An Intuitive, Interactive, Introduction to Biostatistics

Ward, Caitlin E; Nolte, Collin

Ward, C. E., & Nolte, C. (2023). An Intuitive, Interactive, Introduction to Biostatistics. University of Iowa. https://doi.org/10.25820/work.006232

An Intuitive, Interactive, Introduction to Biostatistics

Caitlin Ward and Collin Nolte

Introduction

Welcome! This resource offers an interactive learning experience for introductory biostatistics. This resource was designed with the course BIOS:4120 Introduction to Biostatistics at the University of Iowa in mind, but could be used in any introductory statistics or biostatistics course. We combine textual explanations and embedded interactive Shiny applications, to provide students with an engaging learning resource that offers an intuitive illustration of core statistical concepts. The applications and associated exercises are targeted at *conceptional* understanding, as opposed to calculations.

At this point in time, we provide chapters on:

- Introduction to biostatistics
- Data and data summaries
- Study design and bias
- Randomness, probability, and simulation
- Probability distributions
- Sampling distributions and Central Limit Theorem
- Confidence intervals and hypothesis testing

Acknowledgements

We would like to acknowledge the funding which made the creation of this resource possible, provided by a University of Iowa Libraries OpenHawks grant.

We would also like to thank Patrick Breheny, Knute Carter, and Jacob Oleson, whose course notes helped inform some of the material covered in this textbook.

1 Introduction

> "So never lose an opportunity of urging a practical beginning, however small, for it is wonderful how often in such matters the mustard-seed germinates and roots itself."
>
> — Florence Nightingale

Learning objectives

1. Describe the field of statistics and its importance in scientific inquiry
2. Learn the scientific method, its purpose, and its connection to statistics
3. Understand the concepts in the general statistical framework

1.1 Statistics and Evidence-based Research

There are many definitions of statistics:

- the science of learning from experiences
- the study of collecting, analyzing, and interpreting data
- the mathematics of uncertainty
- a required college course designed to make everyone miserable

All of these definitions hold a degree truth (some more than others – we'll let you guess which 😉). Statistics is a highly versatile field, with methods applicable to and used by a wide variety of disciplines. The importance of the field of statistics lies in it's ability to quantify our uncertainty about a potential outcome – such as the most likely cause and treatment for a headache. Almost everything in life is associated with some uncertainty – the weather, your career, the connection between your genetics and your response to certain medications; however, just because something isn't known with 100% certainty doesn't mean we can't understand it better. Statistics helps us understand the world better by allowing us to quantify our uncertainty, describe associations between phenomena, and make informed decisions.

And how exactly do researchers use statistics to quantify their uncertainty? Well, this is achieved by collecting data, or evidence, about whatever process is of interest. For example, if you want to know how likely it is to rain tomorrow, you might look at the weather in your location the past few days. Since rain can travel, you might also look at the recent weather in the surrounding areas. You could also evaluate historical weather records to determine if this season is known for being particularly rainy or dry in the past. These are all various examples of data you can collect to inform your weather prediction. In scientific research, data is obtained and investigated through statistical inquiry to help answer important questions:

- Which drug should a doctor prescribe to treat an illness?
- Can we predict the longevity of the national population to inform government decisions regarding Social Security?
- What factors increase the risk of an individual developing coronary heart disease?

These questions are too important to be left to opinion, superstition, or conjecture. Consequently, there has been a tremendous push for objective, **evidence-based** decision making in medicine, public health, and policy making. Statistics and biostatistics are the sciences that allows us to make these difficult decisions. When studying humans, we can't control every aspect of their life, such as what they eat or where they work. There are ethical considerations - if we wanted to research the effects of smoking on lung cancer, we would not be able to force people to smoke (or not). Humans are also incredibly diverse and variable, not to mention expensive to perform research upon. Despite these issues, there is a moral imperative to make decisions on potentially life-saving therapies as fast as possible. To make evidence-based conclusions, we need to collect, process, analyze, and interpret data in order to draw conclusions in an objective manner.

Definition 1.1

Evidence-based practice: *Using sound research findings based on observed or collected data to make decisions*

1.2 Scientific Method

In principle, people collect and process information every day of their lives. Since it's something we do frequently, you might think we would be really good at it…but we're not. Unfortunately, humans are not natural statisticians. We are not good at picking out patterns from a sea of noisy data. And, on the flip side, we are *too good* at picking out non-existent patterns from small numbers of observations. We also find it difficult to sort out the effects of multiple factors occurring simultaneously and we are subject to all sorts of biases depending on our personalities, emotions, and past experiences.

In order to mitigate any of our own personal biases when answering important questions about the way the world works (i.e., to do good science), we must be careful to be rigorous in the way we proceed. The **scientific method** is the process used to answer scientific questions.

1. Ask a question
2. Construct a hypothesis
3. Test your hypothesis with a study or an experiment
4. Analyze data and draw conclusions
5. Communicate results

As an example of the scientific method, consider that you would like to start a garden, but you are not sure which type of fertilizer is best. If there are three types of fertilizer you cannot decided between, you might proceed through the scientific method as follows:

1. Ask a question

Of fertilizers A, B, and C, which will yield the fastest plant growth in my garden?

2. Construct a hypothesis

You may have read some reviews for each type of fertilizer online. Fertilizer A has 4.5 stars, fertilizer B as 4.2 stars, and fertilizer C has 4.6 stars. Thus, you may hypothesize that fertilizer C leads to the fastest plant growth, followed by fertilizer A, and that fertilizer B has the slowest plant growth.

3. Test your hypothesis with a study or an experiment

To test your hypothesis, you can section off three areas of your garden. In each area, you use one of the three fertilizers and plant the same number and type of plants. In order for your experiment to be fair, you will want to make sure each of the three areas is equal-size, gets the same amount of sunlight, is watered the same amount, etc.

4. Analyze data and draw conclusions

After 5 weeks, you measure the height of each plant which use each type of fertilizer. Taking the average plant height from each of the three sections of your garden, you find that when fertilizer A was used, the average plant height was the tallest and thus you conclude that fertilizer A is the best for your garden.

5. Communicate results

Other gardeners in your area may be interested in your results. You may share with them your recommendation for fertilizer A, which is backed by a well-controlled experiment.

Even in this relatively simple example, you can see how much care is required to have a valid experiment. There are many other factors that may also influence your decision on the best fertilizer, such as the price and availability. It's also possible that fertilizer A ended up with the fastest average plant growth just due to random chance, and if you repeated the entire experiment again you might find fertilizer B or C to result in taller plants.

Definition 1.2

Scientific method: *Five steps that can be used to acquire knowledge without incorporating personal opinion*

1.3 Statistical framework

As we saw in the fertilizer example, collecting and analyzing data can be complicated. Statistics helps us design studies, test hypotheses, and use data to make scientifically valid conclusions about the world. In general, scientists use the scientific method to make generalizations about classes of people on the basis of their studies. The class of people that they are trying to make generalizations about is called the **population**. Most of the time, it is impractical and expensive to study all individuals in a population - although many governments do the best they can every 10

years. Instead of sampling everyone in the population, or taking a **census**, typically we study only a portion of the population called the **sample**. In order to determine how best to obtain a sample to answer the research questions, we must be cautious about the **study design**. Then, researchers make generalizations, or **inference**, about the entire population based on studying the sample. We can visualize the statistical framework using this diagram:

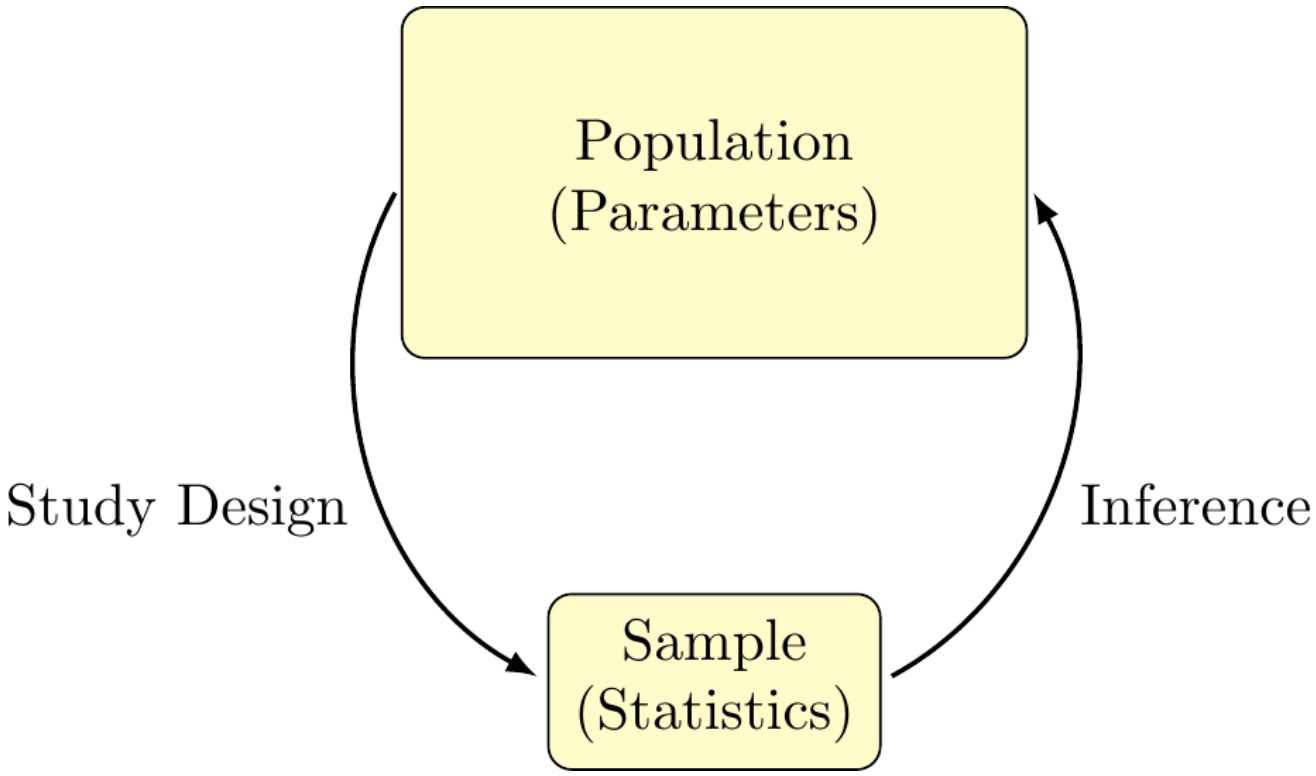


Figure 1.1: Statistical Framework

Throughout this book, we will dive deeper in to each aspect of the statistical framework. We consider two broad categories of statistical analysis: **descriptive statistics**, which deals with methods for summarizing and/or presenting data and **inferential statistics**.

Definition 1.3

Population: *The group towards which scientists attempt to generalize*

Census: *A study which includes all members of the population*

Sample: *A portion of the population which is part of the study*

Study design: *The process of obtaining the best sample to answer the research questions*

Inference: *The process of making generalizations about the population*

Descriptive statistics: *Methods for summarizing and/or presenting data*

Inferential statistics: *Methods for making generalizations about a population using information contained in the sample*

2 Data Summaries and Presentation

> “Numerical quantities focus on expected values, graphical summaries on unexpected values.”
>
> — John Tukey

Learning objectives

1. Define data and categorize and identify different types of data
2. Understand and calculate numerical summaries of different data types
3. Learn about different types of graphs and how they can be interpreted

2.1 Introduction to Data

Data can be virtually anything that is observed and recorded, including how tall you are, the color of all the cars in a town, or the time it takes to drive to work. Just as the different types of data can vary considerably, so can the amount. There is a natural tension between the quantity of data available and our abilities to make sense of it. It is difficult to sort through large streams of data and make any meaningful conclusions. Instead, we can better understand data by condensing it into human readable mediums through the use of *data summaries*, often displayed in the forms of *tables* and *figures*. However, in doing so, information is often be lost in the process. A good data summary will seek to strike a balance between clarity and completeness of information. The focus of this chapter will be on descriptive statistics, utilizing both numerical and graphical summaries of various types of data.

