definition 11.4 (Location (Central Tendency))**.** The location of data, also referred to as the central tendency, is a description of where observations in the dataset tended to fall around. This can be measured as the sample mode, the sample median, or the sample mean, and is typically summarized using all three. Measures of central tendency give a sense as to where the middle of the data were, where various definitions of middle can be used.

Just as with random variables, measures of location come about by asking what we "expect" to see in the data. When we are discussing samples, rather than random variables, we are instead answering questions around what we saw on average, or what we tended to see in the observed data. Location summaries are quite common, and quite intuitive. You may indicate the most common value, which is the sample mode, or the overall average, the sample mean. However, just as with random variables, the measures of central tendency only tell a partial story. The other key feature of the data is the spread.

Definition 11.5 (Spread (Variability))**.** The spread of data is a measure of how separated the data are, and how they tend to be spaced around the location. The spread can be captured using particular measurements, such as the sample variance, sample IQR, or sample range as was done with random variables. The may also refer to the tail behaviour of the data, which looks at how likely values are as they move further and further from the center of the data.

A distribution is said to be **heavy-tailed** if points that are far from the measures of central tendency are quite frequent in the data, and is said to be **light-tailed** otherwise. These concepts can be formalized more rigorously, however, it is often taken to be an informal rather than formal check.[12]

The spread of the data gives a measure as to how concentrated (or not) the observations were. Data which are widely spread out will have large measures of variability compared to those which are less spread out. When combined with the measures of central tendency, as well as a description of the shape of the distribution, it is possible to develop a fairly clear picture of the behaviour of the data, summarized rather succinctly.

11.3.4 The Shape of a Distribution

As indicated, the shape of a distribution is primarily defined by the modality and skewness of the distribution. That is to say, when asked to describe the shape of the distribution, you

[12]Which is to say, we rely on the vibes of the tails, rather than their mathematical behaviour explicitly.

should report on the modality (including the values of the modal points), as well as on the symmetry or skewness of the distribution.

Definition 11.6 (Modality)**.** Modality refers to the number of **local** peaks that a frequency distribution has. That is, the number of times that there are values in the frequency distribution that are higher than those in close proximity to them. If looking at the histogram we are looking for the number of "hills" that exist. Modality is classified by the number, and values, of the different modal points.

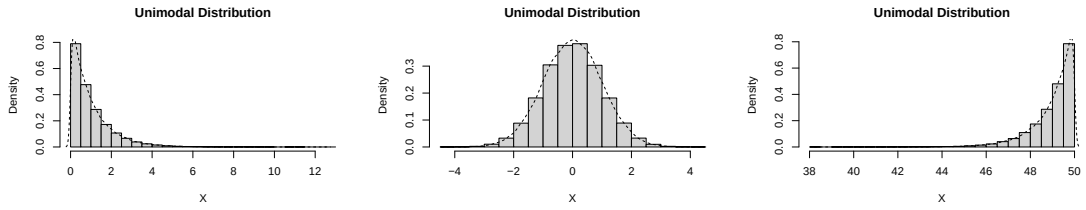
A distribution with one local maximum is considered **unimodal**. A distribution with two local maxima is considered **bimodal**. A distribution with three or more local maxima is considered **multimodal**.

It is important to emphasize that a frequency distribution may have only one mode, but may be multimodal. That is, we do not require each of the peaks to tend to equate to exactly the same level to be considered peaks. Instead, we compare them to only the points that are around them. This way, we can capture the idea of *local behaviour* indicating that certain regions appeared more frequently than others around them, which is often of direct interest to us. It is also important to recognize that, if reading modal points from a histogram, the number of breaks and the number of observations will likely change the perception of the modality. There will often be judgment calls when discussing the number of modes that a histogram exhibits, with reasonable disagreement being possible. As a general rule, you should not consider small noisy peaks adjacent to others to be additional modal points, unless there is a good reason to do so. If you envision drawing a smooth line over the full distribution of data, the modal points will come where you draw the crests of the hills.

Identifying the Modality of Distributions

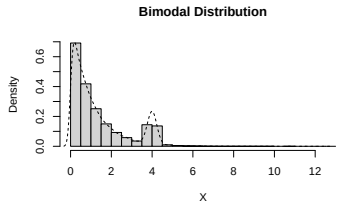
Unimodal

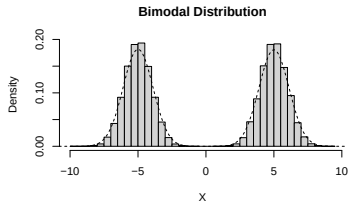
Data which exhibit a single peak, whether this is in the center or off to one side, are considered to be **unimodal**.

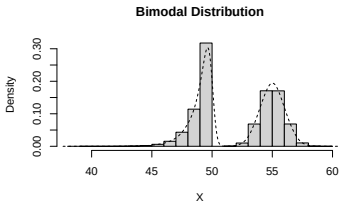


Bimodal

Data which exhibit exactly two peaks are considered to be **bimodal**.

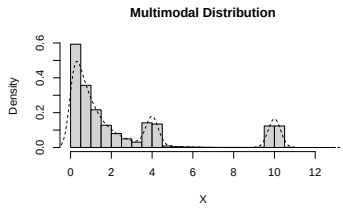


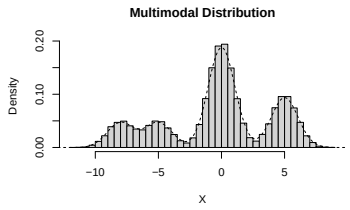


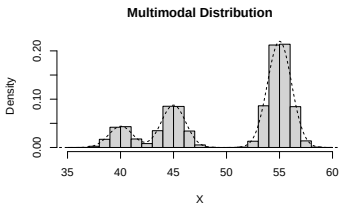


Multimodal

Data which exhibit more than two peaks are considered to be **multimodal**.







Beyond modality is the skewness. Skewness, or conversely, symmetry is a way of describing the **tail behaviour** of a distribution. As you move away from the central values, the most common values, or the middle values of a distribution, the tails are the values that are far from where you started but still present in the data. In general, if the tails going to the positive and negative directions look similar, the data are said to be symmetric. Otherwise, the data are said to be skewed. We differentiate skewness based on the direction that the tail travels.

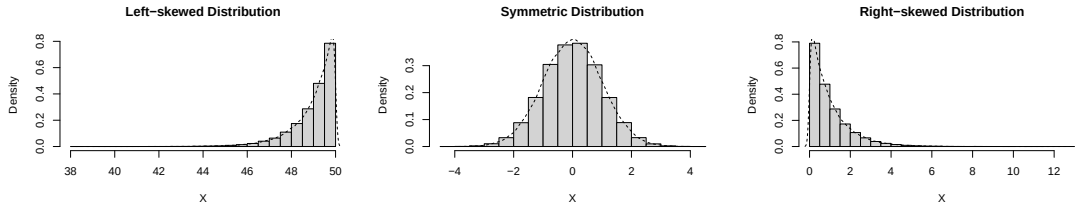
Definition 11.7 (Skewness)**.** Data which are nonsymmetric are said to be skewed. The lack of symmetry can be identified by the behaviour of the tails of a distribution, differentiating between **positive** (or right) skew and **negative** (or left) skew.

Data are right-skewed if the tail is longer to the righthand side of the figure. Data are left-skewed if the tail is longer to the lefthand side of the figure.

Sometimes skewness is quite dramatic, being very evident in which direction the skew will be. In other cases the data are not symmetric, but are also not evidently skewed. In these settings it is worth investigating the data in slightly more depth to try to understand whether the lack of symmetry (or skewness) can be explained based on some particular values, and if the remaining data exhibit a more predictable pattern.

Identifying the Skewness of a Distribution

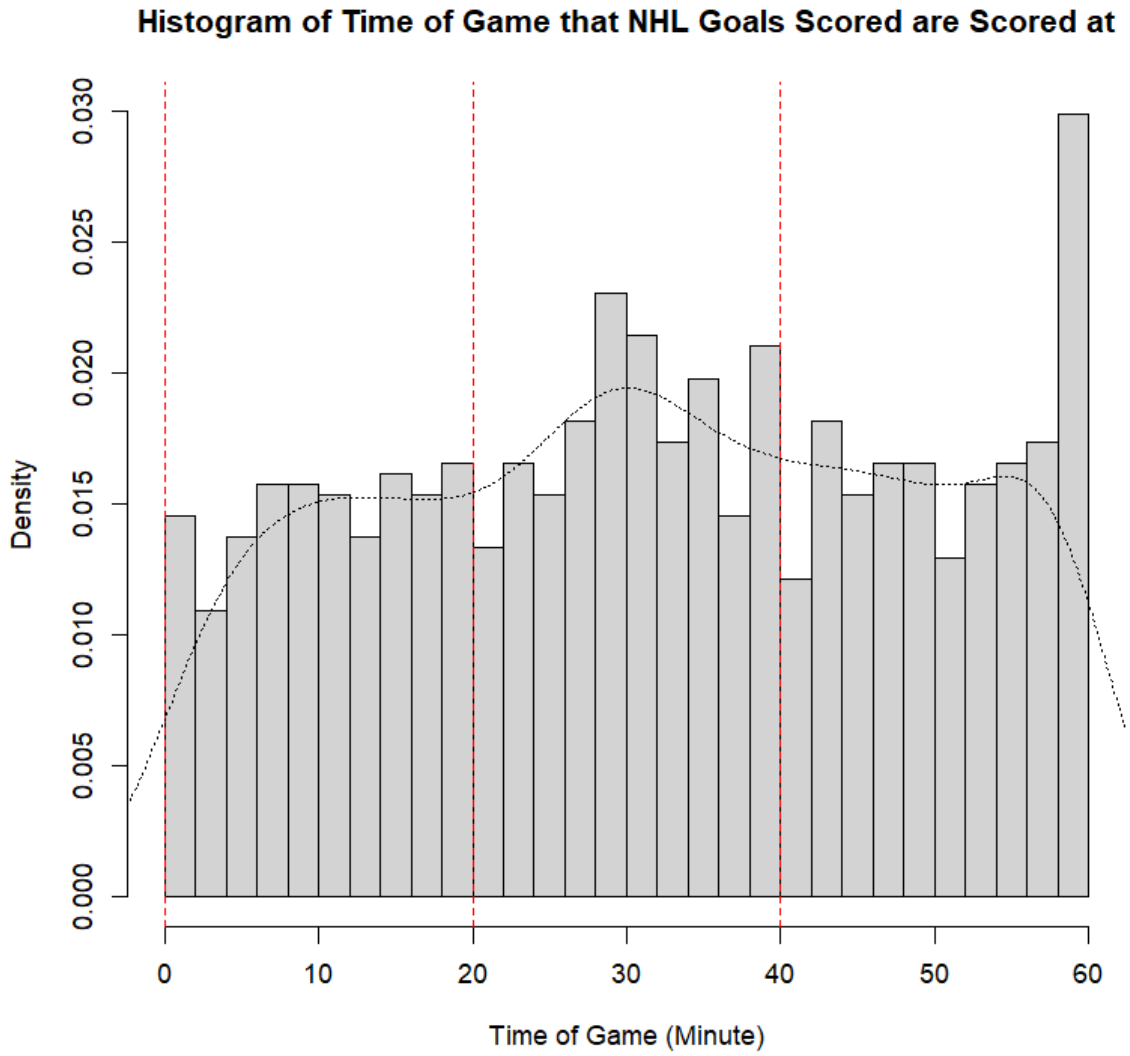
To identify whether data are symmetric, right-skewed, or left-skewed, you consider the beahviour of the tails.