The optimal summary and presentation of data depends on the data’s type. There are two broad types of data that we may see in the wild, which we will call **categorical data** and **continuous data**. As the name suggests, categorical data (sometimes called *qualitative* or *discrete* data) are data that fall into distinct categories. Categorical data can further be classified into two distinct types:

- **Nominal data**: data that exists without any sort of natural or apparent ordering, e.g., colors (red, green, blue), gender (male, female), and type of motor vehicle (car, truck, SUV).
- **Ordinal data**: data that does have a natural ordering, e.g., education (high school, some college, college) and injury severity (low, medium, high)

Continuous data (sometimes called *quantitative* data), on the other hand, are data that can take on any numeric value on some interval or on a continuum. Examples of continuous data include height, weight, and temperature. Categorical and continuous data are summarized differently, and we'll explore a number of ways to summarize both types of data.

Definition 2.1

Categorical data: *Data that takes on a distinct value (i.e., falls into categories)*

Nominal data: *A type of categorical data where the categories do not have any apparent ordering*

Ordinal data: *A type of categorical data where there is a natural ordering*

Continuous data: *Data that takes on numeric values, often measured on an interval*

2.2 Categorical Data

2.2.1 Basic Categorical Summaries

Let's begin by considering a dataset of survey responses for 592 students responding with their sex, hair color, and eye color. This data includes responses from male and female students, with hair colors that are black, brown, red, or blond, and eyes that are brown, blue, hazel, or green. Note that these are *qualitative* measures, suggesting that are dealing with categorical data. Let's take a look at the data for the first ten subjects.

SubjectID	Sex	Hair	Eye
1	Male	Brown	Blue
2	Female	Blond	Blue
3	Female	Blond	Blue
4	Female	Black	Brown
5	Male	Red	Green
6	Male	Blond	Blue
7	Female	Black	Brown
8	Male	Blond	Green
9	Female	Blond	Blue
10	Female	Brown	Brown

Each row indicates a subject possessing the indicated sex, hair, and eye color. For example, the first row indicates a male with brown hair and blue eyes. Trying to make sense of 592 such observations is a daunting task, so we can begin by taking the data we have and summarizing it in a useful way. For categorical data, like we have here, summarizing the data is pretty straightforward – you just count how many times each category occurs. For example, we can count how many of each hair color was observed in our data.

Black	Brown	Red	Blond	Total
108	286	71	127	592

This kind of counting is known as **absolute frequency**, which gives us a single value indicating the total number of observations. In looking at the table above, it is clear that there are far more observations with brown hair than black, blond, and red.

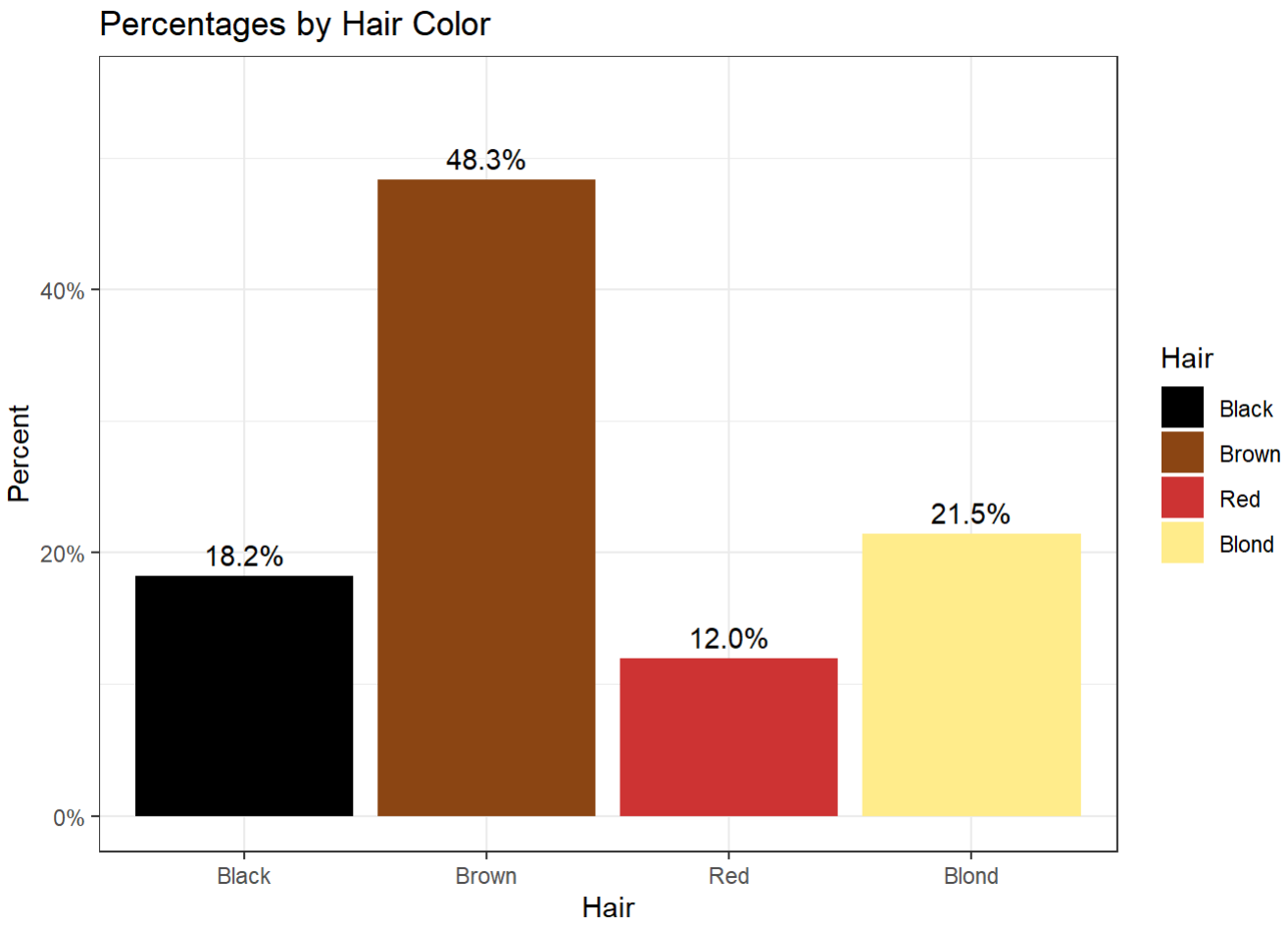
However, suppose somebody asks you how common brown hair is *relative* to other colors. Does it make sense to respond, “Oh, there are 286 individuals with brown hair?” Without knowing the values for the other hair colors, this number alone doesn’t carry much meaning. Is 286 observations a lot? It depends. Were 300 people examined? 3,000? Without knowing anything about the rest of the data, the absolute frequency may not be very useful.

In addition to actual counts of observations in categorical data, we may often be interested in **rates**. A rate can be as simple as taking the total number of a single category observed, and stating it in terms of the total number of observations. For example, instead of saying, “286 subjects who were observed had brown hair,” we might instead say, “286 of 592 subjects surveyed had brown hair.” Rates are also known as **relative frequencies**, because they are relative to a specific number of observations. More commonly, we use **percentages**, also known as **proportions**, which are a special type of rate or relative frequency – the count per 100 observations. That is, 286 of 592 subjects becomes $286/592 = 0.4831 = 48.31\%$. Let’s consider again the same table as above, this time in terms of percentages:

Black	Brown	Red	Blond
18.2%	48.3%	12.0%	21.5%

By considering all observations as rates per one hundred, we can quickly compare the relative counts of our observations. For example, we can quickly note that about half of the observations collected had brown hair, and almost twice as many had blond hair compared to red.

In addition to tables, we can also summarize categorical data visually. The most common figure used to represent categorical data is the **bar plot**. Below is a demonstration of a bar plot for the percentages of hair color in our data.



2.2.2 Advanced Categorical Summaries

Numerical and visual summaries become even more useful as our data becomes more complicated. Let's continue with the data we've been using, but now let us also break down observations with each hair color by sex as well. This process is known as **stratification**.

	Black	Brown	Red	Blond	Total
Male	56	143	34	46	279
Female	52	143	37	81	313
Total	108	286	71	127	592

First, we notice that by taking sums across the columns, we arrive at the same numbers that we had when only hair color was considered. If we sum horizontally across the rows, we also get the total number of observations of each sex. Note that the sum of both marginal totals add up to

592, the total number of observations. In other words, when stratifying the hair color counts by sex, we haven't lost any information related to the hair color, but we have *added information* to our summary about sex.

Getting stratified counts is straightforward; however, we now have several ways in which we might compute the percentages. For the table above, there are three ways we could compute percents.

- How many in each category, relative to the entire sample

Table 2.1: Relative to population

	Male	**Female**
Black	9.5%	8.8%
Brown	24.2%	24.2%
Red	5.7%	6.2%
Blond	7.8%	13.7%

Here, the percentages are computed by diving the count in each inner cell by the total sample size, 592. For example, there were 56 male respondents with black hair, representing 9.5% of the sample. Adding up all of the percentages gives us $\approx$ 100% (due to rounding to the first decimal place we actually get 100.1% here).

- How many of each hair color, within sex

Table 2.2: Proportion of hair color, by sex

	Male	**Female**
Black	20.1%	16.6%
Brown	51.3%	45.7%
Red	12.2%	11.8%
Blond	16.5%	25.9%
Total	100.0%	100.0%

Now our table looks very different. The percentages in the inner cells to not add up to 100%. Instead, the percentages have been computed relative to the total number of respondents in each sex. For example, we still have 56 male subjects with black hair, but relative to the total number of male subjects (279), this is 56/279 = 0.201 = 20.1%. There are 52 black-haired female respondents out of 313 total females, which gives us 53/313 = 0.166 = 16.6%. Since the percentages are computed relative to sex, we now see the percentages in *each column* add up to 100%. In other words, we have information about distribution of hair colors within sex.