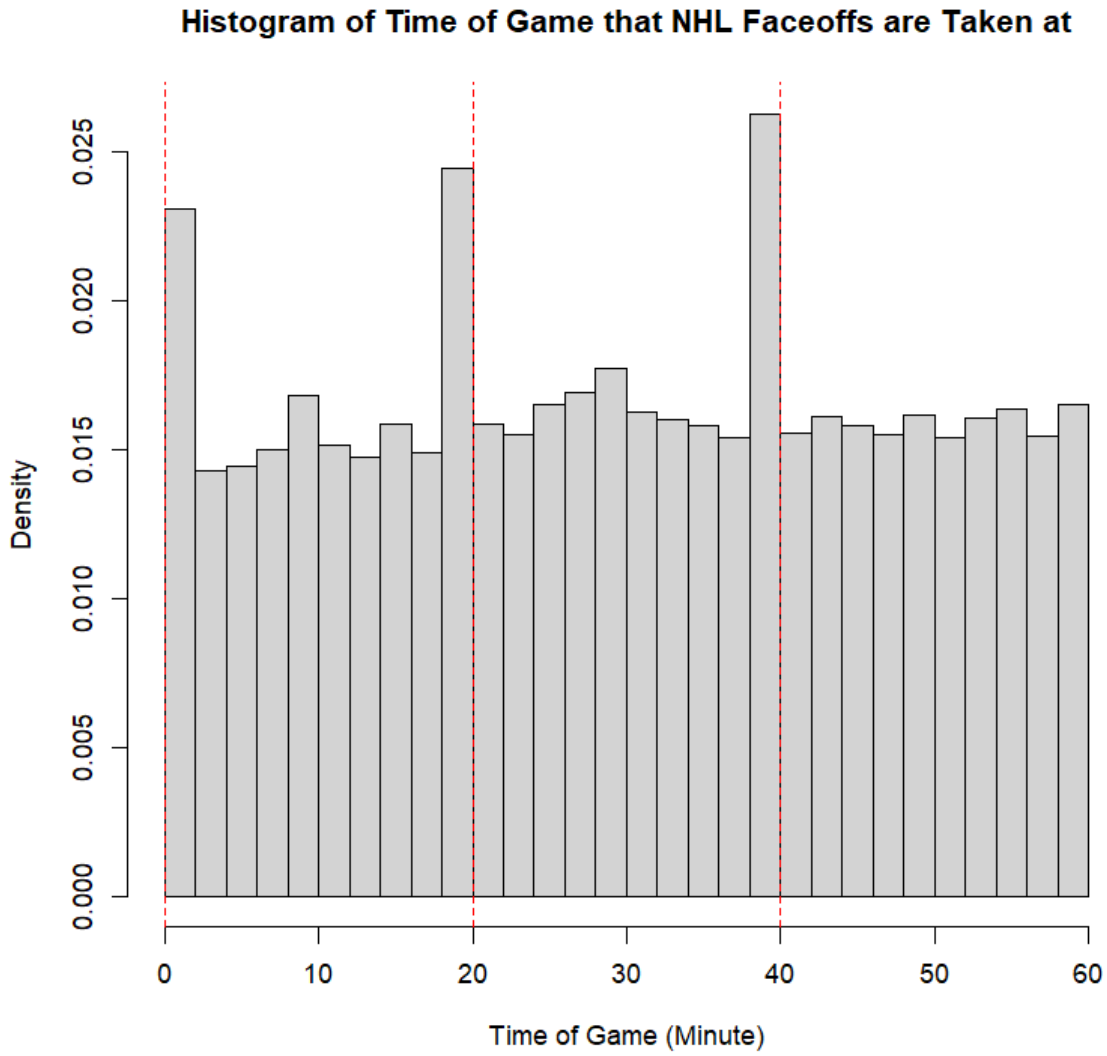


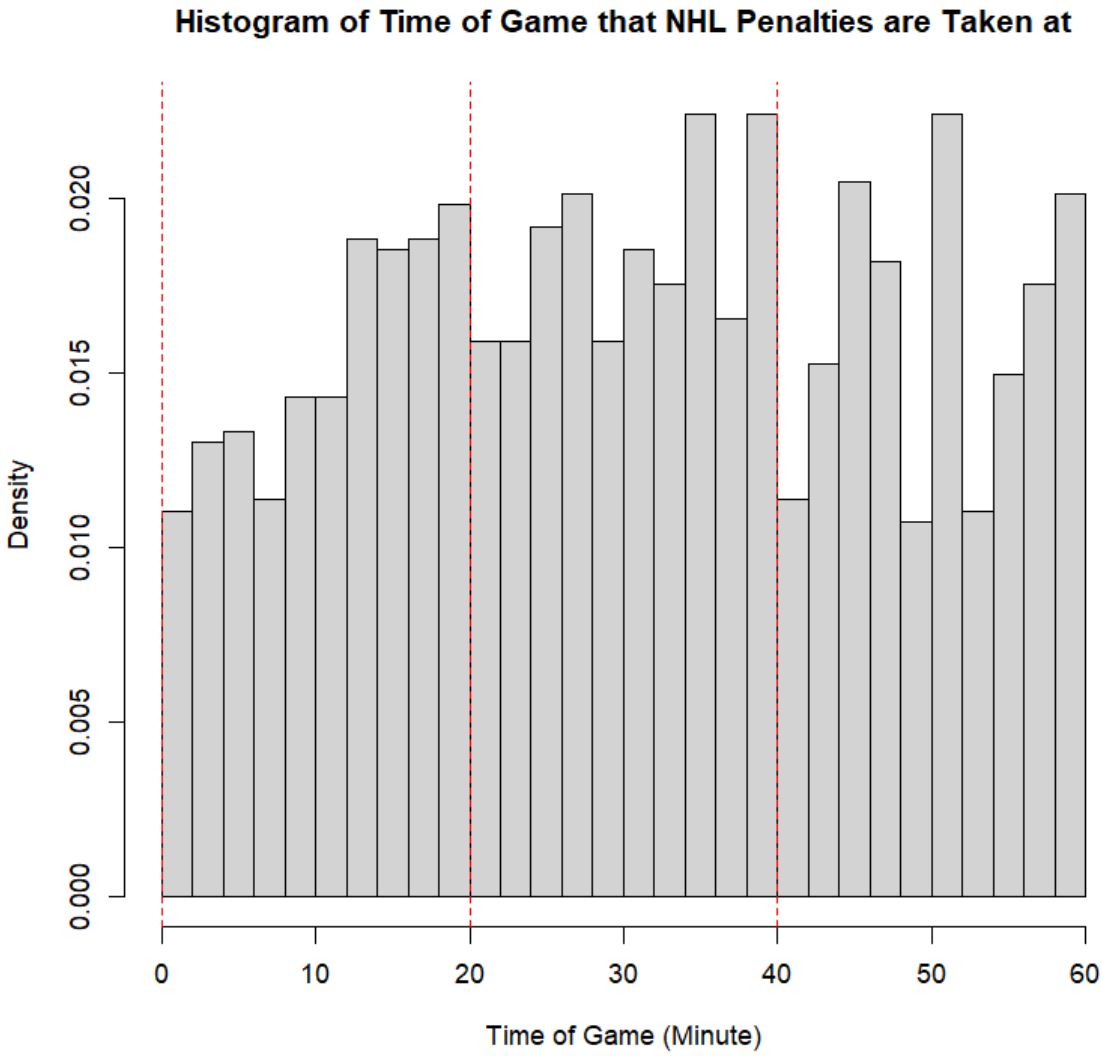
As a result, when asked to identify the shape of a distribution, you are being asked about the key features of the distribution: how many modal points are there, and where are those located, and what is happening with the tails of the distribution? Beyond these points, discussion of the shape of the distribution largely centers around giving added context to the nuance with the provided descriptors. For instance, if the skewness is not partiuclarly pronounced, that can be discussed. If modal points are not clearly delineated, that should be acknowledged. Once described, an individual should be able to sketch out a rough distributional curve that approximately mirrors the behaviour of the frequency distribution.

Example 11.7 (Sadie Records the Timing of Hockey Events)**.** Sadie, being a big sports fan, has started to record and analyze the timing of different events throughout the game. Charles decides to help out by producing histograms for the time that these events happen at. Sadie records the time that faceoffs occur at, that goals are scored at, and when penalties are taken.[13] The histograms are provided below.

[13]For reference, NHL games are 60 minutes long (potentially longer if no one is winning at the end of the time), divided into 3 periods. When play is stopped, a face-off takes place to start the play again. Penalties occur when rules are violated by one of the teams. On the histograms, the start of the three periods are indicated in red dotted lines. These data come from a random sample of 200 games during the 2022-2023 NHL regular season.


Histogram of Time of Game that NHL Goals Scored are Scored at
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
10
20
30
40
50
60
Time of Game (Minute)


Histogram of Time of Game that NHL Faceoffs are Taken at
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
10
20
30
40
50
60
Time of Game (Minute)



a. Describe the shape of the distribution of goal times in NHL regular season games.
b. Describe the shape of the distribution of faceoff times in NHL regular season games.
c. Describe the shape of the distribution of penalty times in NHL regular season games.
d. Indicate any difficulties in describing these distributions.

Solution

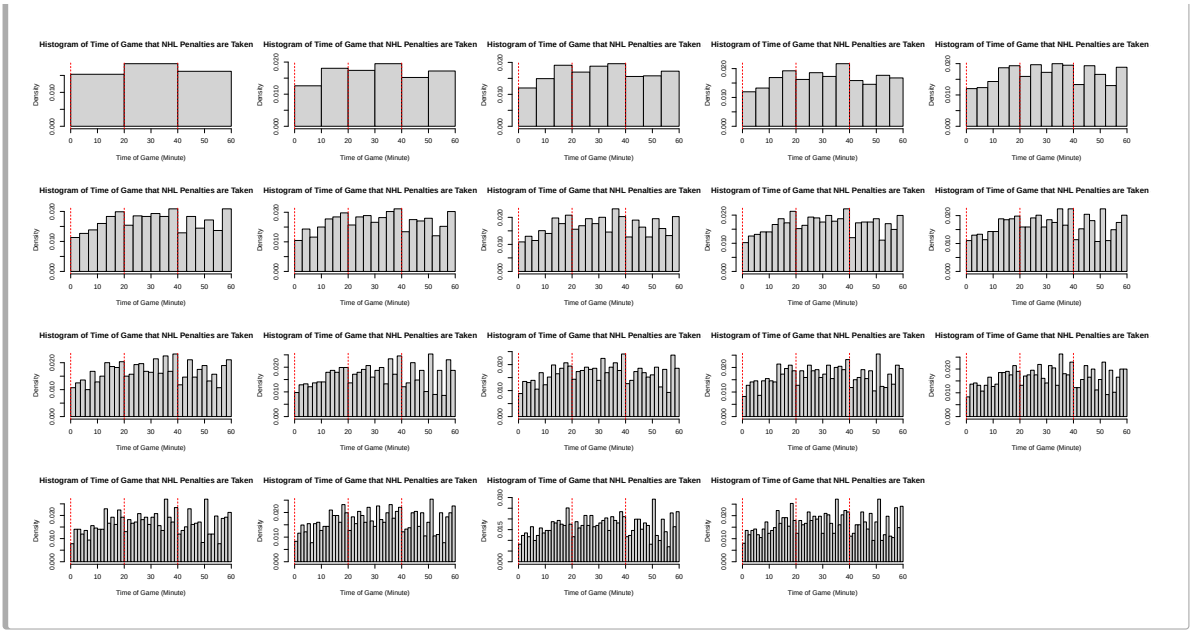
a. The distribution is approximately **bimodal**, with modal points at approximately 30 and 60 minutes. The distribution is non-symmetric, since higher frequency of goals are scored near the end of the game than at the beginning, making it a left-

skewed distribution. If the final bin is ignored, which may be reasonable given that the final minute of game has different dynamics than other times, the distribution appears roughly symmetric.

b. This distribution is approximately **multimodal**. The modes occur at 0, 20, and 40 minutes. Every period starts with a faceoff, and as a result, every game will have faceoffs taken at exactly 0, 20, and 40 minutes into it. There are some points which may be described as modal points around 10 minutes and around 30 minutes, but these are far less obvious than the other three. The distribution is non-symmetric owing to the lack of a modal point at 60 minutes. Because of this, we may say that there is a slight positive skew to the distribution. It seems more sensible, however, to indicate that in the absence of the three modal points that are clearly explicable, it is a roughly symmetric distribution that is approximately uniform across the whole range.

c. The modality of the penalty timings is difficult to describe. It may be reasonable to describe this as multimodal with modes appear around 20, 25, 38, 45, and 60. However, this may also represent some histogram noise that is better to smooth over. In that case, perhaps you suggest that the three modal points are around 20, and 40, and around 50. The distribution does not exhibit perfect symmetry, owing to the spike near the end of the games, however, there seems to be little apparent skewness in the distribution. With the points near the end of the game removed, it is a mostly symmetric plot, however, with those points in there is some negative skewness indicated.

d. While the first two distributions are fairly clear to describe in shape, the penalties are a lot less evident. This becomes even more apparent when we consider how the number of bins selected impacts the overall shape of the distribution. Consider the following histograms (if online, you can click to enlarge them). With few bins, this distribution appears to be approximately symmetric with a single mode in the middle of the distribution. As more and more are added, the pattern remains rather similar until a within-period distribution emerges: the first period has a negative skew, unimodal distribution, the second a fairly uniform symmetric distribution, and the third a bimodal negative skew distribution. As we continue to add bins we see histograms that look similar to the one we considered, before getting to a point that looks more or less consistent throughout the range, with seemingly random spikes at various times. These demonstrate the importance of bin size selection, and illustrate the challenges that can occur when trying to describe real-world data.



11.3.5 Measures of Location

The shape of a distribution is described in fairly general terms. If you know that a distribution is right-skewed and unimodal, with a peak around 10, there are many plausible distributions that could be drawn for what this would look like. While often times this shape captures the most pertinent information for a distribution, sometimes we require more. To add specificity, we can consider the location or central tendency of a frequency distribution. The main measures of location are the sample mean, sample median, and sample mode. These correspond to the exact quantities that we saw in random variables, this time computed on the data directly.

Definition 11.8 (Sample Mode)**.** The sample mode is the most common value observed in a dataset. When there are more than one value which appear equally often, these are all consider modes. If a variable is binned, we may define the mode in terms of the classes rather than the specific value, depending on the context.

Whichever value appears most often is given as the mode. This is analogous to the most probable value for a random variable. The sample mode and the modal points are related to one another. When discussing modality, we contented ourselves with approximations, considering values *near* the actual modal points when defining the peaks. When discussing the sample mode we are looking for a specific value, or a specific category of values, which actually gives the most frequent value.

Definition 11.9 (Sample Median)**.** The sample median is the middle point after ordering the observed data. If there are an even number of observations it is the mean between the two middle points. If we order the observed data as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, then the median is defined as

$$\text{Median} = \begin{cases} x_{([n+1]/2)} & n \text{ is odd;} \\ \dfrac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & n \text{ is even.} \end{cases}$$

That is, it is the middle point when there is an odd number of observations in the data, and it is the average of the two middle points when there are an even number of observations.

The sample median has the same interpretation as the population median we previously saw. There will always be 50% of the observations which are less than or equal to the median, and always \$50% of the observations which are greater than or equal to the median. This puts the median as the center of the distribution, when measured in terms of frequencies.

Definition 11.10 (Sample Mean)**.** The sample mean is the standard arithmetic average. If we observe $x_1, \ldots, x_n$, then we write the mean as $\overline{x}$, and this is calculated as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The mean is a very commonly reported measure for the center of a distribution. it is also referred to as the average. Just as with random variables and the expected value, the mean can be viewed as balancing the mass of observations. If you place equal mass at each of the observations, then the mean would be the point which balances a seesaw holding those masses.

Example 11.8 (Charles' Penguin Bill Lengths)**.** Continuing on in day-dreaming adventures at the Palmer Station in Antarctica, Charles envisions recording the bill lengths of a random sampling of the penguins that are observed. These day-dreamed values are recorded, and Charles would look to summarize the general behaviour of these points, considering the measures of central tendency of them.

49.0	37.8	45.8	39.0	43.2	48.8	37.8	49.1	40.9	37.3

a. What is the sample mean of these data?
b. What is the sample median of these data?
c. What is the sample mode of these data?
d. What is the expected value for the bill length in the population? Explain.

Solution

a. For the sample mean, we compute

$$\begin{aligned}\overline{x} &= \frac{1}{10}(49.0 + 37.8 + 45.8 + 39.0 + 43.2 + 48.8 + 37.8 + 49.1 + 40.9 + 37.3) \\ &= \frac{428.7}{10} \\ &= 42.87\end{aligned}$$

b. For the median, we first consider ordering the values in ascending order.