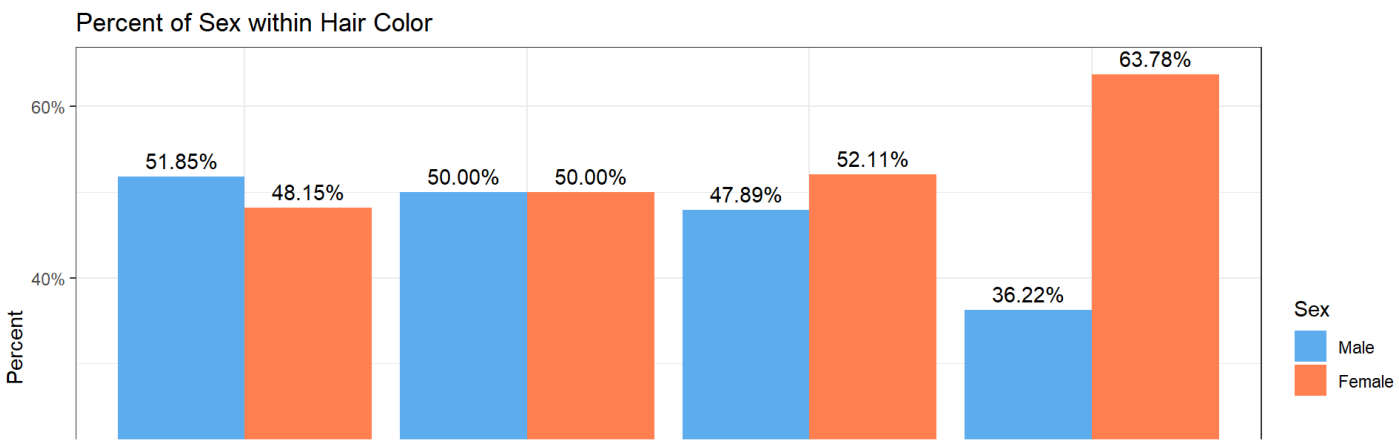
- How many in each sex, within hair color

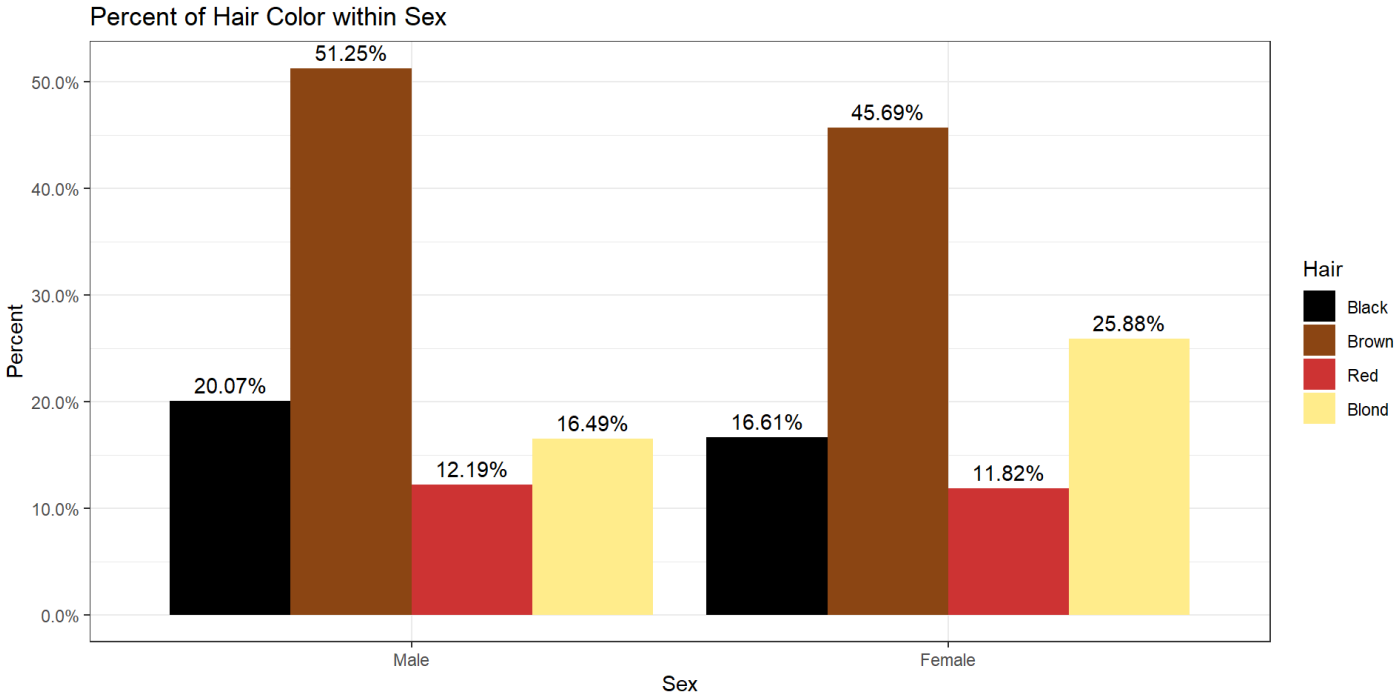
Table 2.3: Proportion of sex, by hair color

	Male	Female	Total
Black	51.9%	48.1%	100.0%
Brown	50.0%	50.0%	100.0%
Red	47.9%	52.1%	100.0%
Blond	36.2%	63.8%	100.0%

Similarly, we can also look at the relative frequencies of each sex within the four hair color categories. We have 56 black-haired males and 52 black-haired females, so 56/108 = 0.519 = 51.9% of the black-haired respondents are male and 53/108 = 0.481 = 48.1% are female. Since hair color is given in the rows of our table, we now have percentages that add up to 100% in each *row*.

All three of these tables are correct and informative. When deciding the best way to calculate percentages for tables like this, it will depend on the research question and the data at hand. We can also use stratification graphically, by creating multiple bar plots for each category of the stratification variable.





Definition 2.2

Absolute frequency: *The number of observations in a category*

Rate/Relative frequency: *The number of observations in a category relative to any other quantity*

Percent/Proportion: *The number of observations per 100*

Bar plot: *Visualization of categorical data which uses bars to represent each category, with counts or percents represented by the height of each bar*

Stratification: *The process of partitioning data into categories prior to summarizing*

2.3 Continuous Data

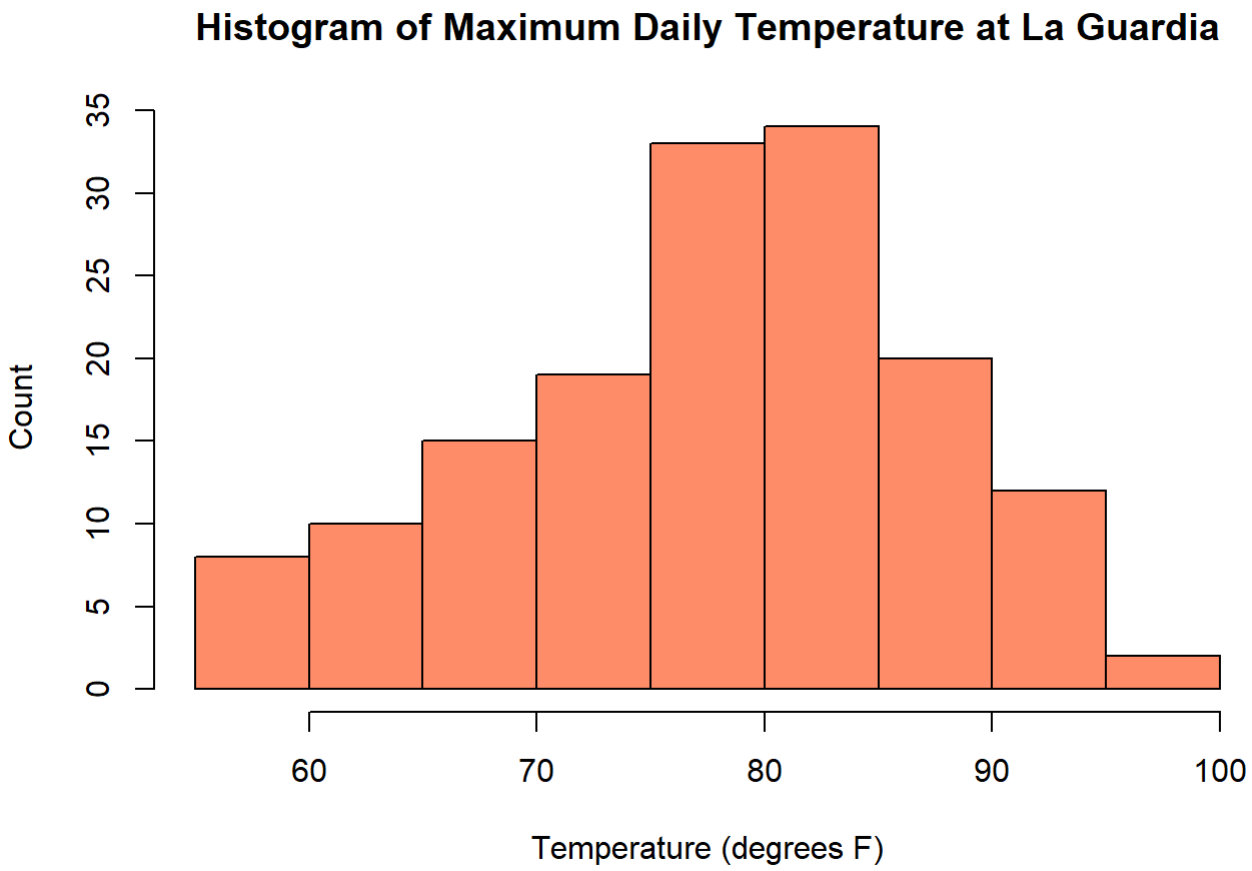
To explore graphical and numerical summaries of continuous data, let's consider a dataset which contains daily air quality measurements in New York from May to September 1973. Let's look at the first 10 rows of the data:

Month	Day	Temp	Wind	Ozone
5	1	67	7.4	41
5	2	72	8.0	36
5	3	74	12.6	12
5	4	62	11.5	18
5	5	56	14.3	9
5	6	66	14.9	28
5	7	65	8.6	23
5	8	59	13.8	19
5	9	61	20.1	8
5	10	69	8.6	25

This data contains continuous measurements on the air quality:

- maximum daily temperature in degrees Fahrenheit at La Guardia Airport (Temp)
- average wind speed in miles per hour (mph) at La Guardia Airport (Wind)
- average ozone in parts per billion (ppb) from 1:00pm - 3:00pm at Roosevelt Island (Ozone)

The tables and bar charts we introduced in Section 2 are great summaries when the data can be categorized and counted. With continuous data, however, there are no natural discrete categories into which our data can be organized. Instead, one way we can visualize the data is by first creating *bins* which partition an interval of possible values and then counting the number of observations that fall into each bin. For example, when considering a range of temperature values from 50°F - 100°F, we might construct bins at intervals of 5°F. Plotting the absolute or relative frequency of observations in each bin gives us a **histogram**. Below is an example of a histogram representing the maximum daily temperatures from our New York data:



In the histogram of temperature, we can readily see that on most days between May and September, the maximum temperature at La Guardia was between 75°F - 85°F. Some days were particularly chilly, with temperatures below 60°F and some days were quite hot, with temperatures above 90°F.

Histograms also provide a nice picture of the *distribution*, or "shape," of the observed data. Distributions of data can look very different for different sets of data, and we will consider them in more detail in Chapter 6. For now, its only relevant to know that distributions of data are often summarized with two types of measures – those that describe where the centers (peaks) of our data are and those that describe how spread out the data is about that peak. As these will be important concepts for the rest of this book, let's take a moment now to go into a bit more detail into each.

2.3.1 Measures of Centrality

While nothing can replace a picture, sometimes it is preferable to summarize our data with one or two numbers characterizing the most important information about the distribution. Often, we are most interested in information that describes the 'center' of a distribution, where the bulk of our

data tends to aggregate. The two most common ways to describe the center are with the *mean* and the *median*.