37.3	37.8	37.8	39.0	40.9	43.2	45.8	48.8	49.0	49.1

Then we note that the two middle values are 40.9 and 43.2, so we get that

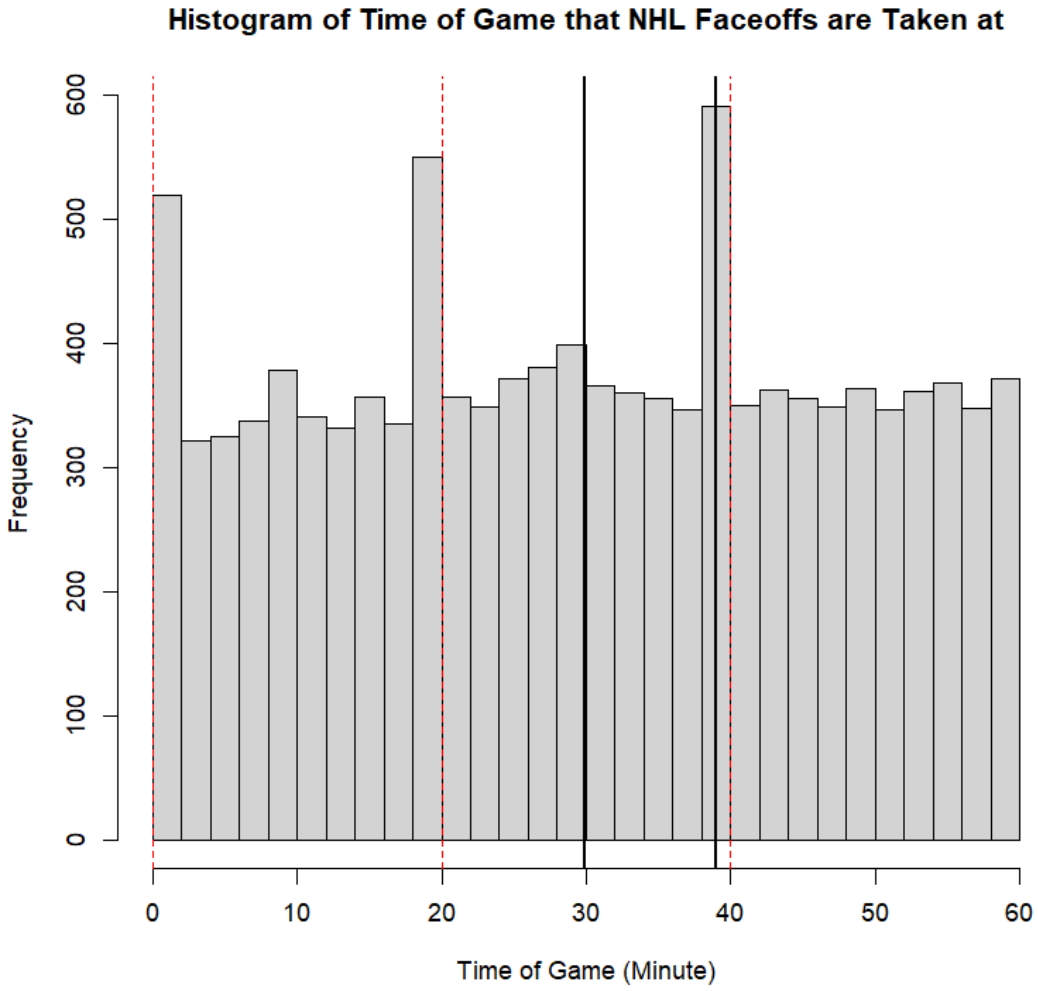
$$\text{Median} = \frac{1}{2}(40.9 + 43.2) = 42.05.$$

c. The only repeated value in these data is 37.8, and so that makes it the mode.
d. We do not know, given this information, what the population expected value will be. We do not have access to enough information to compute the parameter, and instead must rely on using only our sample statistics. These are measures of the data that are observed, rather than the full population.

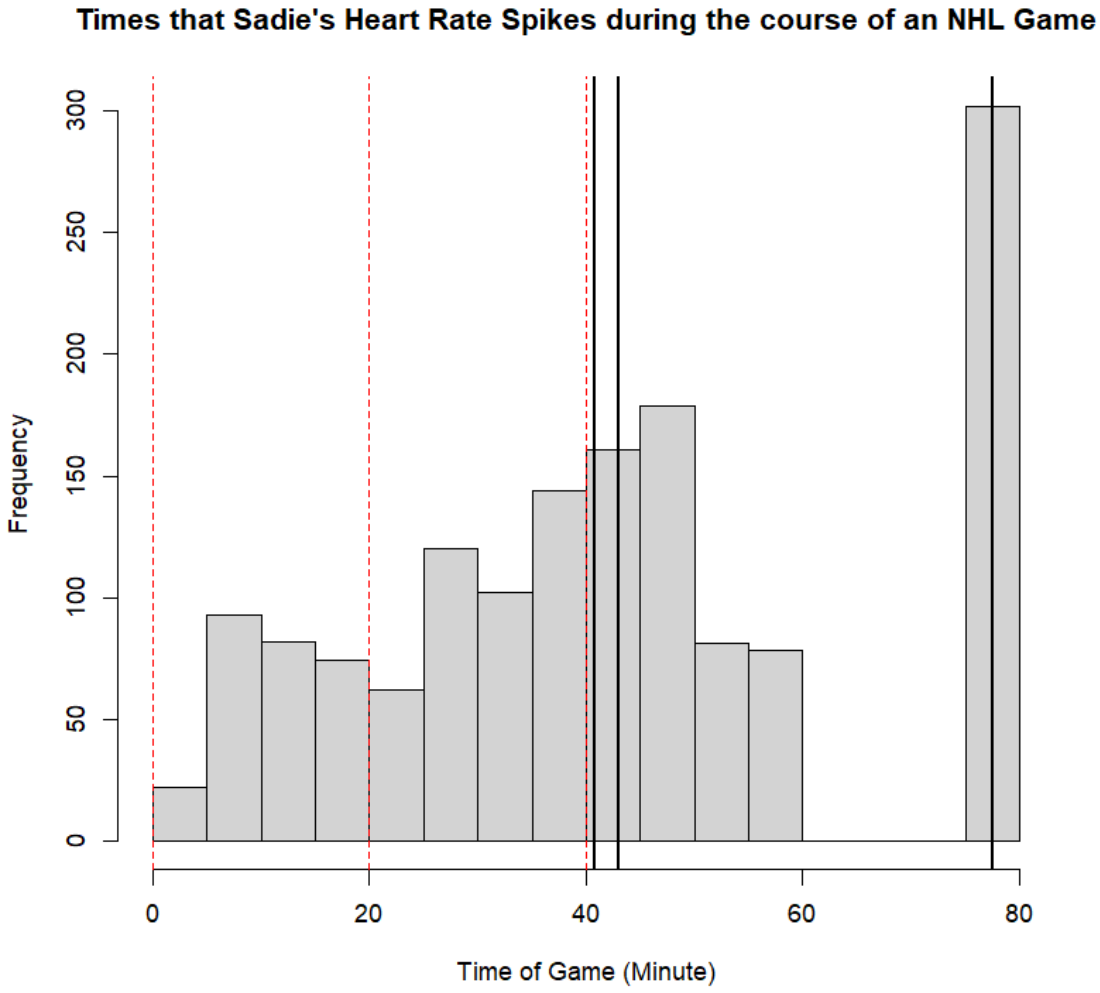
The three common measures of central tendency will often be explicitly computed and reported when access to the data is directly available. These can also be approximately indicated using a histogram of a frequency distribution. The mode will be the bin with the highest frequency, or equivalently, the highest point on the histogram. The median will be found in the bin which contains the middle observation. This can be challenging to find exactly without counting, but an approximation is likely possible. The mean will be found in the bin which balances the mass of the distribution. You can imagine asking yourself: where would the fulcrum need to go in order to balance a seesaw with these weights on it. The answer will tell you where the mean is. Note that this process, without explicit observations, will not be exact. Instead, it is in our interest to attempt to find approximately correct solutions to these questions, getting a general sense of the measures of central tendency from a graphical representation.

Example 11.9 (Unknown Histogram Markings)**.** Charles wanted to help Sadie with the hockey analysis from before. To do so, Charles worked out the mean, median, and mode for each of the distributions, and indicated this on the histograms with black vertical markings. Unfortunately, Charles does not remember which marking is which. For each of the following

graphics, indicate which of the three solid vertical markings corresponds to the mean, median, and mode, or describe why it is not possible to tell.



a.



b.

Solution

a. In this plot only two markings are differentiable from one another: one in the bin immediately before 30 and one in the bin immediately before 40. We know that the mode of the distribution falls into the bin around 40, and so the two lines here indicate the mean and the median. With the plot we should expect that the mean is slightly higher than the median, since the tail extends ever so slightly beyond symmetry to the positive side.

b. Here we can discern the mode to be marked around the 80 minute mark. The other two markings, for the mean and median, are marked in the bin just beyond 40. Here we know that the mean should be higher than the median, since the outlying spike later on will have the effect of pulling-up the observed mean, without impacting

the median. Thus, we observe in order the median, mean, then mode.

It is also important to remember that, for each of these quantities, when computed on a dataset, we refer to them as *sample* measures. That is, we call the mode the **sample mode**, the median the **sample median**, and the mean the **sample mean**. This terminology emphasizes that our calculations are not with respect to a theoretical random variable with some assumed probability distribution, but rather from a sample of data that was actually observed. Recall that we view our sample as being **realizations** of a random quantity, either as a random process or from a larger population. That is, we can think of these measures as **statistics** computed on a sample, rather than the **parameters** that could be computed on the population.

11.3.6 Measures of Spread or Variability

When we introduced the concept of the expected value, and other measures of location of a random variable, we indicated that these would be insufficient to accurately summarize the behaviour of a random quantity. The same is true of a frequency distribution. Once again, by complementing the measures of central tendency with measures of spread, we are better able to understand the data which were actually observed, and use this to inform our understanding of the data. Combining variability, with central tendency, and distributional shape gives a good overall picture of what was observed, in a digestible summary form. The primary measures of variability for a dataset's distribution are the same quantities used to measure variability of a random variable: the (sample) variance, (sample) IQR, and (sample) range.

Definition 11.11 (Sample Range)**.** The sample range is the observed range in the data set. We define the sample range to be the difference between the maximum observed data point and the minimum observed data point. That is, if the ordered data is observed as

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

then the sample range is

$$\text{Range} = \max\{x\} - \min\{x\} = x_{(n)} - x_{(1)}.$$

Just as with the range of a random variable, the range may be reported as either the distance between the minimum and maximum, or else as the minimum and maximum points themselves.

Just as with random variables, the sample range gives a rather coarse view of variability within a datset. The sample range does give information regarding the values that are *possible*, based on what was observed within the data, but it does not necessarily provide a reasonable representation of which values were commonly expressed throughout the data. A single outlying point can drramatically impact the range, without meaningfully changing the observed patterns. For this reason, we will often consider the sample IQR instead.

Definition 11.12 (Sample Interquartile Range (IQR))**.** The sample interquartile range is the difference between the first and third quartiles in the dataset. That is, it is the length that spans the middle 50% of observations within the variable. The sample IQR is computed as $\text{IQR} = Q3 - Q1$, where $Q3$ and $Q1$ are the third and first quartiles.

In order to compute $Q1$ you compute the median of the first half of the data. In order to compute $Q3$ you compute the median of the second half of the data. If there are an odd number of points, the central point is computed in both. That is, taking

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

then for the first quartile,

$$Q1 = \begin{cases} \text{Median}\{x_{(1)}, x_{(2)}, \ldots, x_{(n/2)}\} & n \text{ even} \\ \text{Median}\{x_{(1)}, x_{(2)}, \ldots, x_{([n+1]/2)}\} & n \text{ odd} \end{cases}.$$

The third quartile, $Q3$, is computed similarly as

$$Q3 = \begin{cases} \text{Median}\{x_{(n/2+1)}, x_{(n/2+2)}, \ldots, x_{(n)}\} & n \text{ even} \\ \text{Median}\{x_{([n+1]/2)}, x_{([n+1]/2+1)}, \ldots, x_{(n)}\} & n \text{ odd} \end{cases}.$$

The interquartile range has the same benefits when compared to the range in a sample as it did for random variables. Outlying points that substantially deviate from the trends that are actually observed do not make a large difference on the sample IQR, where they will on the sample range. This can be desirable for understanding the variability in *most* of the data. Just as with random variables, the range and IQR are analogous in that they give a full representation of how spread out the data are. It is also possible to conceive of variability as how far from *average* data tend to be. For this, we consider the sample variance (and standard deviation).

Definition 11.13 (Sample Variance)**.** The sample variance is an analogue to the population variance, measuring how far data are from the sample mean, on average. To compute the sample variance we take the following form

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Note that the division here is by $n-1$ rather than by n.[14] Notice that if n is large, dividing by n or by $n-1$ will give roughly the same results.

[14] This makes the sample variance **unbiased** for the true variance. If you conceive of the sample variance as a random quantity, one that depends on what sample you actually take, dividing by $n-1$ rather than by n will make it so that $E[S^2] = \text{var}(X)$, where if you divide by n this will not be the case.

The sample variance has the same underlying concern as the variance of a random variable: namely, it is a squared quantity. As a result, we will often consider the **sample standard deviation**, which is given by the square root of the sample variance, as an alternative representation of the sample variability.

Definition 11.14 (Sample Standard Deviation)**.** The sample standard deviation is an analogue to the population standard deviation, giving an approximate measure of the mean deviation from the sample mean, measured in the same units. The sample standard deviation is given by the square root of the sample variance, which is to say

$$\text{SD} = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

These measures of spread are also useful to supplement the understanding of the behaviour of a dataset that is provided by the measures of central tendency. Specifically, by reporting measures of central tendency, alongside measures of spread, and a description of the shape of the distribution, you are able to describe and summarize the beahviour of a dataset in a concise manner in such a way so as to allow for a deep understanding of the patterns that have emerged.

Example 11.10 (Sadie Questions the Spread in Bill Lengths)**.** After having described the central tendency of the bill length data that Charles day dreamed about, Sadie inquires about the variability in the data. Charles, so focused on the penguins themselves, had not even stopped to consider how much variability may be present in these data. Sadie decides to help out by computing measures of sample variability.

49.0	37.8	45.8	39.0	43.2	48.8	37.8	49.1	40.9	37.3

a. What is the sample range of these data?
b. What is the sample IQR of these data?
c. What is the sample variance of these data? What is the sample standard deviation?
d. What is the variance of the bill lengths? Explain.

Solution

The solution is made easier by first ordering the data. Consider

37.3	37.8	37.8	39.0	40.9	43.2	45.8	48.8	49.0	49.1

a. The largest value is 49.1 and the smallest is 37.3. As a result the sample range is $49.1 - 37.3 = 11.8$.
b. To find the sample IQR we find $Q1$ and $Q3$. There are 10 points, so $Q1$ is the median of the first five, and $Q3$ is the median of the last 5. The first five points are $37.3, 37.8, 37.8, 39.0, 40.9$ so $Q1 = 37.8$. The last 5 points are $43.2, 45.8, 48.8, 49.0, 49.1$ so $Q3 = 48.8$. Thus, the IQR is $48.8 - 37.8 = 11$.
c. For the sample variance, we first note that (from Example 11.8), we know that $\overline{X} = 42.87$. Thus, we get

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{9}\left((37.3 - 42.87)^2 + (37.8 - 42.87)^2 + (37.8 - 42.87)^2 + (39 - 42.87)^2\right. \\ &\quad +(40.9 - 42.87)^2 + (43.2 - 42.87)^2 + (45.8 - 42.87)^2 + (48.8 - 42.87)^2 + \\ &\quad \left.(49 - 42.87)^2 + (49.1 - 42.87)^2\right) \\ &= 24.6156 \end{aligned}$$

For the sample standard deviation, we simply take the square root fo this, giving $s = \sqrt{24.6156} = 4.9615$.
d. We do not know, given this information, what the population variance will be. We do not have access to enough information to compute the parameter, and instead must rely on using only our sample statistics. These are measures of the data that are observed, rather than the full population.

11.3.7 The Five Number Summary and Boxplots

Histograms display a substantial amount of information for the entire observed distribution. They display the shape of the distribution, as well as the details as to which observations are likely or unlikely values, allowing this to be seen in one place. This amount of detail is often very useful, but on occasion it can obscure the larger picture. This becomes particularly apparent when we wish to compare the distribution of two different variables, or perhaps the same variable across two or more categories. In these situations, the numeric summaries that we have discussed end up holding more weight. While it is very often to report the mean along with the standard deviation, as the two values complement eachother well, it is also very common to report the so-called **five number summary** of a data.

Definition 11.15 (Five Number Summary)**.** The five number summary is a method for reporting a set of descriptive statistics for a set of observed data. The five number summary consists of five numbers, listed in order. This is given by

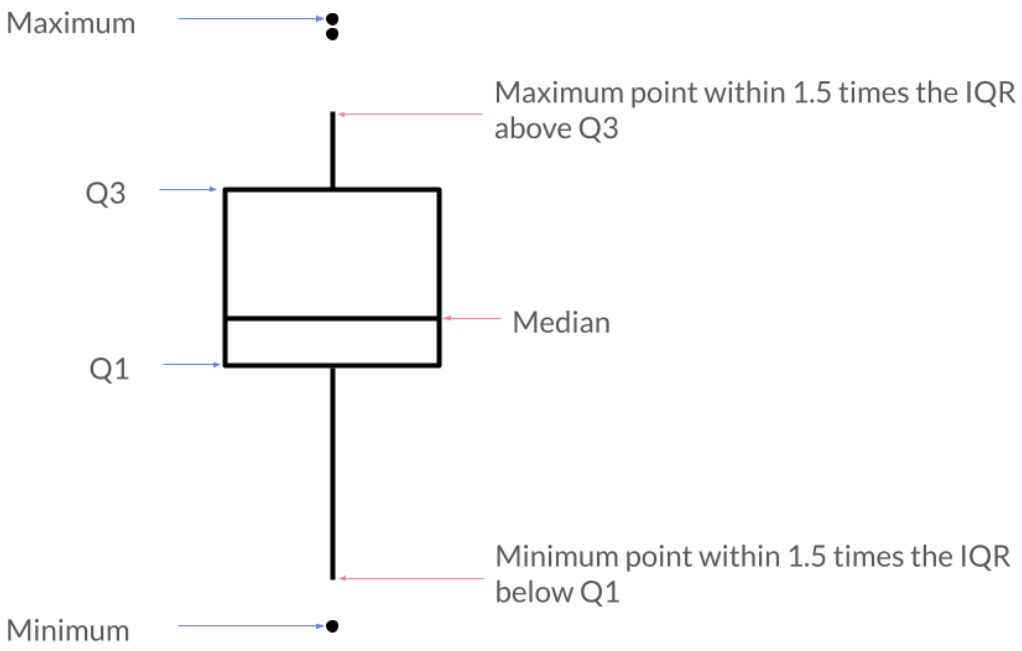
$$\min(x), Q1, \text{Median}(x), Q3, \max(x).$$

That is, the five number summary reports the minimum, the first quartile, the median, the third quartile, and the maximum value from a dataset. Doing so provides a succinct summary of both the location as well as the spread of observed data.

From the five number summary we also immediately know the range, and the IQR. While it is useful to specifically report the values of the five number summary, it can be even more effectively to display this graphically. Boxplots are a graphical display which leverage this idea. For a variable, the boxplot displays the minimum, the maximum, the median, as well as the first and third quartiles $Q1$ and $Q3$, in ascending order, for a given variable. In order to do this, a box is drawn starting at $Q1$ and going up to $Q3$. Then, the median is marked in the middle of this box. Extending from the box are the *whiskers*.

Each whisker is drawn out a length of 1.5 times the observed IQR, stopping at the highest (or lowest) point within that range. Thus, if all points fall within 1.5 times the IQR of either $Q1$ or $Q3$, then the whiskers will stop at the minimum or maximum point observed. If not, there are points beyond those included in the whiskers: these are typically referred to as outliers. The outliers are drawn beyond the whiskers of a boxplot, drawing a single dot for each point.

Figure 11.2: A visual representation of a boxplot. The five number summary is included, along with an indication of which values fall outside of an expected range by using the whiskers to indicate 1.5 times the IQR.



Example 11.11 (Boxplots and Numeric Summaries)**.** Charles and Sadie have gotten very

much into the summarizing data from the penguins. For the first sample of penguins, they have measurements of bill lengths in mm, with the following observations.

49.0	37.8	45.8	39.0	43.2	48.8	37.8	49.1	40.9	37.3

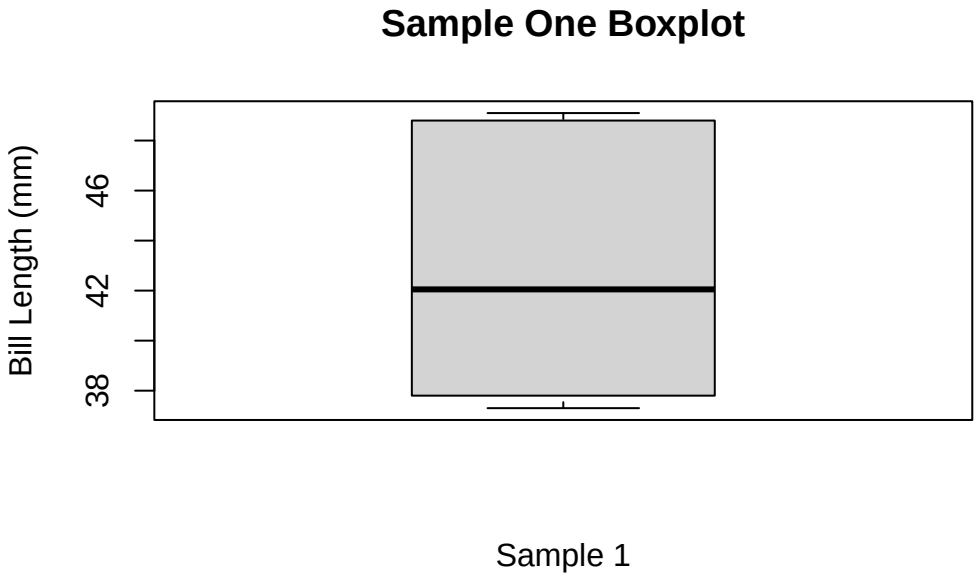
They took another two samples with other bill length measurements, but seemed to have lost the data directly. Fortunately, they have the five number summaries. These are given by the following.

Sample	Min	Q1	Median	Q3	Max
2	32	40	43	45	52
3	34	37	41	49	50

a. Write down the five number summary for the first sample.
b. Sketch a boxplot for the first sample, or explain why it is not possible.
c. Sketch a boxplot for the second sample, or explain why it is not possible.
d. Sketch a boxplot for the third sample, or explain why it is not possible.

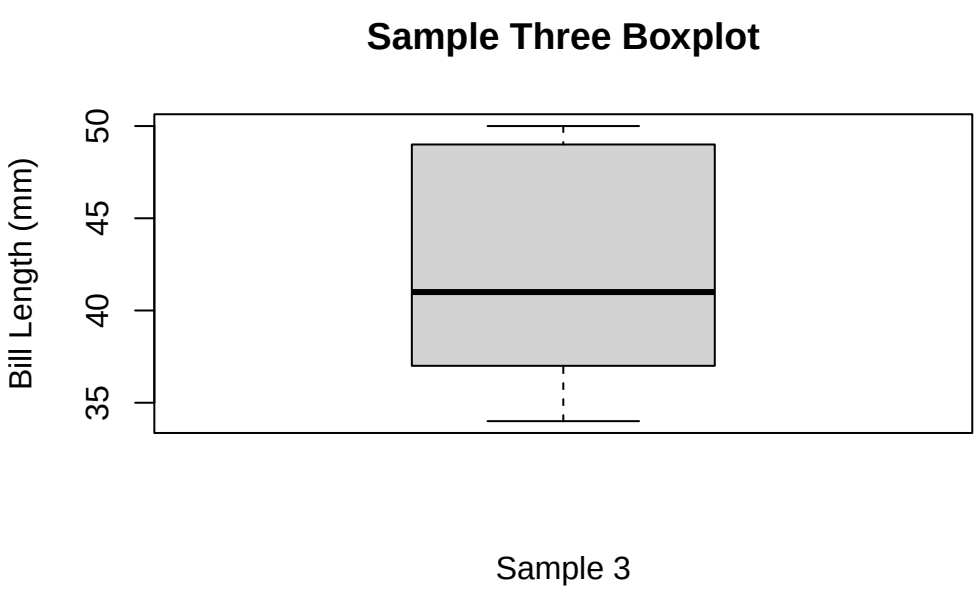
Solution

a. In the previous examples, Example 11.8 and Example 11.10, we found that the five number summary would be given by $37.3, 37.8, 42.05, 48.8, 49.1$.
b. The boxplot here will have a box drawn from 37.8 to 48.8, with a median line drawn at 42.05. The whiskers will extend for 37.3 which is only 0.5 beneath $Q1$ and so there will be no outlier points drawn. The upper whisker will be drawn out to 49.1 which is only 0.3 above $Q3$ and so it will be drawn without outlier points. This can be pictured as follows.



c. In this case we will be unable to draw a boxplot for the sample. The issue is that the minimum point is 32 which is 8 below $Q1$. The IQR is $45 - 40 = 5$, and so $1.5 \times 5 = 7.5$. As a result, the minimum would need to be drawn as an outlier point, which means that we cannot know where the whiskers will stop. This concern is not present on the upper side, where the whisker would be extended to 52.

d. Here we can use the 5 number summary to draw the boxplot directly. The box would be drawn from 37 to 49, with the medina marked at 41. The whiskers would extend down to 34 (3 below $Q1$) and up to 50 (1 above $Q3$). This can be pictured as follows.



It is important to note that the boxplot is inspired by the five number summary, but it encodes slightly more information. It will be precisely the same whenever the maximum and minimum fall within 1.5 times the IQR of the first and third quartiles, but it will include further information in all other cases. This is done to indicate points which are **outliers**, those which appear to deviate from expected trends (of nicely behaved data).[15]

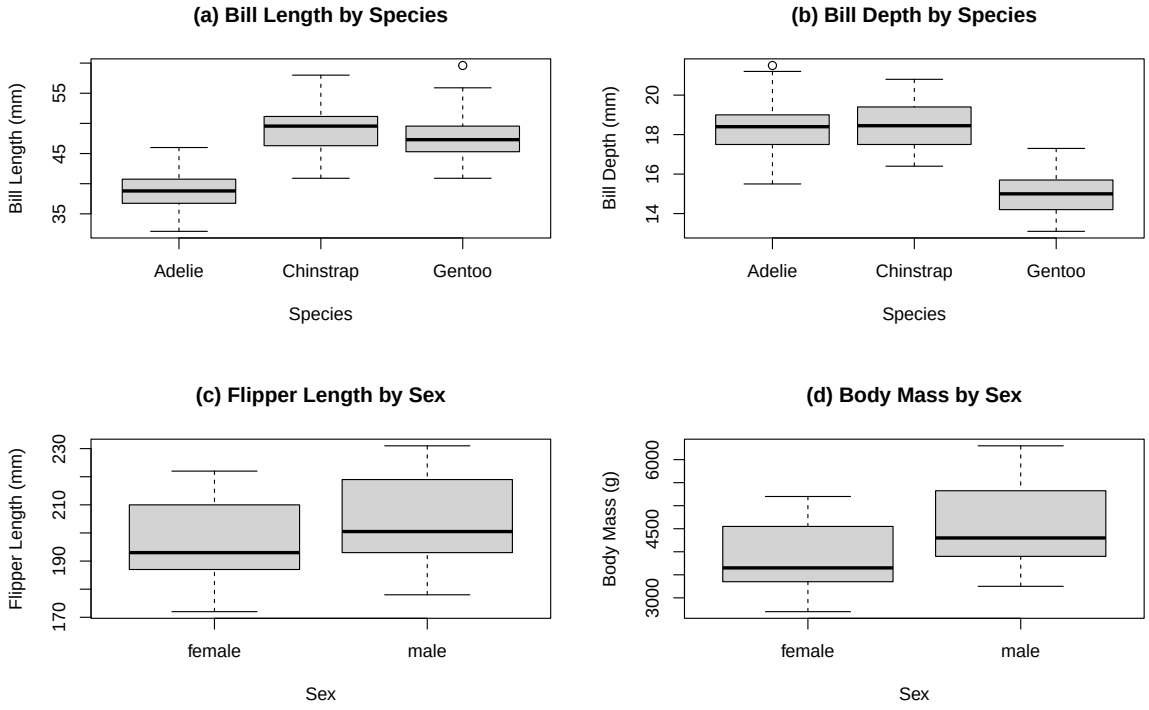
Typically, the boxplots will be drawn so that multiple plots are shown on the same graph. In order to read a boxplot you can compare the medians, and then the spread. The typical variability of the quantity is contained within the box portion, and plots which have largely overlapping boxes are often thought to behave similarly. The whiskers represent the outer limits of what is expected within the data: if both whiskers are roughly the same length, the distribution appears to be mostly symmetric. If one is longer than the other, the distribution exhibits either positive or negative skew. The outliers can contribute to the illustration of skewness on the distribution, but are typically less representative of the distribution itself. A boxplot with a lot of outliers is suggestive of a dataset with far heavier tails than is typical for most well-behaved data, and any analysis on these data should proceed cautiously.