The **mean** is the most commonly used measure of the center of a distribution. Simply enough, the mean is found by taking the sum of all of the observations, and dividing by the total. With n observations, $x_1, x_2, \ldots, x_n$, we can mathematically express the mean, denoted as $\bar{x}$ (x-bar), in the following way:

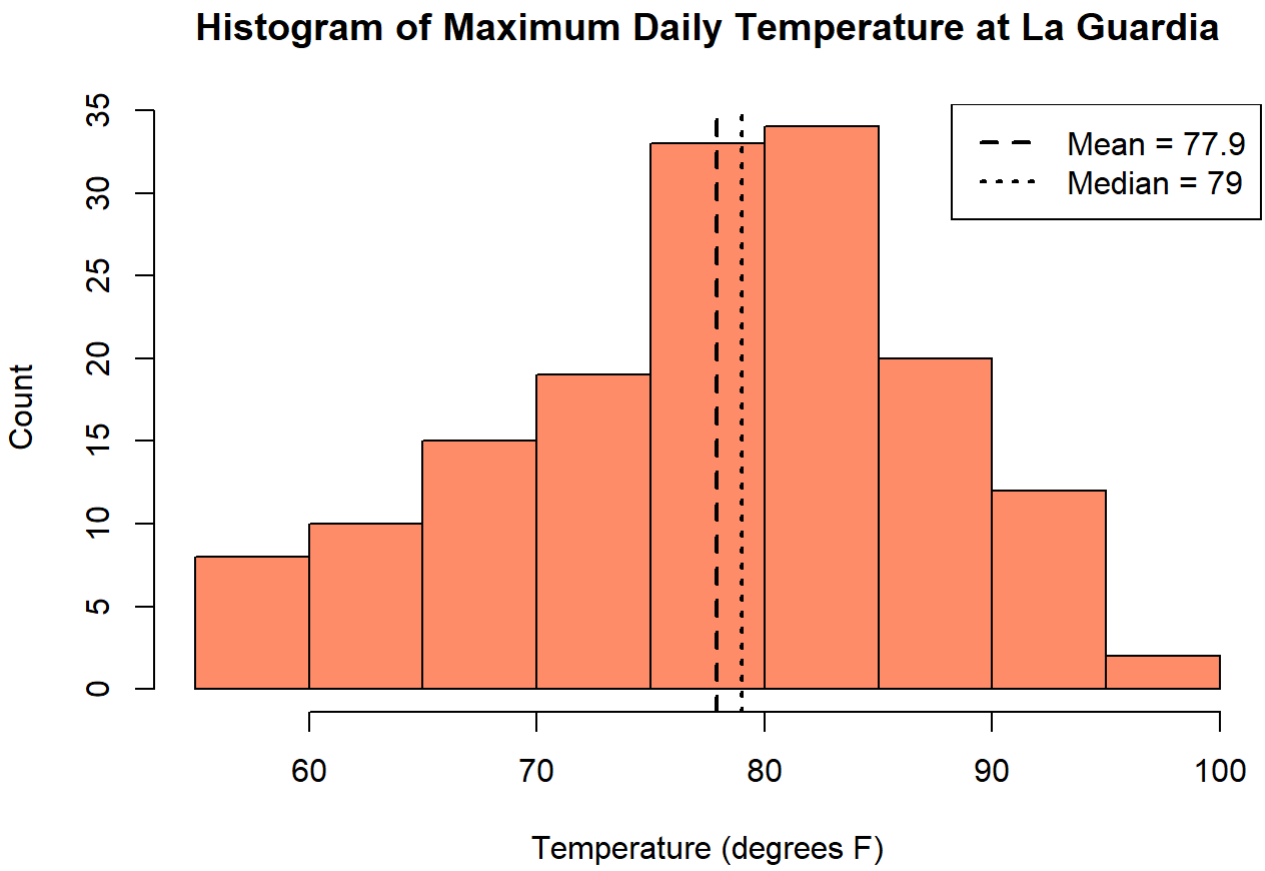
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The **median** is another common measure of the center of a distribution. In particular, for a set of observations, the median is an observed value that is both larger than half of the observations, as well as smaller than half of the observations. In other words, if we were to line our data up from smallest to largest, the median value would be right in the middle. Indeed, to find the median, we begin by arranging our data from smallest to largest. If the total number of observations, n, is odd, then the median is simply the middle observation; if n is even, it is the average of the middle two.

Examples:

- $1, 2, 2, 3, \color{red}{5}, 7, 9, 10, 11 \quad \Rightarrow \quad \text{Median} = 5$
- $1, 2, 2, 3, \color{red}{5}, \color{red}{6}, 7, 9, 10, 11, \quad \Rightarrow \quad \text{Median} = \frac{(5+6)}{2} = 5.5$

For the New York temperature data, the mean is 77.9°F and the median is 79°F. These values are very similar to each other, and both fall near the peak in the histogram.



However, it is not always the case that the mean and median will be similar. Let's consider an example in Table 2.4 where we collect $n = 10$ samples of salaries for University of Iowa employees:

For our sample, we find that the mean is $569,266, but the median is ($48,962 + $50,879) / 2 = $49,921. It turns out our sample included the highest paid university employee – the head football coach. This extremely high salary has caused the mean to be very large – larger than the remaining 90% of the salaries in our sample *combined*. The median, on the other hand, ignoring the extremes ends of our distribution and focusing on the middle, is not impacted by the football coach's salary. Consequently, in this case, it is a much better reflection of the typically university employee's salary than the estimate found with the mean.

Table 2.4: University of Iowa Salaries

$31,176	$130,000
$50,879	$37,876
$34,619	$144,600
$103,000	$48,962
$36,549	$5,075,000

This one high salary, which is not representative of most of the salaries collected, is known as an **outlier**. From the example above, we have seen that the mean is highly sensitive to the presence of outliers while the median is not. Measures that are less sensitive to outliers are called **robust**

measures. The median is a robust estimator of the center of the data.

We have seen an example where the mean and median are quite close and an example where they are wildly different. This begs the broader question – when might we expect these measures of central tendency to be the same, and when might we expect them to differ? Here, we consider a collection of histograms showing us different "shapes" that the distributions of our data may take.

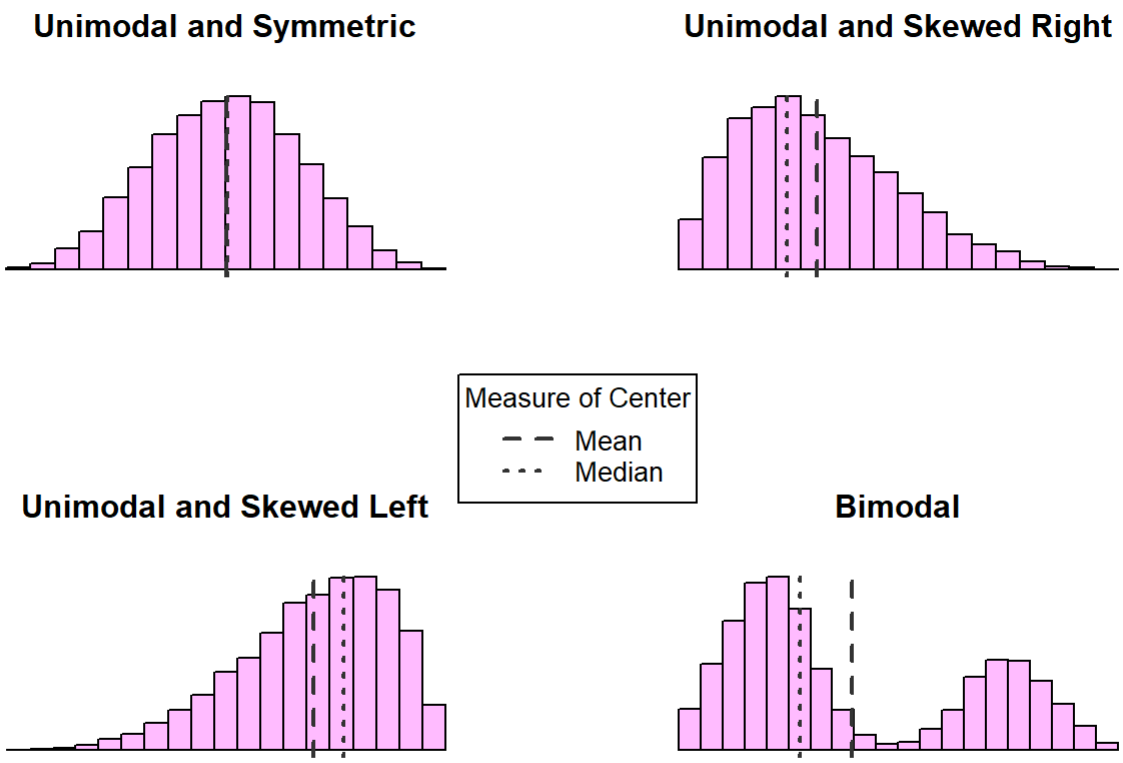


Figure 2.1: Examples of modality and skew

The shape of a distribution is often characterized by its **modality** and its **skew**. The modality of a distribution is a statement about its modes, or "peaks." Distributions with a single peak are called *unimodal*, whereas distributions with two peaks are call *bimodal*. *Multimodal* distributions are those with three or more peaks. The skew on the other hand describes how our data relates to those peaks. Distributions in which the data is dispersed evenly on either side of a peak are called *symmetric distributions*; otherwise, the distribution is considered skewed. The direction of the skew is towards the side in which the tail is longest. Examples of modality and skew are presented in Figure 2.1.

When the data is unimodal and symmetric, the mean and median are indistinguishable. However, when there is skew or multiple peaks, we see the mean and median start to differ. When the distribution is skewed, the mean is pulled towards the tail. On the other hand, the number of very large/small observations is relatively small, so the median remains closer to the peak where the majority of data lies. When the distribution is bimodal, neither the mean or median can summarize the distribution well. In that case, it might be better to characterize the center of each peak individually.

Definition 2.3

Mean: *The average value, denoted $\bar{x}$ and computed as the sum of the observations divided by the number of obervations*

Median: *The middle value or 50th percentile, the value such that half the observations lie below it and half above*

Outlier: *Extreme observations that fall far away from the rest of the data*

Robust: *Measures that are not sensitive to outliers*

Unimodal: *Characterization of a distribution with one peak*

Bimodal: *Characterization of a distribution with two peaks*

Multimodal: *Characterization of a distribution with three or more peaks*

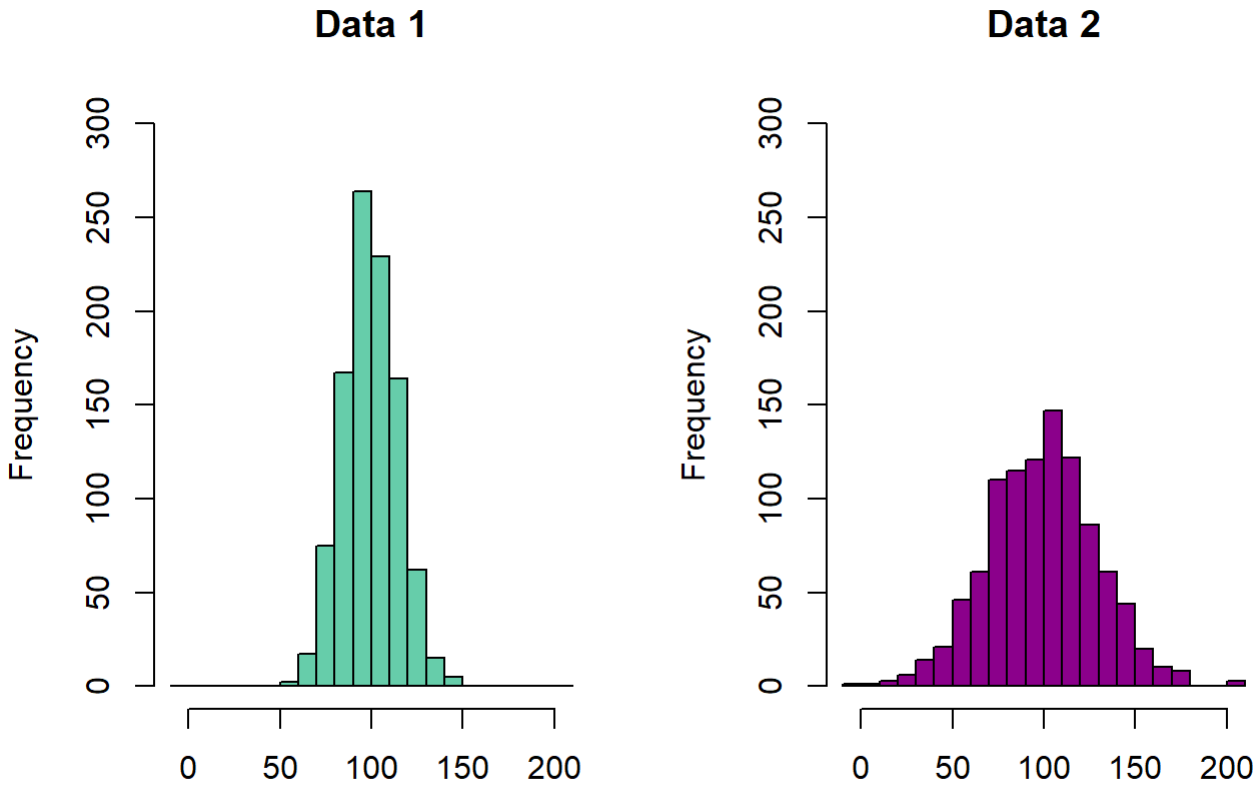
Symmetric: *Characterization of a distribution with equal tails on both sides of the peak*

Skewed right: *Characterization of a distribution with a large tail to the right of the peak*

Skewed left: *Characterization of a distribution with a large tail to the left of the peak*

2.3.2 Measures of Dispersion

In addition to measuring the center of the distribution, we are also interested in the *spread* or *dispersion* of the data. Two distributions could have the same mean or median without necessarily having the same shape. For example, consider the two distributions of data shown below. Each distribution represents a sample of 1,000 observations with mean of 100.