Example 11.12 (Penguin Species Comparisons)**.** The enthusiasm that Charles and Sadie had for the penguin data lead them to reaching out to some researchers who have actually studied

[15]Outliers are a topic that necessitate a great deal of discussion to approach with care. As a general rule, I would be skeptical of any analysis you see which excludes outliers on the basis of a statistical test. Outliers, especially as assessed from these types of rules, are better understood as points that demonstrate that the data are heavy-tailed rather than points which should be ignored.

the penguins. The researchers, grateful to share their work with enthusiastic individuals, sent a series of boxplots comparing multiple different measurements broken up by penguing species and by the sex of the penguins. Charles and Sadie begin to study these various boxplots, comparing the distributions illustrated by each of the boxplots, and trying to determine differences in observations. For each of the following boxplots, compare the location and spread of the various distributions represented, and briefly describe what is observed. The following boxplots are given for:

a. Bill length by species.
b. Bill depth by species.
c. Flipper length by sex.
d. Body mass by sex.



Solution

a. By way of comparison, the adelie pegnuins appear to have less long bills than both the chinstrap and the gentoos, even when accounting for variability. We can see this since the median is substantially lower, and the box does not appear to overlap at all. The chinstrap and gentoo have more comparable bill length,s with the chinstrap being slightly larger in general, but with the gentoo having extreme observations

that are the longest observed.

The adelie have a median of just below 40, will an interquartile range from about 37 to around 41, and all observations falling between about 32 and 46. The chinstrap have a median length around 50, with an interquartile range from about 46 through 51, and a full data span between about 41 and 58. The gentoo have a median of around 47, with a fairly small IQR, spanning from around 46 to around 50. There are no negative outliers, all points falling above about 41, but the highest point extends to nearly 60, sitting beyond the outlier limit of around 56.

The specific five number summary (not easily read directly from the boxplots are):

	Adelie	Chinstrap	Gentoo
Min	32.10	40.90	40.90
Q1	36.75	46.30	45.30
Median	38.80	49.55	47.30
Q3	40.75	51.15	49.55
Max	46.00	58.00	59.60

b. The adelie an chinstrap penguins exhibit roughly the same location, with the chinstrap exhibiting slightly more variability in terms of IQR and slightly less variability in terms of the overall range. The gentoo have less deep bills than either of the other species, with very little overlap between them at all.

The adlie have a median depth of around 18.5, with an IQR ranging from just under 18 to around 19. There is one outlying observation, sitting above 21, with no outliers in the negative direction. The smallest observed bill depth is around 15.5. The chinstrap have a similar median, again around 18.5, with an IQR spanning from just below 18 to just over 19. The range of the data, however, sit from around 16.5 through to approxiamtely 21, with no extreme outlying observations. The gentoo have a median that is a little ways above 15, with an IQR from just above 14 to around 16. The range of the data in total is from around 13 to around 17.

The specific five number summary (not easily read directly from the boxplots are):

	Adelie	Chinstrap	Gentoo
Min	15.5	16.40	13.1
Q1	17.5	17.50	14.2
Median	18.4	18.45	15.0
Q3	19.0	19.40	15.7
Max	21.5	20.80	17.3

c. Male penguins have slightly longer flippers, on average, but with fairly substantial overlap. Both males and females tend to exhibit similar spread, both in terms of the IQR and the range of the data, and the it quite a lot of overlap for both. For

each species, the median sits closer to the first quartile than the third quartile, which suggests that there is likely some skewness in the positive direction, at least throughout the bulk of theo bservations.

To summarize each distribution, you should be able to determine the approximate locations of each of the five numbers from the summary. There are no outliers for any of the measurements. The specific five number summary (not easily read with specificity from the boxplots are):

	Female	Male
Min	172	178.0
Q1	187	193.0
Median	193	200.5
Q3	210	219.0
Max	222	231.0

d. Male penguins weigh more on average than the females. The male penguins also seem to exhibit a wider spread, both in terms of the IQR and the overall range. There is a substantial amount of overlap between the main data, however, less so than for flipper length. Just as with flipper length, the median sits closer to the first quartile than the third, suggesting that there is a slight positive skew in the data.

   To summarize each distribution, you should be able to determine the approximate locations of each of the five numbers from the summary. There are no outliers for any of the measurements. The specific five number summary (not easily read with specificity from the boxplots are):

	Female	Male
Min	2700	3250
Q1	3350	3900
Median	3650	4300
Q3	4550	5325
Max	5200	6300

Exercises

Exercise 11.1. In a survey conducted among students, they were asked about their favorite colors. The options were red, blue, green, yellow, and orange. Below are the responses from 10 students. Use these data to construct a frequency distribution.

Red, Blue, Green, Yellow, Orange, Blue, Red, Green, Green, Yellow

Exercise 11.2. A group of students was surveyed about their preferred leisure activities. The options included reading, playing sports, watching movies, listening to music, and playing video games. Below are the responses from 15 students. Construct the frequency distribution for these data.

> Reading, Playing Sports, Watching Movies, Listening to Music, Playing Video Games, Reading, Playing Sports, Watching Movies, Watching Movies, Listening to Music, Playing Video Games, Reading, Reading, Playing Sports, Playing Sports

Exercise 11.3. In a class of 30 students, each student was asked to choose their favorite genre of music from rock, pop, hip-hop, jazz, and classical. Below are the responses for 10 of them.

> Rock, Pop, Pop, Hip-Hop, Jazz, Classical, Rock, Pop, Pop, Hip-Hop

a. Use these responses to construct a frequency distribution for the data.
b. What proportion of the 30 students preferred rock?

Exercise 11.4. A survey was conducted to find out how many pets each household owns. Below are the responses from 20 households.

> 2, 1, 3, 0, 2, 1, 4, 2, 0, 1, 2, 3, 2, 1, 5, 2, 3, 1, 0, 2

a. What type of variable is this?
b. Write down the frequency distribution for these data. What bin width should you use?
c. Find the five number summary for these data.
d. What is the mode for this variable?

Exercise 11.5. The number of books read by students over the summer break was collected. Below are the responses from 8 students.

> 4, 6, 2, 1, 3, 0, 1, 3

a. Write down the frequency distribution for these data.
b. Find the five number summary for these data.
c. Find the mean and standard deviation for these data.

Exercise 11.6. The time taken (in minutes) by students to complete a quiz was recorded. Below are the times taken by 20 students.

> 20, 25, 30, 35, 40, 22, 28, 33, 37, 42, 24, 29, 31, 36, 39, 21, 26, 32, 38, 41

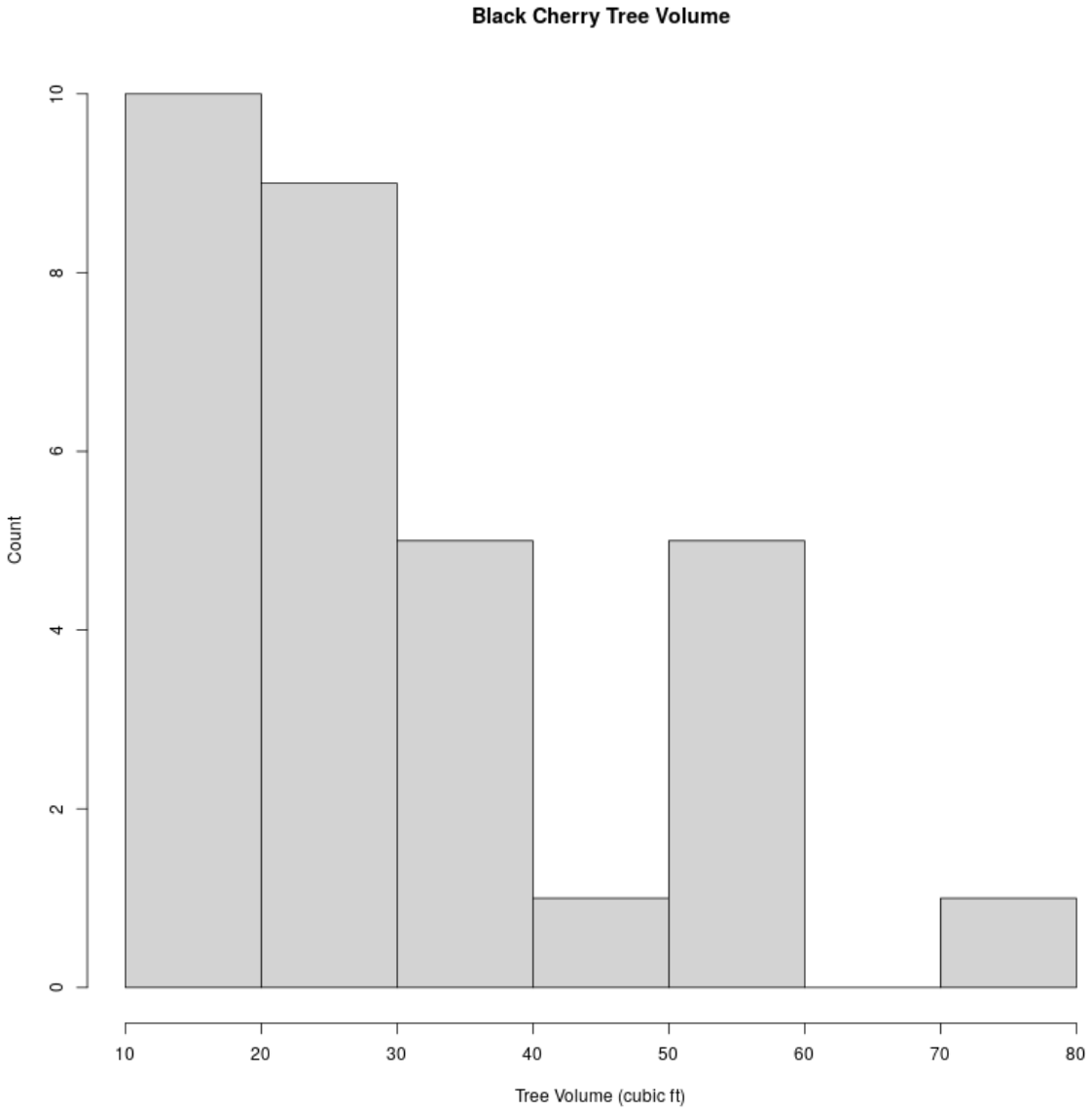
a. What type of variable is this best treated as?
b. Write down a frequency distribution for these data. What bin width did you use?
c. Find the five number summary for these data.
d. What is the mode for this variable?

Exercise 11.7. A survey asked participants about their monthly expenses on groceries, ranging from \$100 to \$1000. Below are the reported expenses from 5 participants.

251, 326, 182, 509, 427

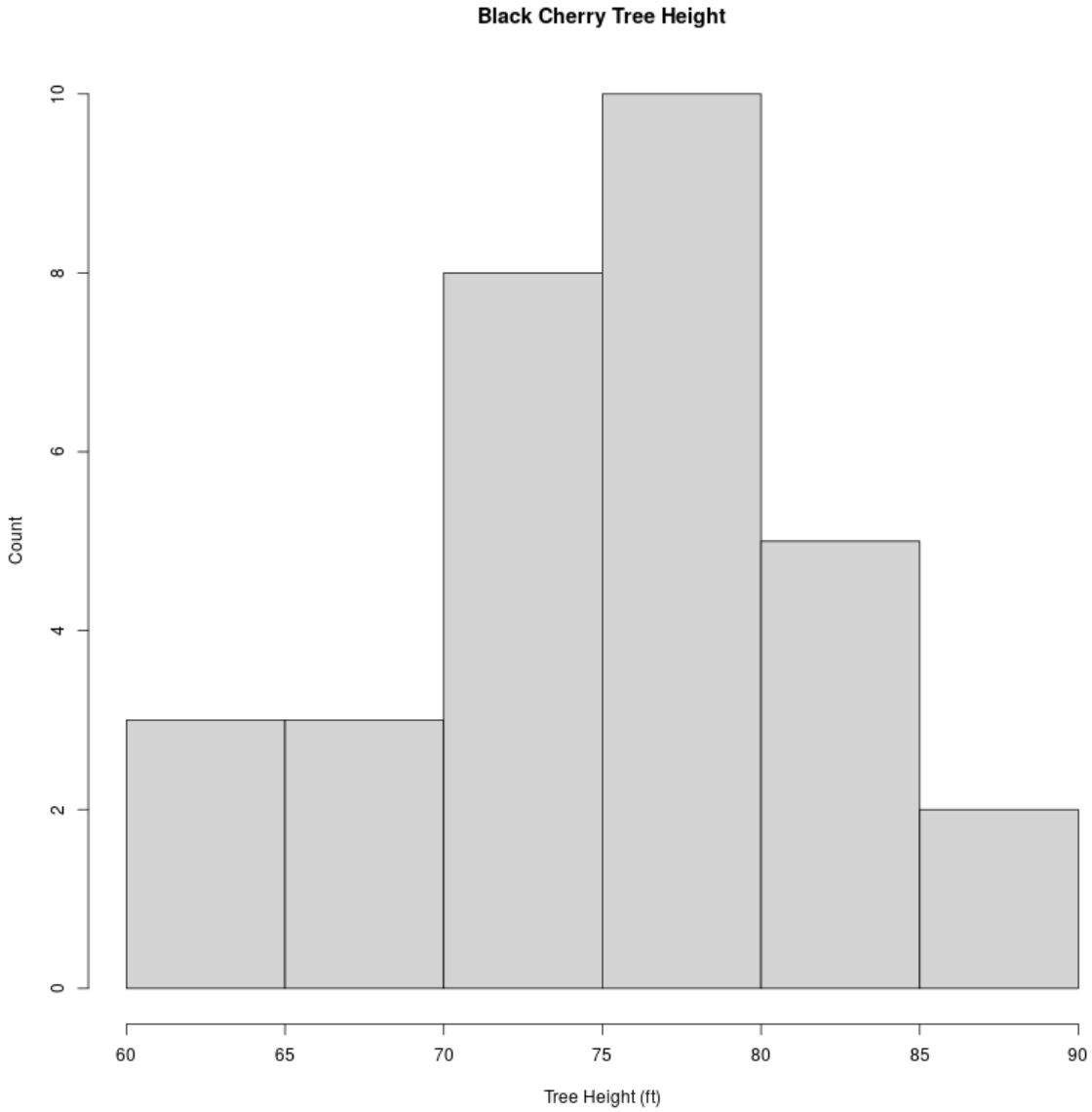
a. Write down the frequency distribution for these data.
b. Find the five number summary for these data.
c. Find the mean and standard deviation for these data.

Exercise 11.8. Consider the following histogram displaying the distribution of volumes for black cherry trees.



a. Describe the distribution of the data set.
b. What is the bin width used?
c. What is the relative frequency of trees between 10 and 20 cubic feet?
d. Are there any notable outliers, (points which otherwise seem to deviate from the overall pattern)? Describe why this may be the case.
e. How would the relative frequency between 10 and 20 change if 4 observations between 60 and 70 were added?