Despite the mean values of each of these distributions being the same, we can clearly see that they are different. On the left, we see that nearly the entire range of the observed data falls between 50 and 150. For the distribution on the right, the data is much more spread out, taking on values near 0 and 200. In order to accurately capture these differences, we need a second numerical summary describing the degree to which the data is spread about its center. Here, we consider two broad categories – those based on percentiles and those based on the variance.

2.3.2.1 Percentiles and IQR

Perhaps the most intuitive methods of describing the dispersion of our data are those associated with **percentile**-based summaries. Formally, the pth percentile is some value V_p such that

1. $p\%$ of observations are less than or equal to V_p
2. $(100 - p)\%$ of observations are greater than or equal to V_p

Informally, percentiles quantify where observations fall, relative to all of the other observations in the data. Two of the best known percentiles are the 1^{st} and 99^{th} percentiles, more commonly referred to as the minimum and maximum values. Together, these two numbers describe the **range** of our data. The next most common value is the 50^{th} percentile – the median – which we

recall describes the value for which half of our observations and greater (or lesser) than. We are also often interested in determining the 25^{th} and 75^{th} as well, which, respectively, mark the midpoints between the minimum and the median and the median and the maximum. Along with the median, these three percentiles make up the *quartiles* of our data, denoted

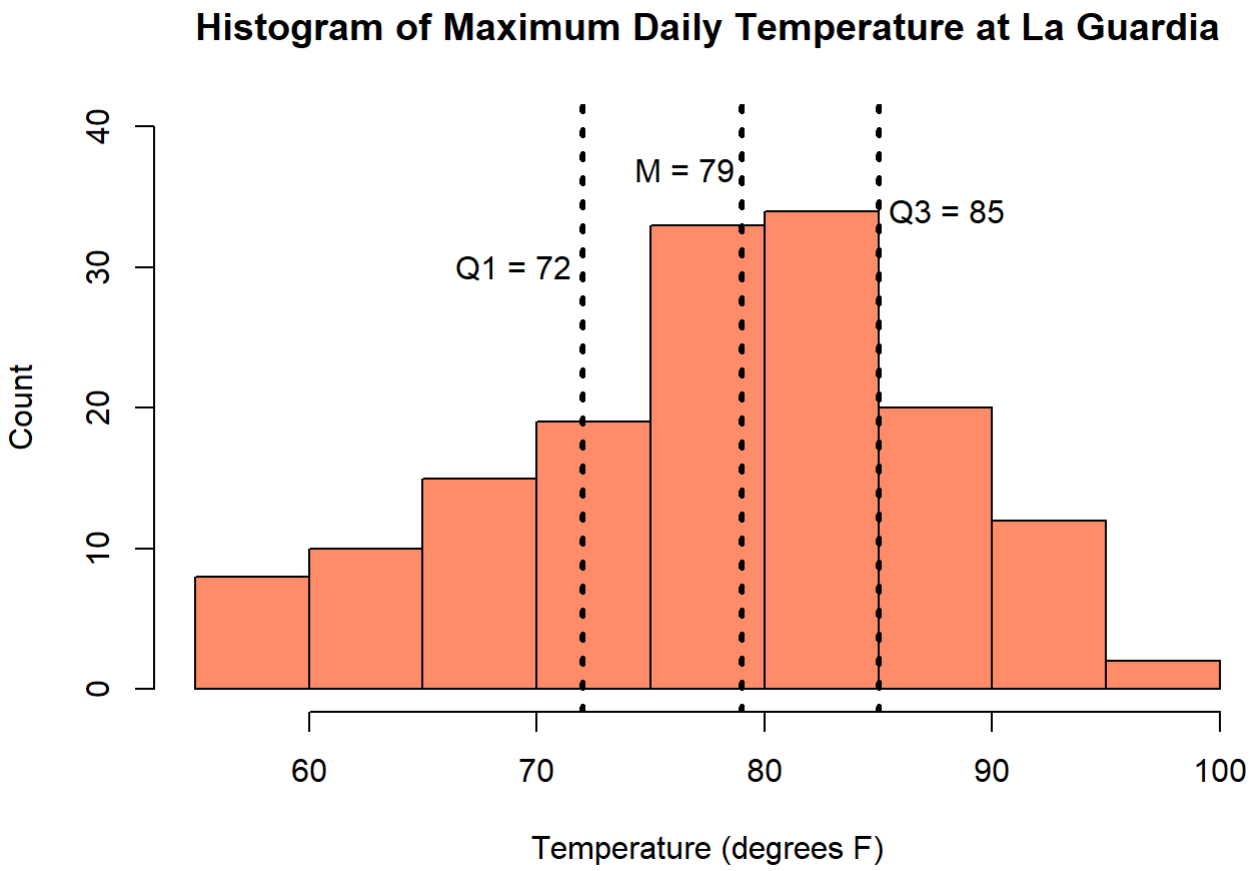
$$\begin{aligned} Q_1 &= 25^{th} \text{ percentile} = 1^{st} \text{ or lower quartile} \\ Q_2 &= 50^{th} \text{ percentile} = 2^{nd} \text{quartile or median} \\ Q_3 &= 75^{th} \text{ percentile} = 3^{rd} \text{ or upper quartile} \end{aligned}$$

We know the median, M, is the value such that 50% of observations are less than the median and 50% are greater than it. Q_1 can be said to represent the median of the smaller 50% of all observations, while Q_3 can be said to be the median of larger 50%. In other words, 25% of the data is below Q_1, 25% of the data is between Q_1 and M, 25% of the data is between M and Q_3, and the remaining 25% of the data is larger than Q_3.

A commonly used percentile-based measure of spread combining these measures is the **interquartile range (IQR)**, defined as

$$\text{IRQ} = Q_3 - Q_1.$$

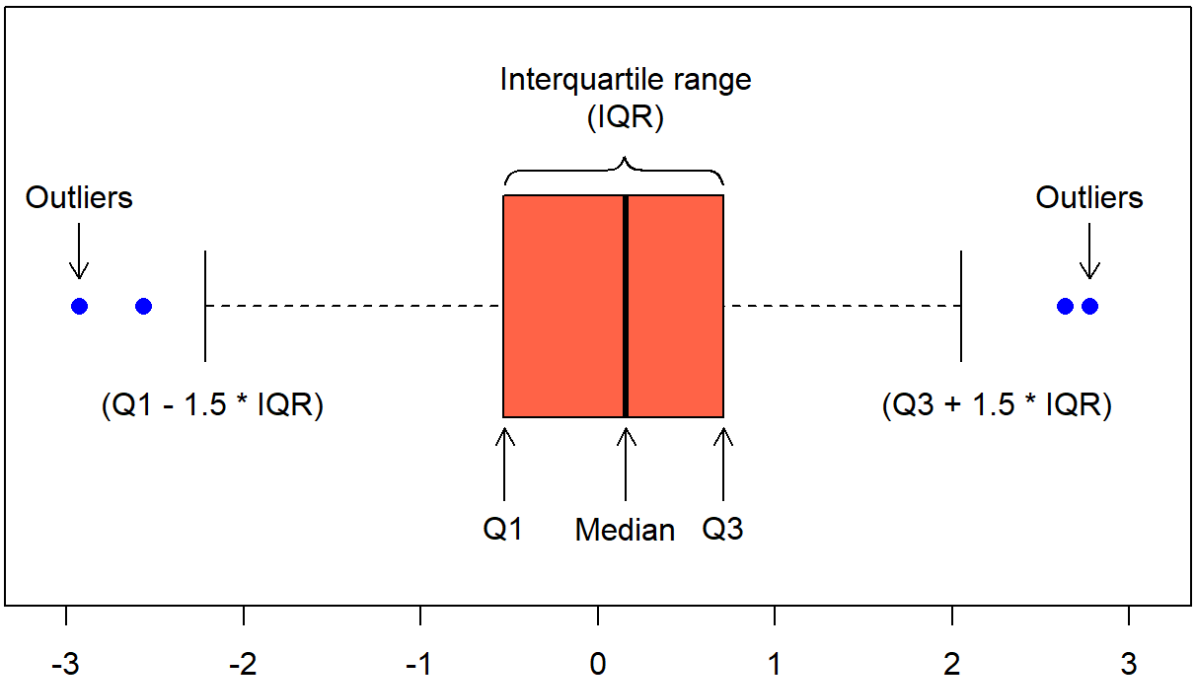
Because it is the difference between the upper and lower quartile, it represents the distance covering the middle 50% of the data. We may also report the range as an interval as (Q_1, Q_3). For the New York temperature data, $Q_1 = 72$, $Q_3 = 85$. The IQR is therefore $85 - 72 = 13$ and tells us that 50% of the days between May and September had temperatures between 72°F and 85°F.



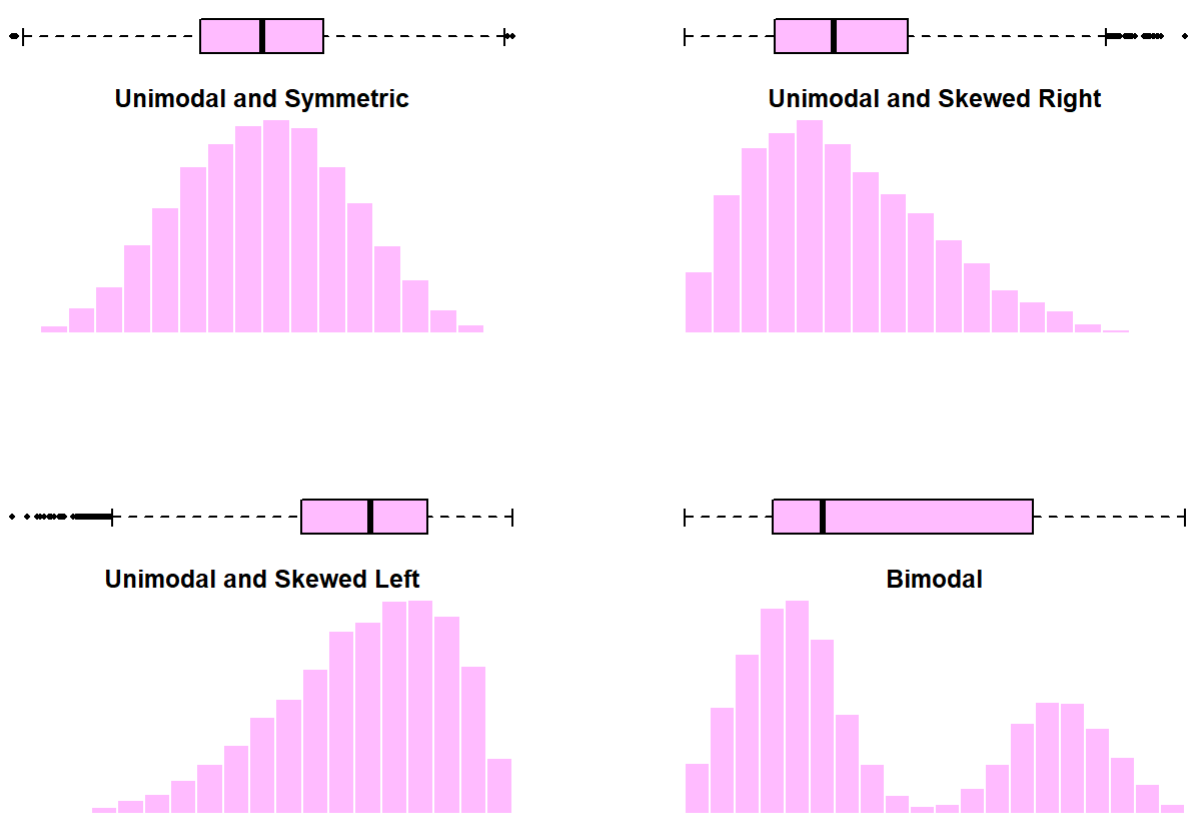
The IQR is not impacted by the presence of outliers, so it is considered a robust measure of the spread of the data. So, like the median, it enjoys the quality of being a robust measure of the data.

Percentiles are also used to create another common visual representation of continuous data: the **boxplot**, also known as a **box-and-whisker plot**. A boxplot consist of the following elements:

- A box, indicating the Interquartile Range (IQR), bounded by the values Q_1 and Q_3
- The median, or Q_2, represented by the line drawn within the box
- The “whiskers,” extending out of the box, which can be defined in a number of ways. Commonly, the whiskers are 1.5 times the length of the IQR from either Q_1 or Q_3
- Outliers, presented as small circles or dots, and are values in the data that are not present within the bounds set by either the box or whiskers



Just like histograms, boxplots can also illustrate the skew of a data. In a histogram, the skewed was named after the location of the tail and in a boxplot, this corresponds to the side with a longer whisker. Here we can see histograms and boxplots for various distributions of data.



Definition 2.4

Percentile: *A value, V_p such that p% of observations are smaller than V_p and $1-p$% of observations are larger than V_p*

Range: *The distance between the minimum and maximum values in a dataset*

Interquartile range (IQR): *The middle 50% of the data, the difference between the upper and lower quartiles*

Boxplot/box-and-whisker plot: *Visualization of continuous data which is based on percentiles*

2.3.2.2 Variance and Standard Deviation

The **variance** and the **standard deviation** are numerical summaries which quantify how spread out the distribution is around its mean. This is done by calculating how far away each observation is from the mean, squaring that difference, and then taking an average over all observations. For a sample of n observations $x_1, x_2, \ldots, x_n$, the variance, denoted by s^2, is calculated as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

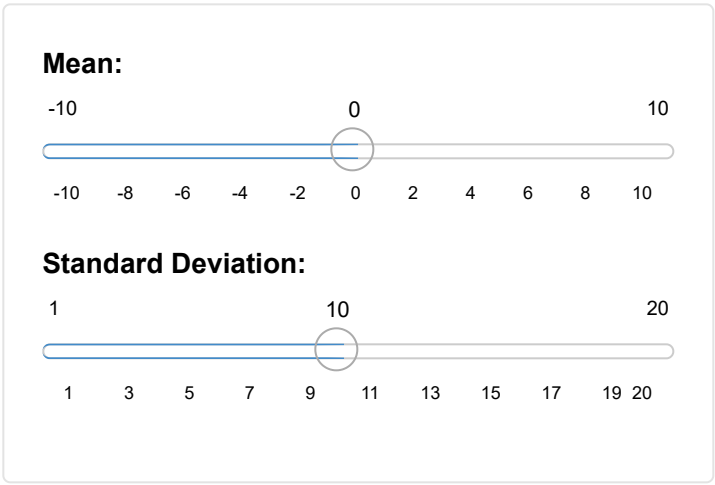
The standard deviation, denoted s, is a function of the variance. Specifically, it is the square root of the variance $s = \sqrt{s^2}$. Because the variance uses the squared differences between each observation and the mean, its units are the square of the units of the original data. For example, the New York temperature measurements have a mean of 77.9°F and a variance of 89.6°F^2, a value that does not readily lend itself to interpretation. The standard deviation, on the other hand, takes the square root, putting the units back on the original scale. For the temperature data, the standard deviation is 9.5°F. Because of this, the standard deviation is often preferred as a measure of spread over the variance.

Finally, unlike the median and the IQR, which are based the percentiles of the observed data, both the variance and standard deviation are calculated based on the mean. Recall that the mean is *not* a robust outlier and is highly sensitive to skew or the presence of outliers. Consequently, the variance and the standard deviation are also very sensitive. When the data are unimodal and symmetric, choosing between the mean/standard deviation and the median/IQR is largely a matter of preference. However, when the data is skewed or has large outliers, more robust statistics such as the median and IQR are preferred.

Exercise 2.1 Exercises

The applet below is designed to help illustrate the properties of the mean and standard deviation. You can vary the mean between -10 and 10 and the standard deviation between 1 and 20. The histogram on the right displays the unimodal and symmetric distribution of data with the specified mean and standard deviation.

Mean and Standard Deviation



1. Set the standard deviation to 8 and use the slider to vary the mean of the distribution. What properties of the histogram change as the mean is varied? What stays the same?
2. Now, set the mean to 0 and vary the standard deviation. What properties of the histogram change as the standard deviation is varied? What stays the same?
3. Let's examine more closely how the standard deviation affects the distribution of data. Keep the mean at 0, and set the standard deviation to 4, 8, 12, 16, and 20.

a. For each standard deviation, fill out the following table with the approximate minimum and maximum values of the data.

Standard deviation	Minimum and maximum
4	
8	
12	
16	
20	

b. What do you notice about the relationship between the standard deviation and the minimum/maximum values observed in the data? Is there any noticeable pattern in your table?

Definition 2.5

Variance: *The average of the squared differences between each data value and the mean, denoted* s^2

Standard deviation: *The square root of the variance, denoted* s

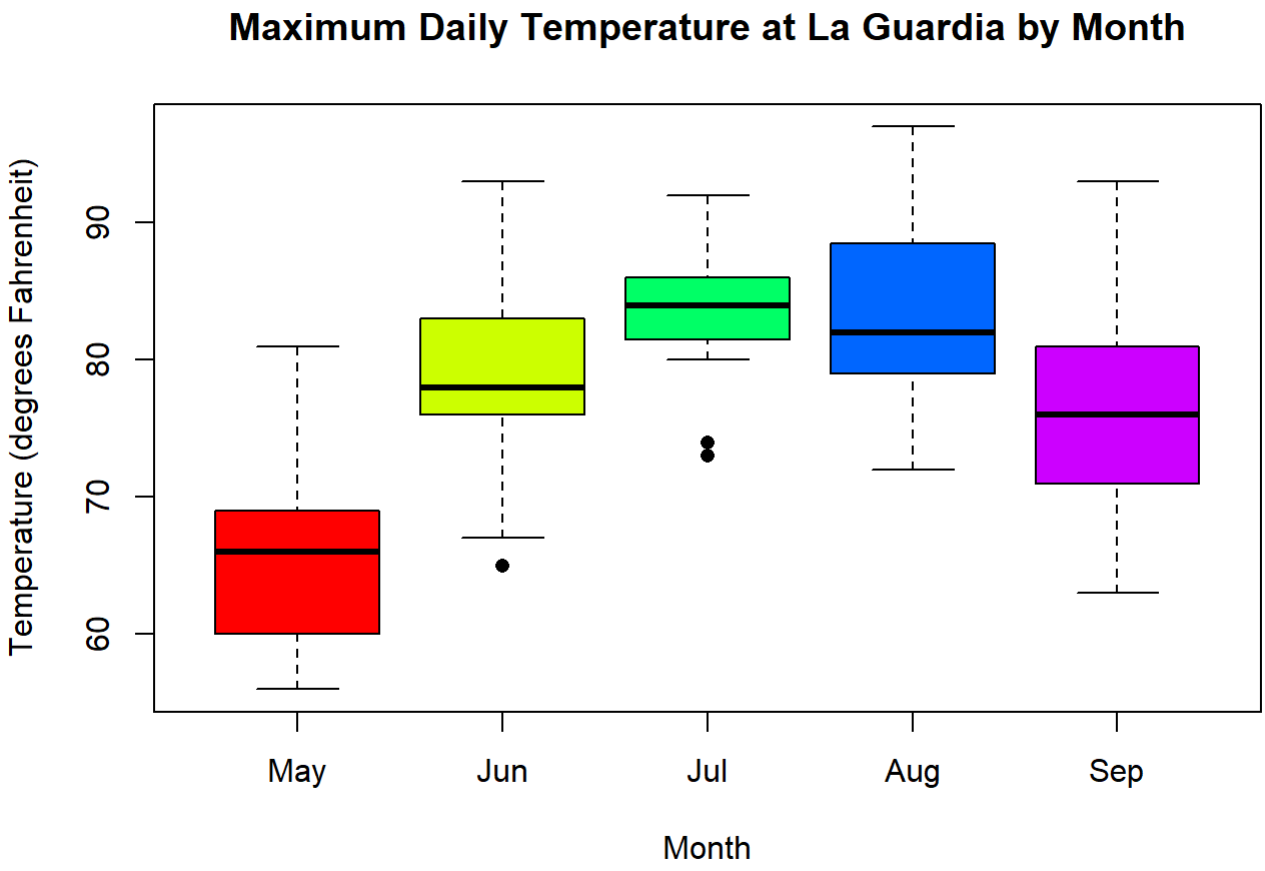
Percentile: *Summaries providing the location of observations relative to all other observations in the data*

2.4 Advanced Data Visualizations

As we saw in the last section, histograms and boxplots are useful when describing the distribution of one continuous measurement. What if we have both categorical and continuous measurements or multiple continuous measurements? Let's return to the New York air quality data:

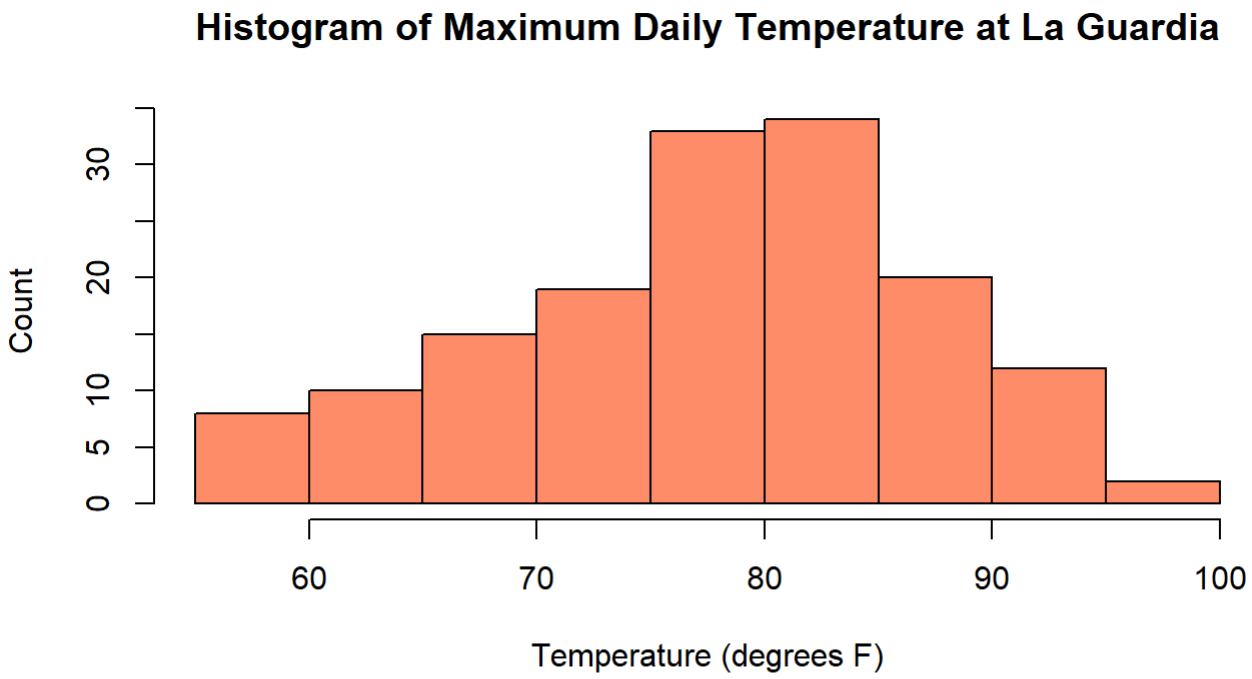
Month	Day	Temp	Wind	Ozone
5	1	67	7.4	41
5	2	72	8.0	36
5	3	74	12.6	12
5	4	62	11.5	18
5	5	56	14.3	9
5	6	66	14.9	28
5	7	65	8.6	23
5	8	59	13.8	19
5	9	61	20.1	8
5	10	69	8.6	25

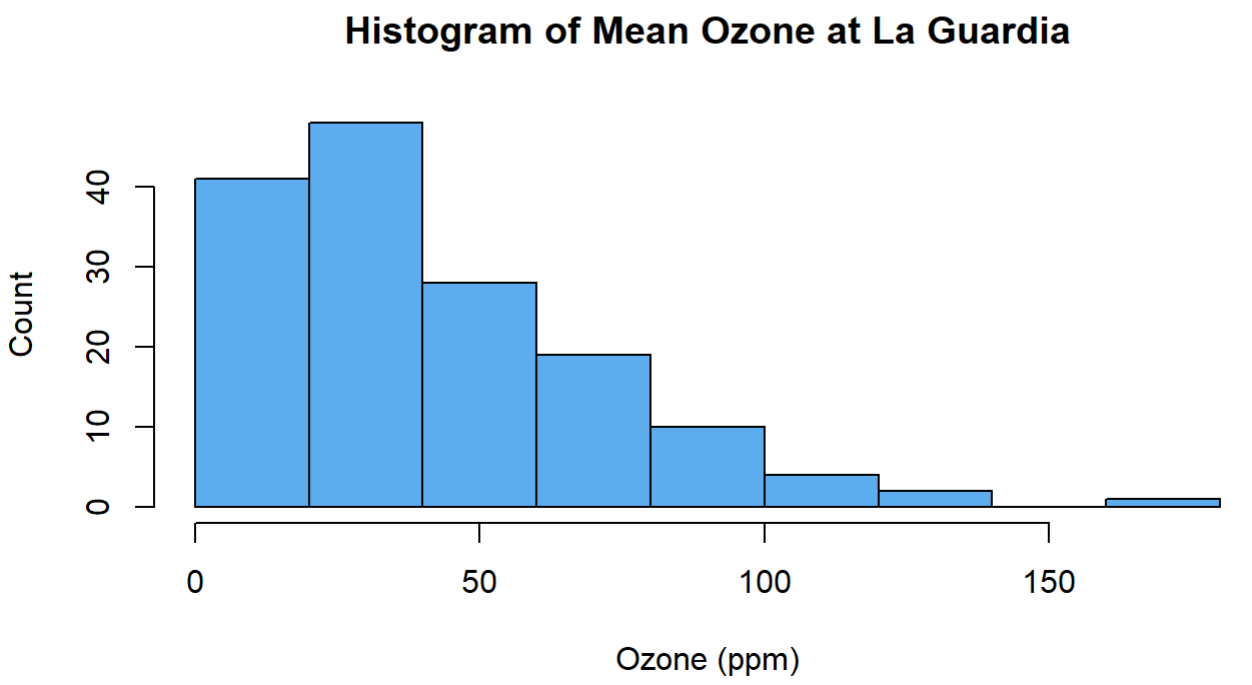
One thing we might be interested in is the distribution of temperatures in each month. Month can be considered an ordinal categorical variable, and temperature is continuous. When we want to see the distribution of a continuous variable across multiple categories, we can look at side-by-side boxplots.



By putting all of the boxplots next to each other with the same temperature range, we can quickly detect trends and differences in temperature in the five months. Median temperature peaks in July/August and is lowest in May. Temperatures in May and September are more variable, whereas the range of temperatures in July is less spread out. There is some overlap between all of these boxplots, which means that although July and August have higher temperatures more often, there are days in the other months that are just as hot or days in July that are cooler.

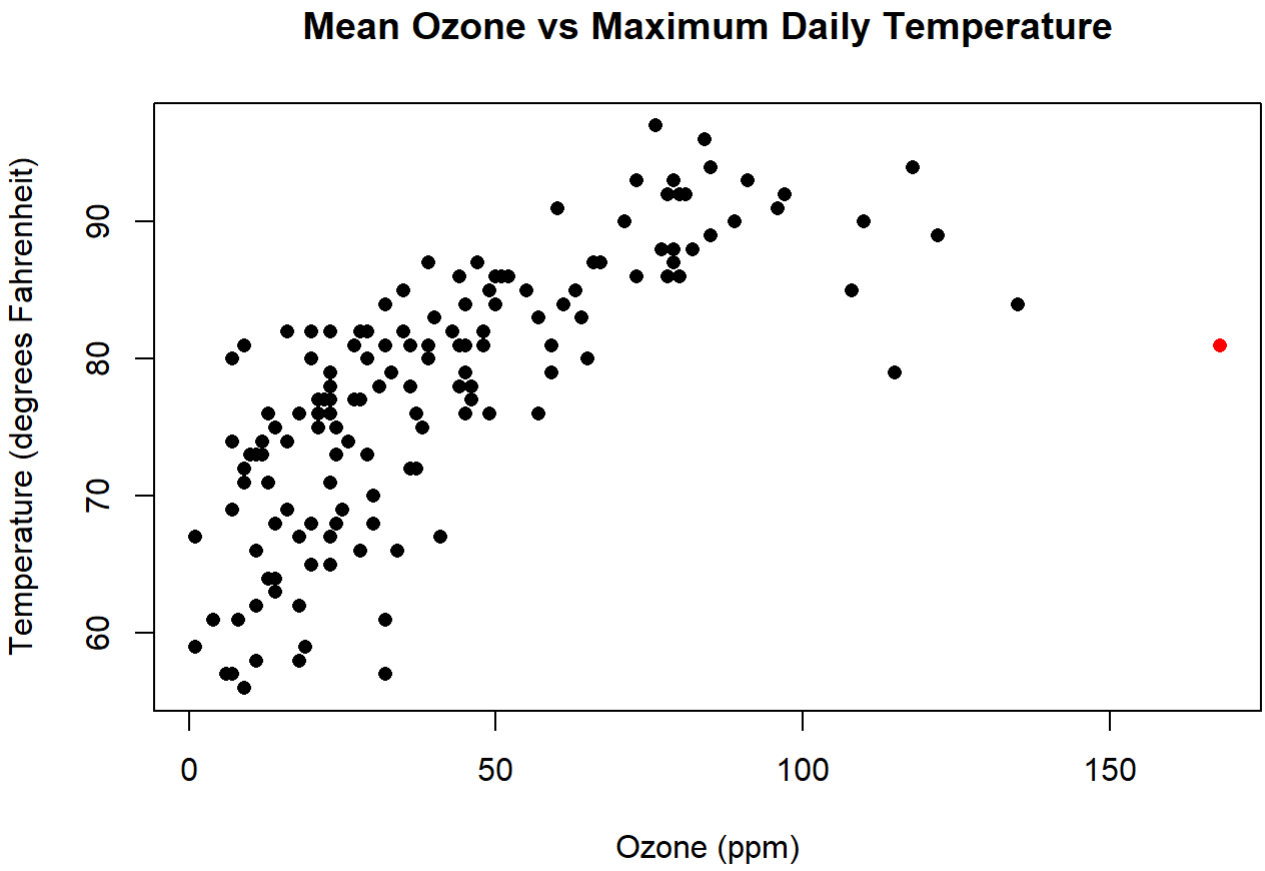
Now let's consider the case in which we have two continuous variables. Suppose, for example, that we are interested in both temperature and ozone levels, as well as the relationship between the two of them. We might first consider them separately; presented below is a visualization of the distribution of both temperature and ozone levels in the New York air quality data:





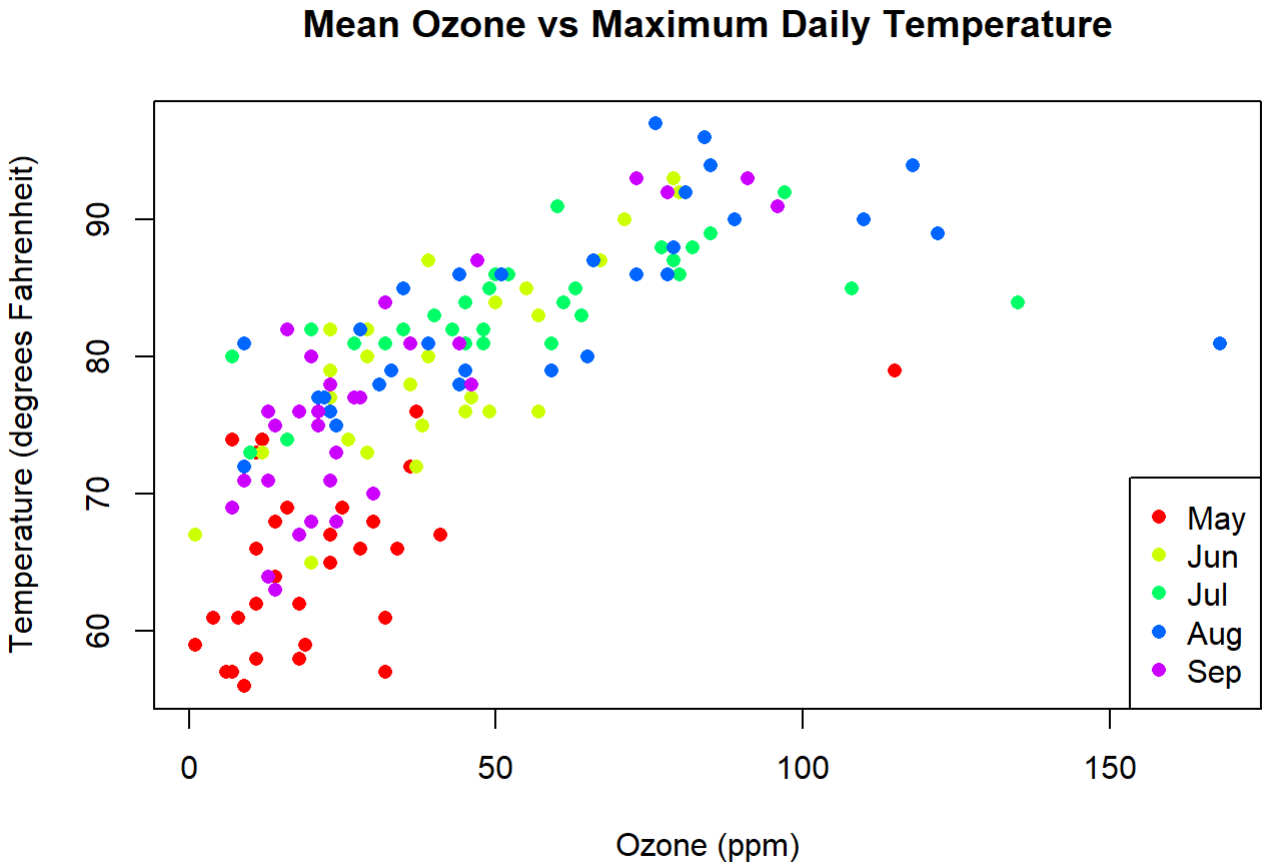
While temperature is unimodal and symmetric in distribution, the ozone concentration measurements are skewed right. These plots are useful in studying each of these variables individually, but they do not provide any information about the relationship between temperature and ozone concentration, which we may expect to be related.