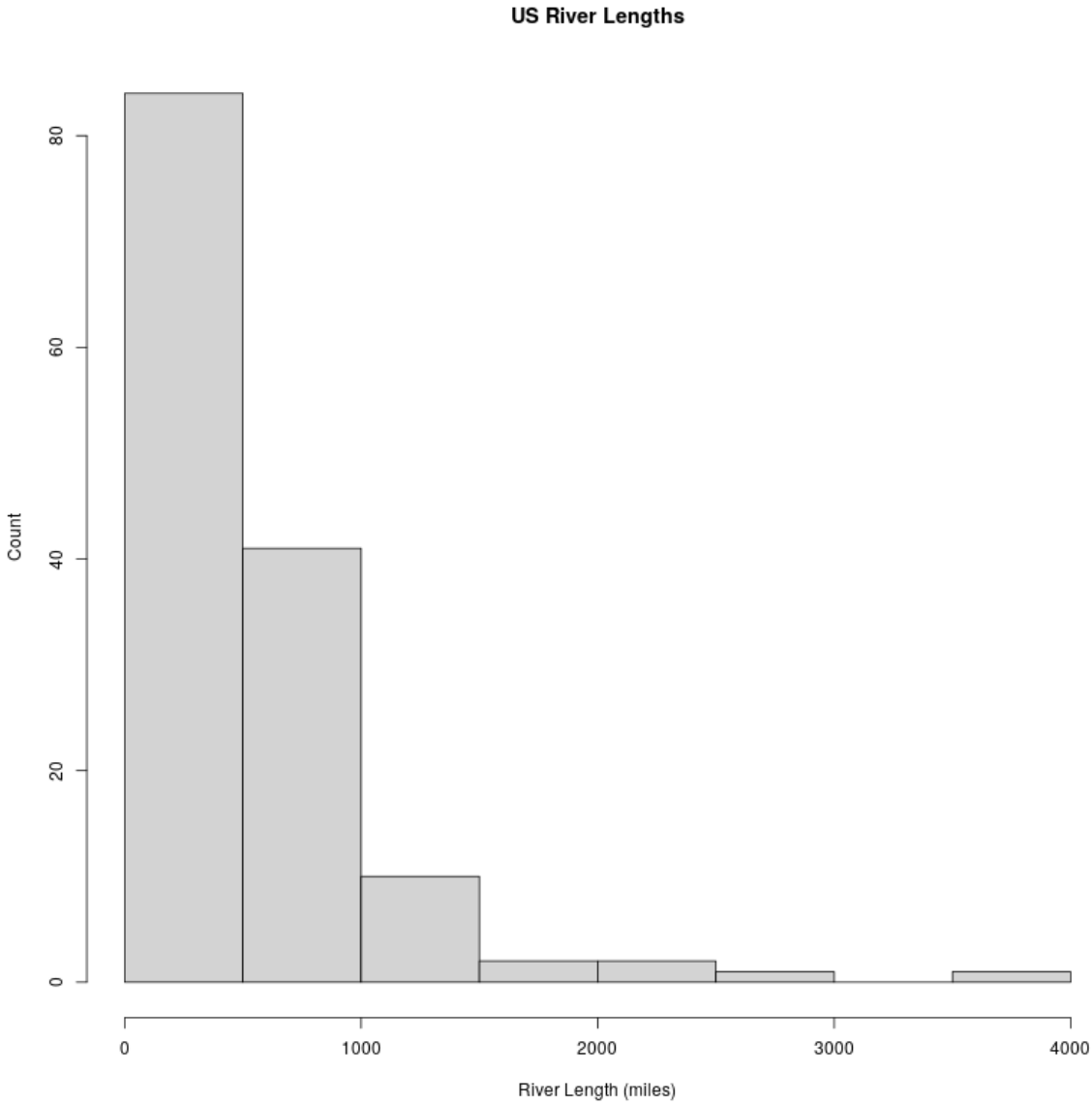
Exercise 11.9. Consider the following histogram displaying the distribution of heights for



black cherry trees.

a. Describe the distribution of the data set.
b. What is the bin width used?
c. What is the relative frequency of trees between 70 and 75 cubic feet?
d. Are there any notable outliers, (points which otherwise seem to deviate from the overall pattern)? Describe why this may be the case.

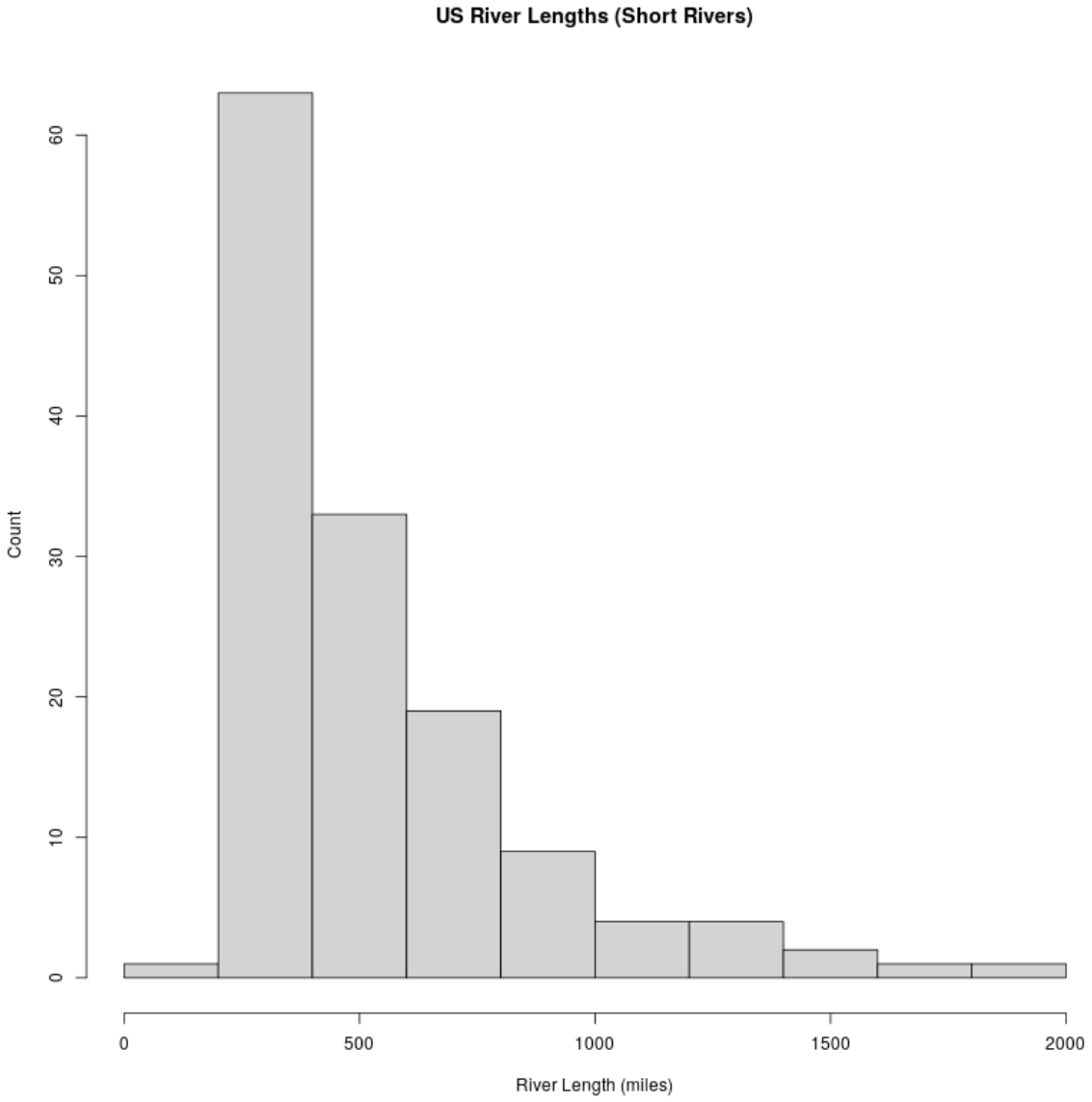
Exercise 11.10. Consider the following histogram displaying the distribution of lengths of ma-



jor rivers in the US.

a. Describe the distribution of the data set.
b. What is the bin width used?
c. Are there any notable outliers, (points which otherwise seem to deviate from the overall pattern)? Describe why this may be the case.
d. When might this be an effective plot? When might this plot be ineffective?
e. Approximately what percentage of observations were longer than 2000?

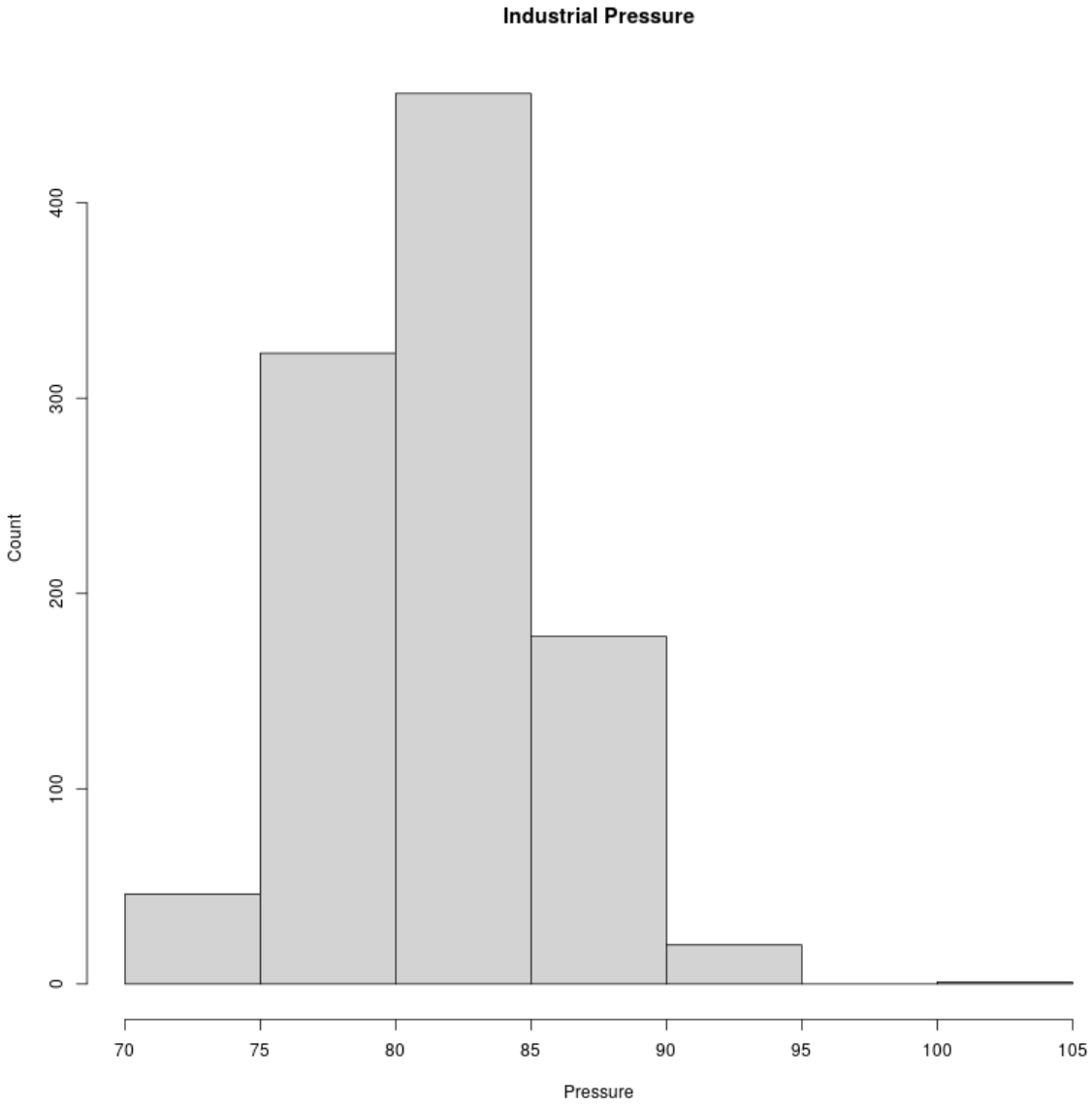
Exercise 11.11. Consider the following histogram displaying the distribution of lengths of major rivers in the US, containing only the data on the rivers under 2000 miles.



a. Describe the distribution of the data set.
b. How many bins are there?
c. Are there any notable outliers, (points which otherwise seem to deviate from the overall pattern)? Describe why this may be the case.
d. Describe how the previous histogram (from Exercise 11.10) would change if the bin width for this graph were used.
e. Approximately what percentage of observations fell between 1000 and 1500?

Exercise 11.12. Consider the following histogram depicting industrial pressure measure-

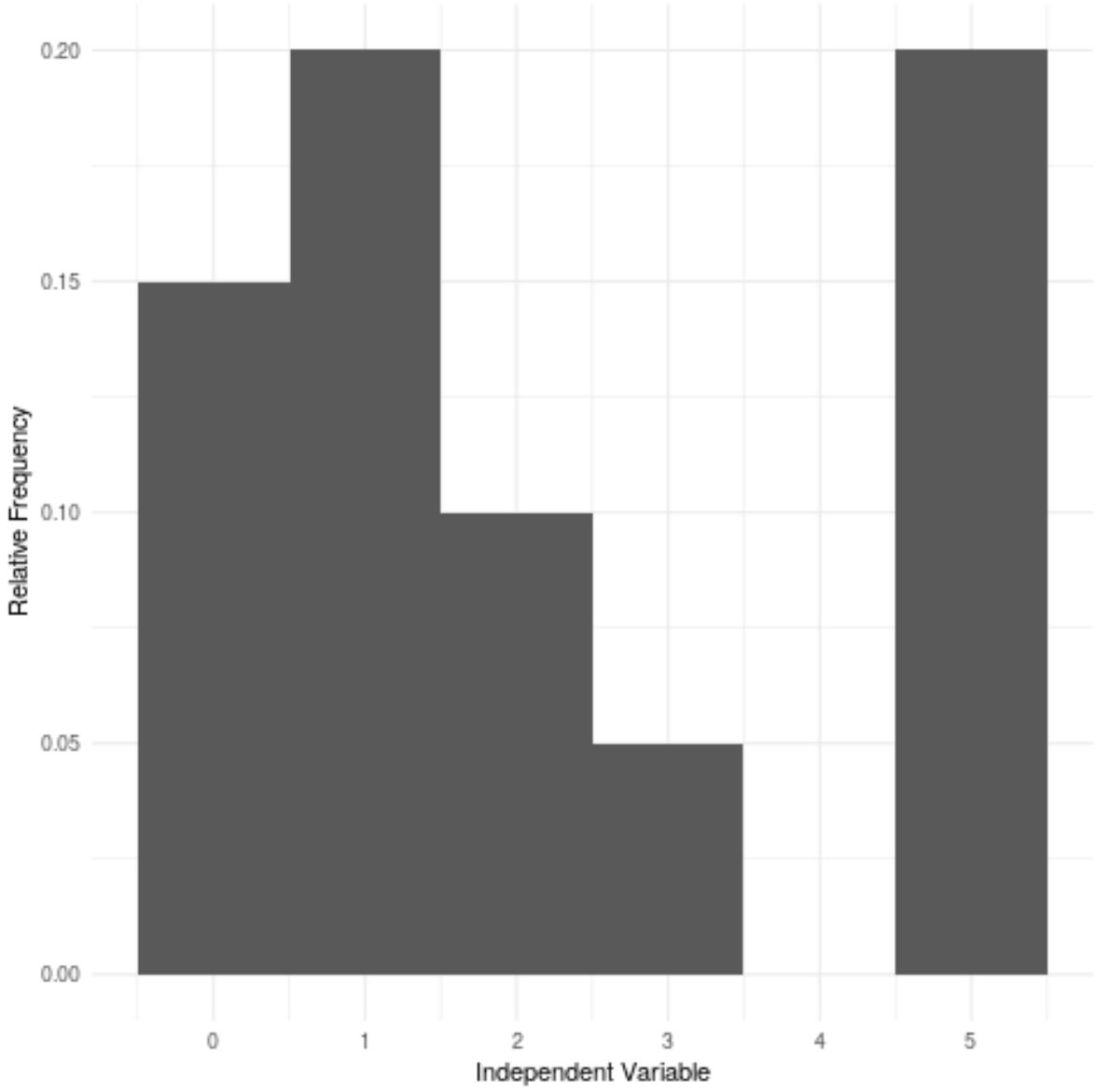
ments.



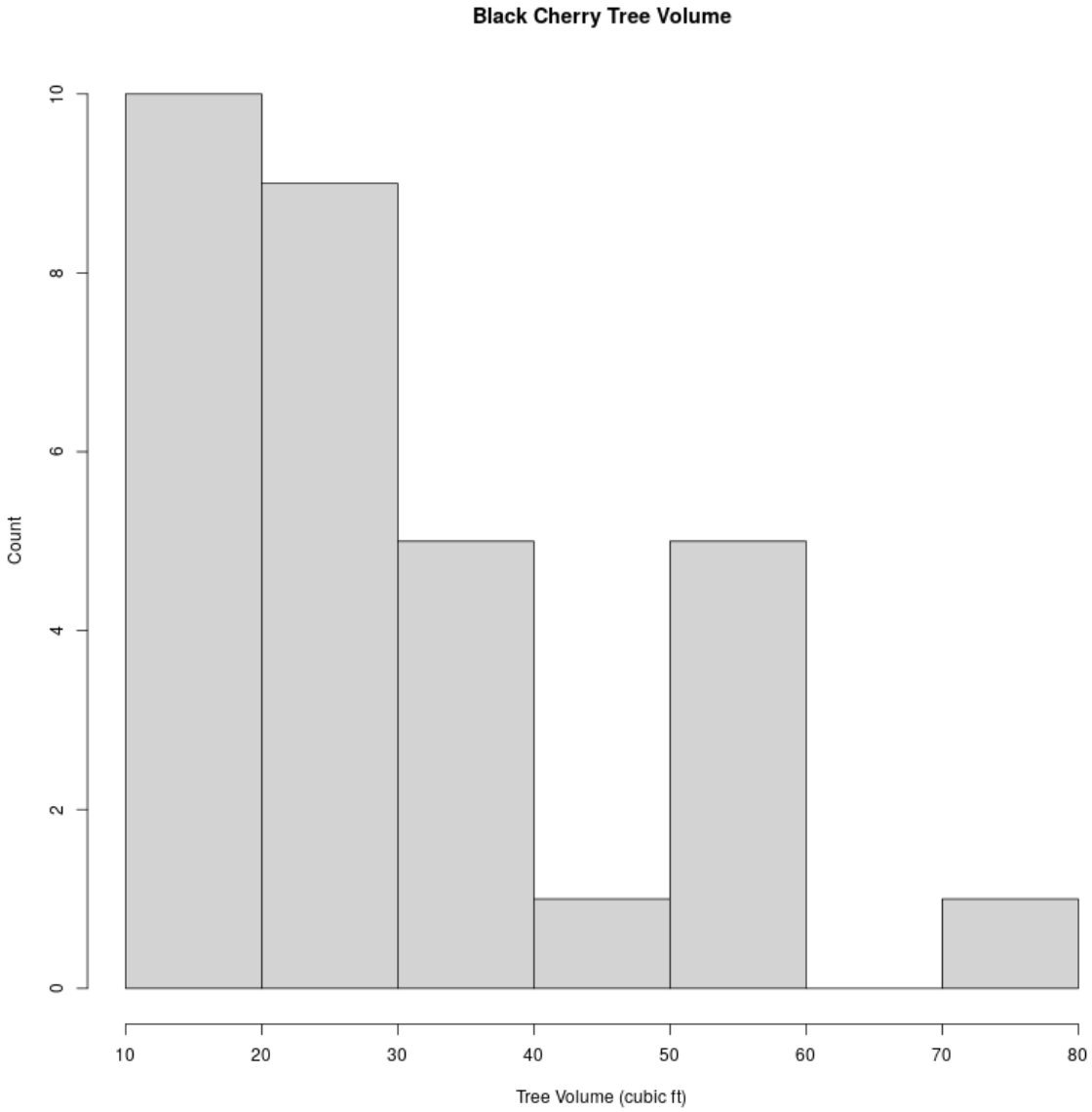
a. Describe the distribution of the data set.
b. What is the bin width used?
c. Are there any notable outliers, (points which otherwise seem to deviate from the overall pattern)? Describe why this may be the case.
d. Suppose that measurements beyond 100 are known to be the result of a transcription error, and they actually should have been recorded between 90 and 95. Describe the distribution that would arise.

e. Approximately what proportion of observations were below 75 or above 90?

Exercise 11.13. Data from the file used to generate the following histogram was accidentally deleted where observations from the independent variable equaled 4. If there were 240 total observations, how many observations were for 4?



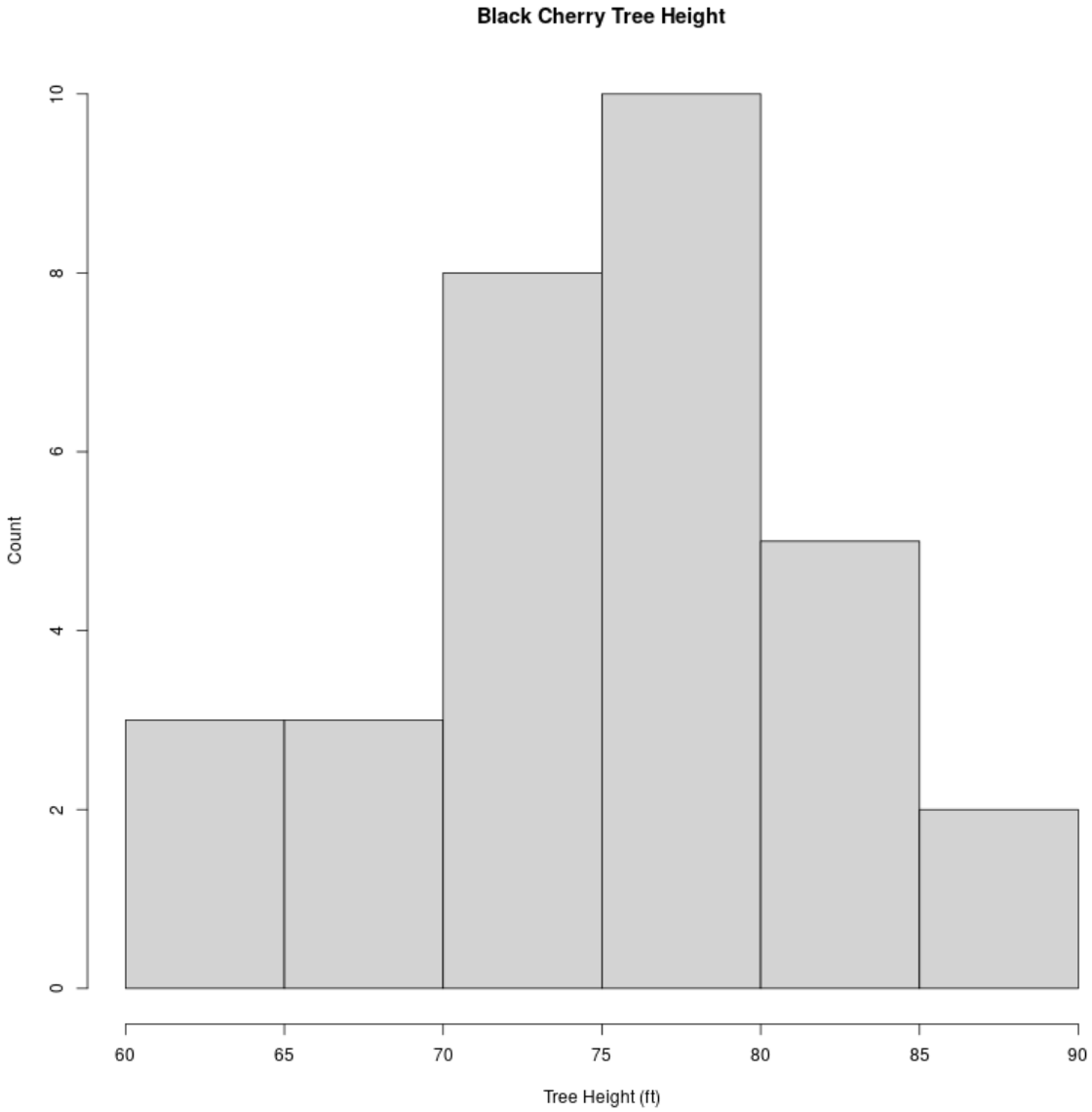
Exercise 11.14. Consider the following histogram displaying the distribution of volumes for black cherry trees.



Using the plot, answer the following:

a. What is the median of the distribution?
b. What is the approximate mean of the distribution?
c. What is the mode of the distribution?
d. Which measure of central tendency is the best for the distribution?
e. Calculate (or approximate) several measures of variation for the distribution.

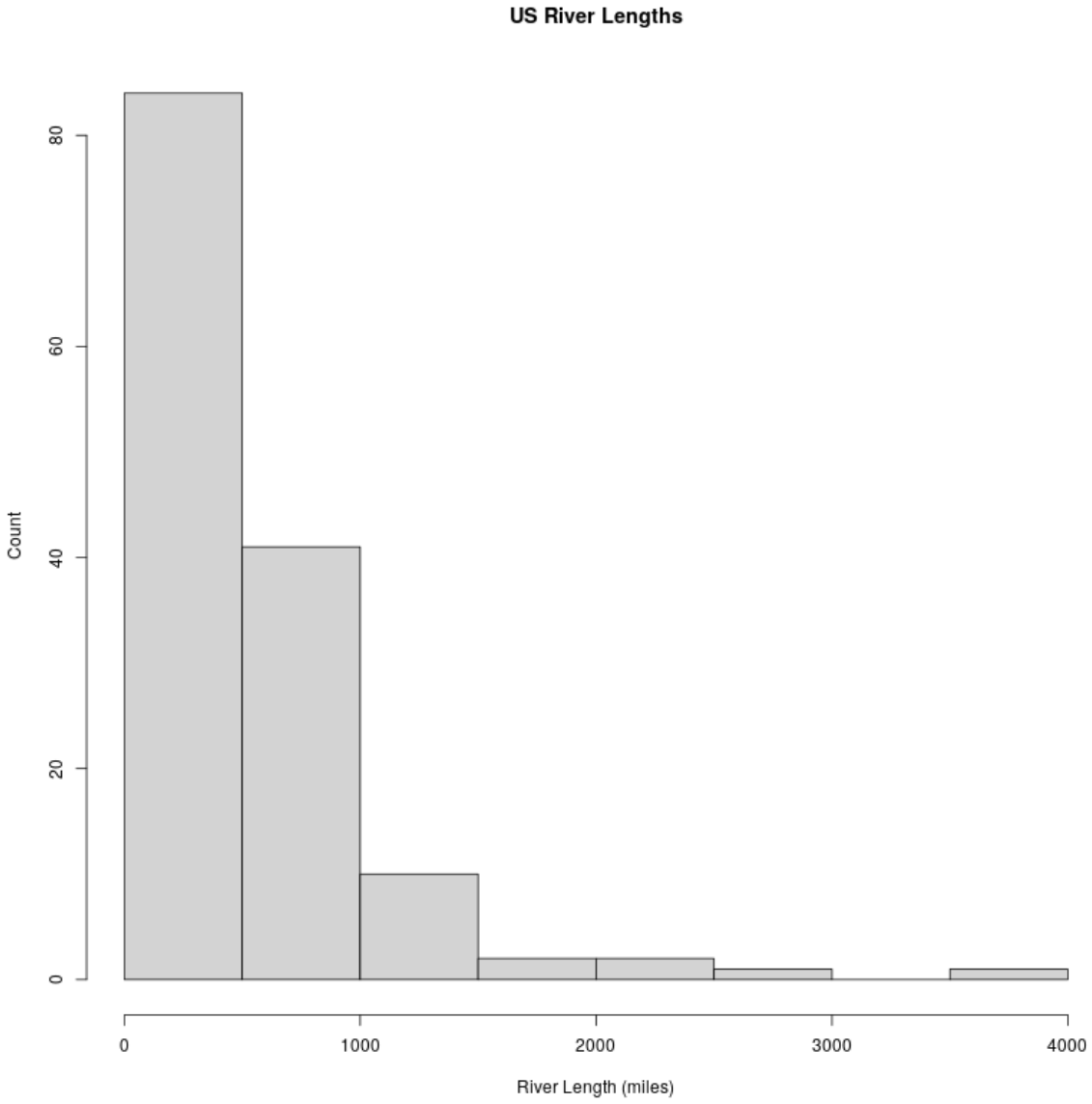
Exercise 11.15. Consider the following histogram displaying the distribution of heights for black cherry trees.



Using the plot, answer the following:

a. What is the median of the distribution?
b. What is the approximate mean of the distribution?
c. What is the mode of the distribution?
d. Which measure of central tendency is the best for the distribution?
e. Calculate (or approximate) several measures of variation for the distribution.

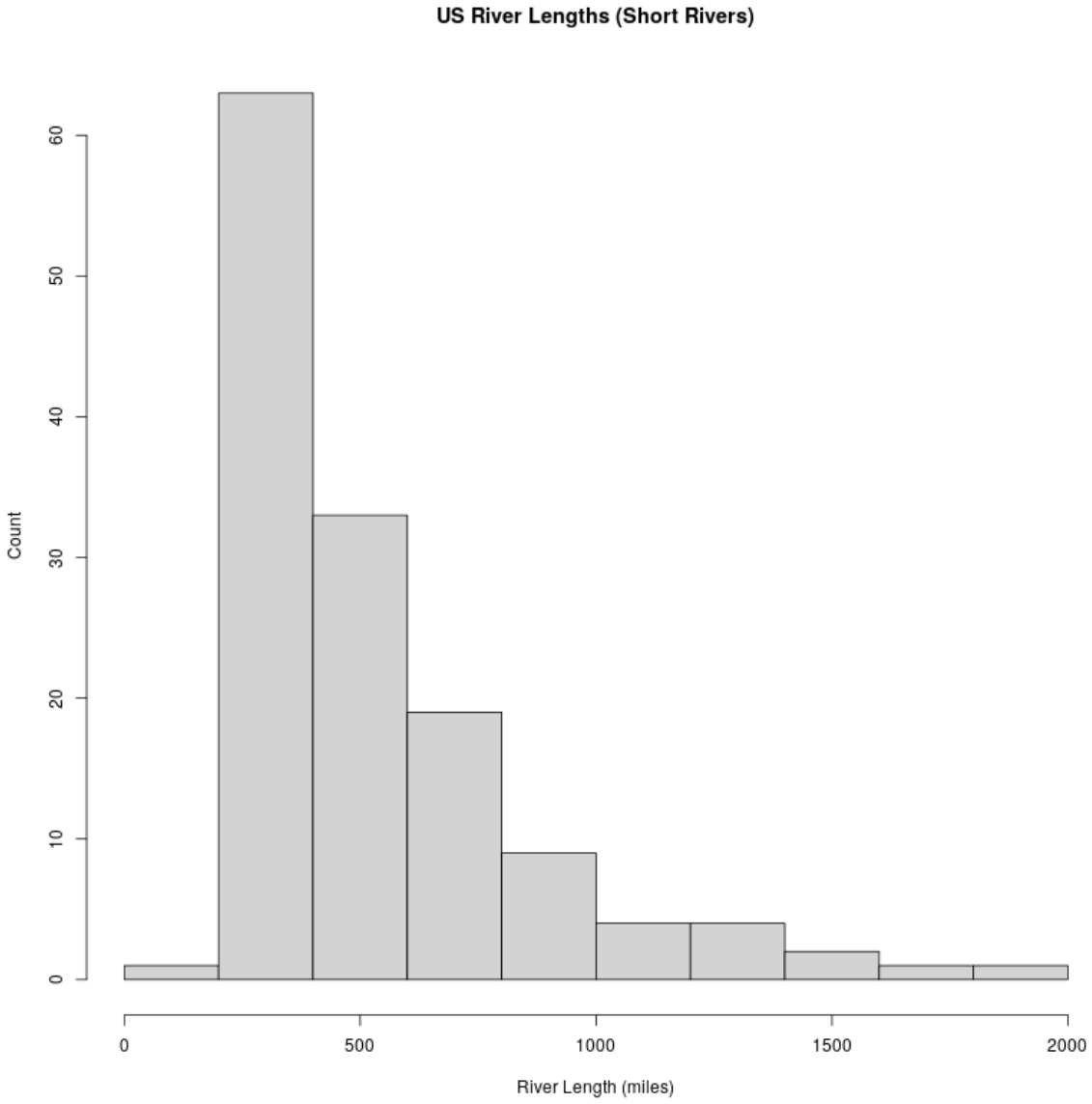
Exercise 11.16. Consider the following histogram displaying the distribution of lengths of major rivers in the US.



Using the plot, answer the following:

a. What is the median of the distribution?
b. What is the approximate mean of the distribution?
c. What is the mode of the distribution?
d. Which measure of central tendency is the best for the distribution?
e. Calculate (or approximate) several measures of variation for the distribution.

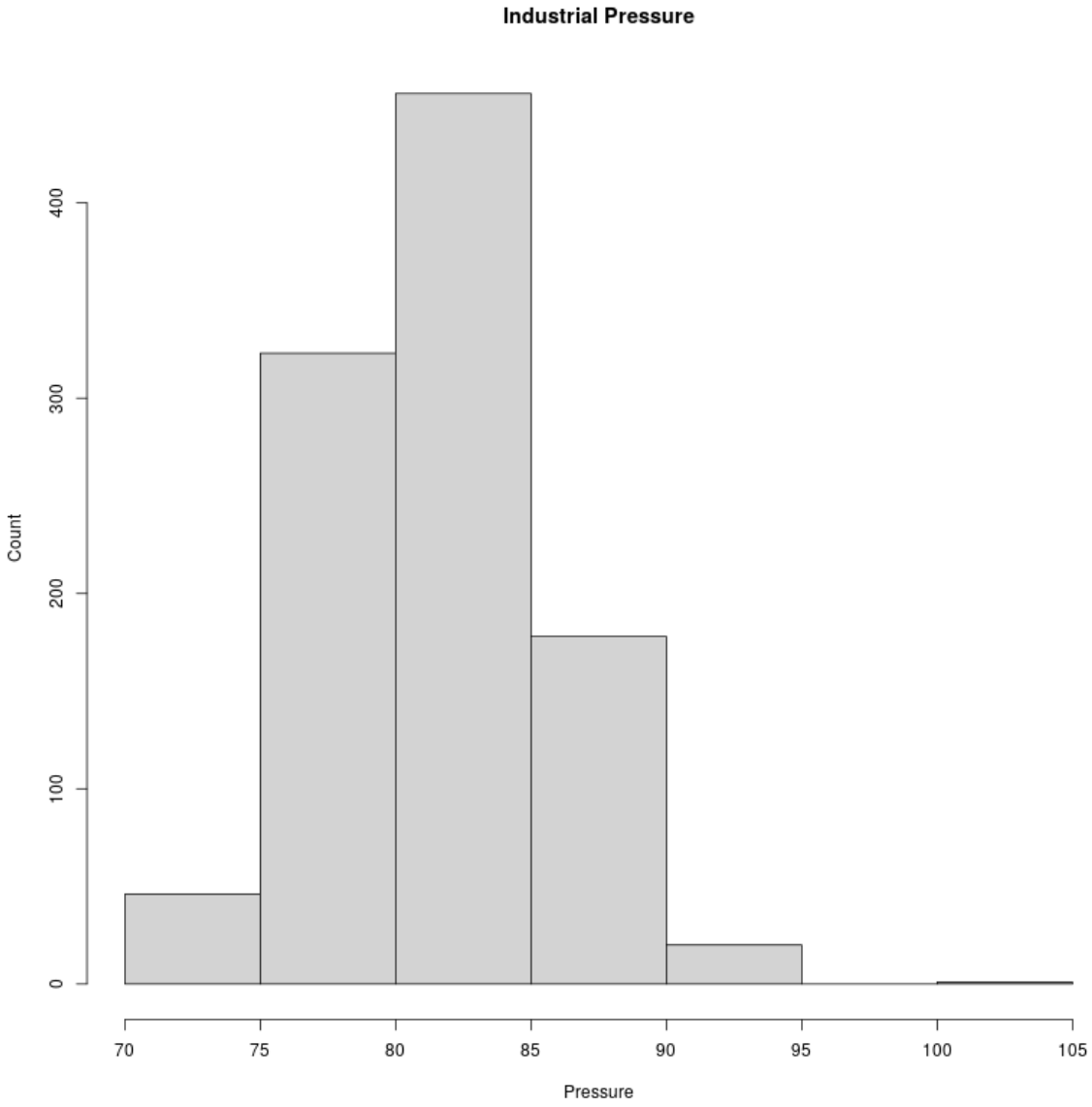
Exercise 11.17. Consider the following histogram displaying the distribution of lengths of major rivers in the US, containing only the data on the rivers under 2000 miles.



Using the plot, answer the following:

a. What is the median of the distribution?
b. What is the approximate mean of the distribution?
c. What is the mode of the distribution?
d. Which measure of central tendency is the best for the distribution?
e. Will the measure of variation be larger or smaller when you restrict to the shorter rivers?

Exercise 11.18. Consider the following histogram depicting industrial pressure measurements.



Using the plot, answer the following:

a. What is the median of the distribution?
b. What is the approximate mean of the distribution?
c. What is the mode of the distribution?
d. Which measure of central tendency is the best for the distribution?
e. Calculate (or approximate) several measures of variation for the distribution.
f. Suppose that measurements beyond 100 are known to be the result of a transcription error, and they actually should have been recorded between 90 and 95. What measure

of central tendency is best for the distribution?

Exercise 11.19. Consider a quantitative variable measured in a dataset.

a. Is the sample mean always equal to one of the values in the sample? If. so, explain why. If not, give an example.
b. Is the sample median always equal to one of the values in the sample? If. so, explain why. If not, give an example.
c. Is the sample mode always equal to one of the values in the sample? If. so, explain why. If not, give an example.

Exercise 11.20. In one group of 20 individuals, the mean height was 178cm. In a second group of 30 individuals the mean height was 164cm. What is the mean height for both groups, when they are put together?

Exercise 11.21. There are 10 employees in a particular division of a company. Their salaries have a mean of $70,000$, a median of $55,000$, and a standard deviation of $20,000$. The largest number on the list is $100,000$.

A clerical error is made which enters the maximum value as a $1,000,000$ rather than $100,000$.

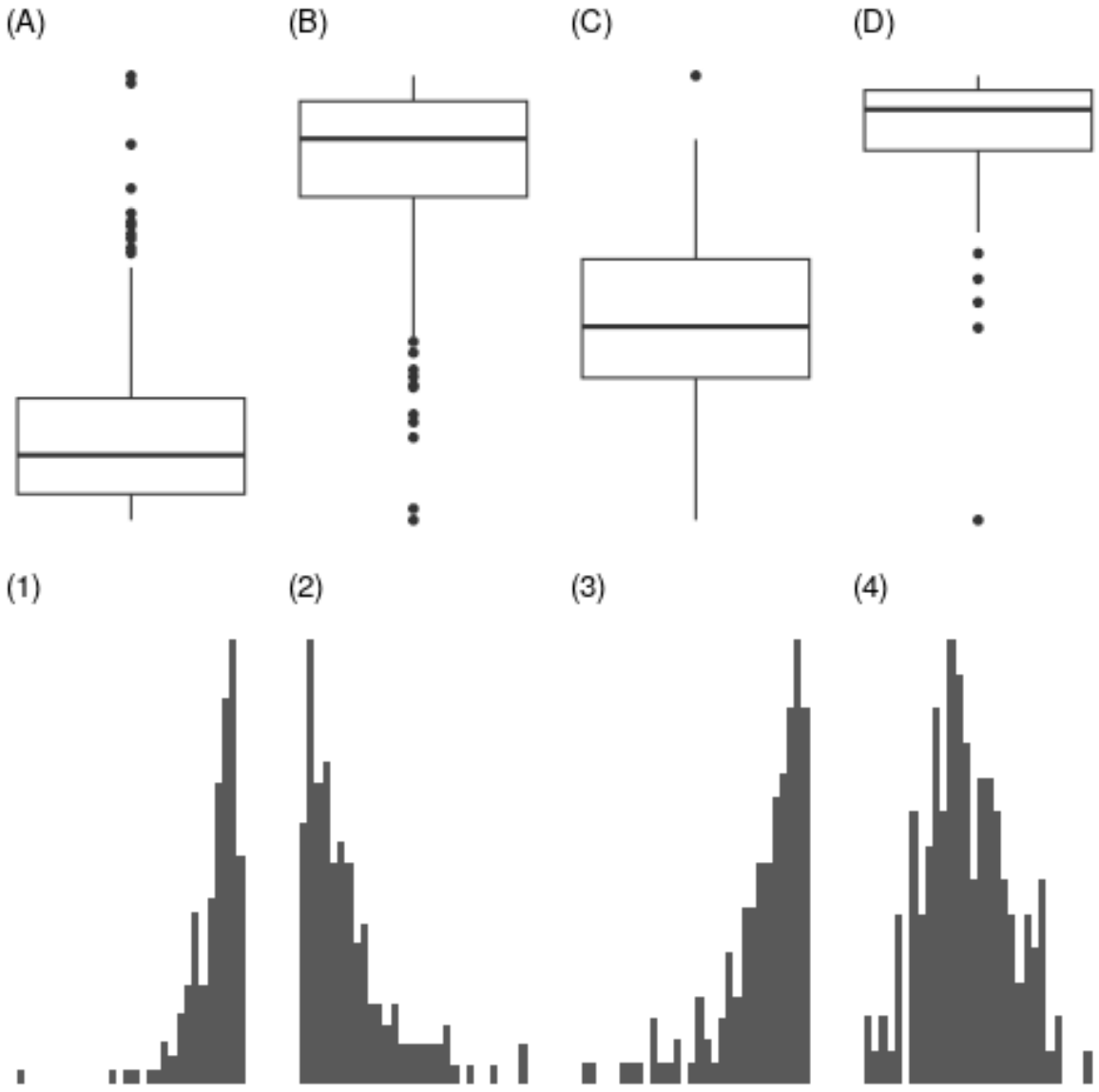
a. What is the mean of the altered data?
b. What is the median of the altered data?
c. What is the standard deviation of the altered data?

Exercise 11.22. Four research teams caught groups of ganges dolphins, consisting of 15, 20, 10, and 18 dolphins each. The reported mean weights (in pounds) of the groups were 162, 148, 153 and 140, respectively. What was the overall mean of all the dolphins?

Exercise 11.23. A certain company is considering giving employees a weekly raise. Currently, the average weekly salary at the company is 1000, with a standard deviation of 100.

a. What would happen to the mean and standard deviation if a \$50 raise were given to every employee?
b. What would happen to the mean and standard deviation if a 5% raise were given to every employee?

Exercise 11.24. Match each of the following boxplots (A)-(D), with the histograms (1)-(4).



Bartoš, František, Alexandra Sarafoglou, Henrik R. Godmann, Amir Sahrani, David Klein Leunk, Pierre Y. Gui, David Voss, et al. 2023. "Fair Coins Tend to Land on the Same Side They Started: Evidence from 350,757 Flips." https://arxiv.org/abs/2310.04153.

Gigerenzer, Gerd, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V Katsikopoulos. 2005. "'A 30% Chance of Rain Tomorrow': How Does the Public Understand Probabilistic Weather Forecasts?" *Risk Analysis: An International Journal* 25 (3): 623–29.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–54.