To look at the relationship between two continuous variables, a common visualization is the **scatter plot**. A scatter plot is a two-dimensional plot of data pairs, with one variable represented on the x-axis, the horizontal range of the plot, and the other variable on the y-axis, or the vertical range. Each dot on a scatter plot represents a single observation. In the plot below, we have put the mean ozone on the x-axis and temperature on the y-axis, with each dot representing one day between May and September. For example, on August 25th, the maximum temperature was 81°F, and the mean ozone concentration was 168ppm. For illustration, this day is colored red on the scatter plot.



Plotted together, the relationship between ozone concentration and temperature becomes apparent – lower temperatures tend to be associated with lower ozone concentrations while higher temperatures are associated with higher ozone levels. The scatter plot is able to illustrate an important relationship that wouldn't be detectable just by looking at histograms of each variable separately.

We may further describe our data by including a third categorical variable. In our scatter plot of ozone and temperature, we could also color the points by the month on which they occurred.



After coloring the points by month, we can see some clusters forming as temperature and ozone concentration on days closer together are more likely to be similar. By adding information about a third variable, we are able to illustrate more information about our data in one figure.

Numerical summaries and graphical visualizations provide us with tools to capture important features of our data. Understanding the distribution and skew of a variable help us know the best way to summarize it numerically. Summaries and figures can become increasingly complex as they incorporate more variables, but these illustrations can be informative for answering and producing important research questions and hypotheses.

Definition 2.6

Scatter plot: *Visualization of two continuous measures which uses dots to represent each data pair*

3 Study Design

> “Randomization is too important to be left to chance.”
>
> — J. D. Petruccelli

Learning objectives

1. Be familiar with the general principles of study design
2. Understand how poor study design can bias results
3. Learn about various sampling biases, such as selection and non-response bias
4. Be familiar with other types of bias, such as extrapolation and confounding
5. Know what a randomized controlled double-blind experiment is and why it is the gold standard

3.1 Importance of Study Design

There are many avenues by which data may land on the desk of a statistician. Many times, data is simply *observational*, that is, the data is collected by researchers without any control over independent variables. Such was the case with the La Guardia air quality data explored in Chapter 2, where in addition to temperature and ozone data, information was also collected for month, date, windspeed, and solar radiation. While this kind of study is useful for exploring the relationship between processes, it is often limited in its ability to make causal statements and inference.

More useful to researchers, and especially in biomedical studies, are those in which clinicians take a more proactive approach, manipulating independent variables in an attempt to determine a more direct relationship with an outcome, or the dependent variable. This is particularly important in studies that seek to determine the efficacy of a new drug or treatment. While there are a number of different designs that are used by medical researchers for a variety of purposes, we

limit our discussion in this chapter to a broad class of studies known as *clinical trials*. More specifically, we will discuss a class of studies known as *randomized controlled clinical trials*, which are considered the gold-standard in medical research.

We will begin by addressing issues with variability and bias, followed by an outline of the procedures that typically go into a clinical trial. Finally, we will explore a list of specific biases that statisticians and researchers attempt to mitigate through effective study design.

3.2 Variability and Bias

In Chapter 1, we defined statistics by its ability to quantify uncertainty regarding the potential outcome of a random process. In order to do this most effectively and make accurate inference, researchers and statisticians do their best to remove potential sources of variability. Variability can manifest in a number of ways, ranging from the standard biological variability associated with a given population (genetic variation), as well as imprecision associated with measurement (measuring blood pressure, weight, etc.,). While these sources of variation do impact the ability to make precise inference, the effects associated with this kind of variability often “even out” in the end. In other words, though it may make inference less certain, it does not make it incorrect.

Similar to variability is the concept of **bias**. Whereas variability is used to describe the effects of random variation and imprecision, bias refers to a characteristics of a process or technique in which the measured outcome is *systematically different* from the true value it is meant to represent. For example, a researcher may be interested in taking the weight of patients at different times during the day. Due to fluctuations in metabolism, an individual patient may have a slightly different weight at different times of the day, resulting in variability. However, if the scale that is used is miscalibrated and always reports a weight that is ten pounds less than the true value, this would be bias. Unfortunately, there are very few tools at a statisticians disposal to address bias, and there is frequently no way to test if it is present or not. However, a well designed study can reduce this risk and put researchers and statisticians in the best possible place to perform successful science.

As a critical piece of the statistical framework, **study design** is the process of organizing, conducting, and analyzing a scientific experiment with the intention of resolving a hypothesis or achieving research goals. There are many types of study designs with various strengths and

limitations, the entirety of which could not possibly be covered here. Instead, we will focus on some of the main considerations when designing a study, as well as the statistical implications associated with common ways in which study design can go wrong.

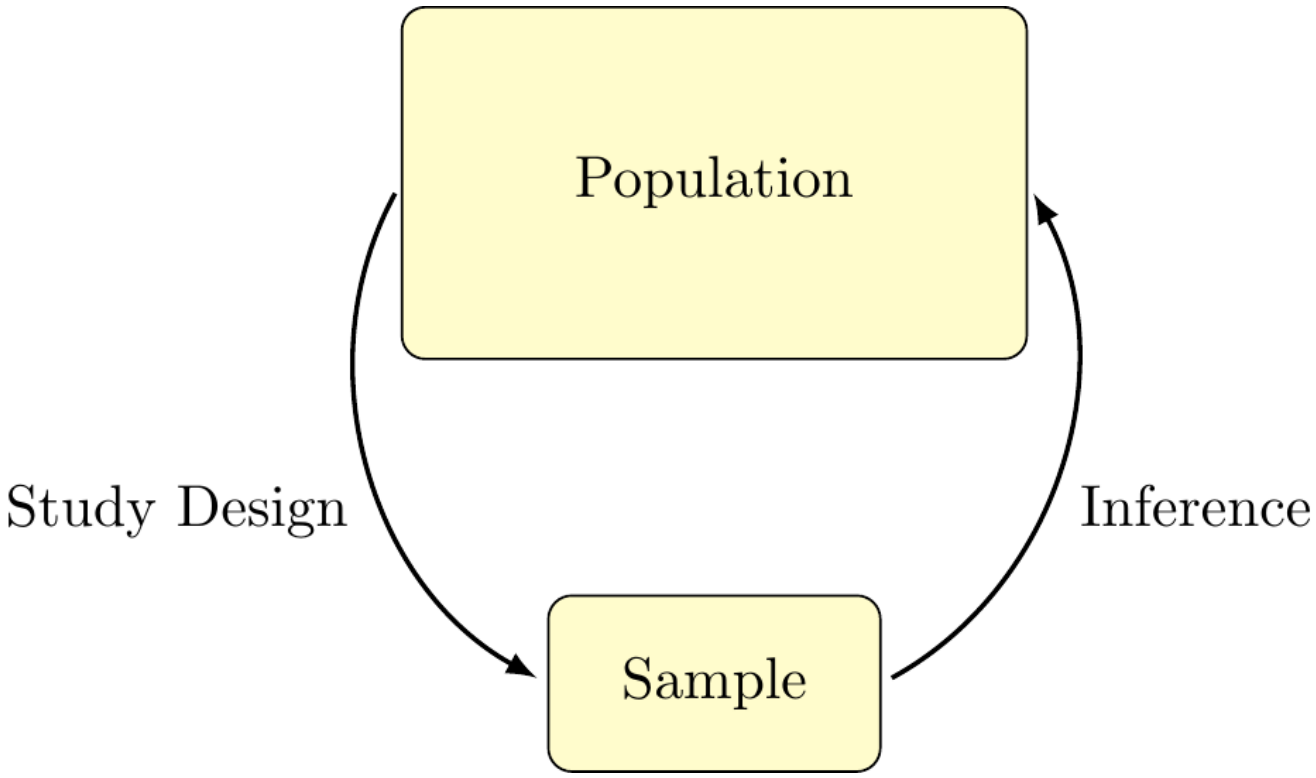


Figure 3.1: Statistical Framework

In an ideal world, the study design process might look a little bit like this:

1. We start with a list of everyone in our population of interest.
2. Each person in this population is equally easy to sample, and the cost of sampling everyone is the same
3. We use our list to randomly select a proportion of individuals in the population who will participate in our study
4. Everybody sampled is eager to help us with our study, and nobody refuses to participate
5. The outcome of interest can be accurately measured on each subject and directly answers the research question.

Of course, real life situations are rarely ideal.

Bias can occur at various points during a research study and materialize in many forms. The ability to identify bias is a crucial skill for researchers to have when evaluating the merit of other studies in the literature or designing studies of their own. In this chapter, we focus on defining several common types of bias common in medical research, how each type can impact study results, and how they can be avoided. This is not meant to be an exhaustive list – if you are

interested in learning about other types of bias, check out the Catalogue of Biases. We will categorize different types of bias broadly by when they occur: pre-trial, during trial, and post-trial, although all types of bias should be considered when planning a study.

Definition 3.1

Study design: *The process of obtaining the best sample to answer the research questions*

Scientific/Research Bias: *Any process or technique whereby the actual outcome is systematically different from the true value it is intended to estimate*

3.3 Clinical Trials

There are several phases involved in a formal clinical trial, with early stages focused on determining safety and evidence for potential efficacy of a treatment. Often, these trials are small, usually with numbers between 10 and 100 patients. Once enough evidence has been collected to demonstrate the safety and possible efficacy of an intervention, larger Stage 3 clinical trials are conducted, often with hundreds or even thousands of patients, and usually at multiple clinics or sites. Yet, in all cases, the general framework is the same. This helps ensure consistency across studies and sites and prevents researchers from introducing sources of bias that may unintentionally influence results. What follows are brief principles that are used to guide each step of the study process.

Study Protocol and Human Subjects

Before any study on human subjects begin, researchers are tasked with writing and proposing a *study protocol* that outlines who will be studied, what metrics will be collected, how the data will be analyzed, and a set of procedures to actively monitor the safety of the participants, as well as address how specific problems will be handled, if they arise. It is critical that these protocols are written and adhered to before a study begins; decisions that are made once the study has begun may be influenced by data that has already been seen. Further, specifying the main analysis in advance ensures that the hypothesis in question is being adequately answered. Without this safeguard in place, investigators may let the data observed direct the methods of analysis. This is known as *data snooping*, which can drastically impact the validity of the results.

Once the study protocol has been written, it is then submitted to an ethics/review board to ensure that subjects are treated fairly and humanely, and that all participants are able to give informed consent. This also helps evaluate the potential risks faced by subjects against any potential benefits from the trial. In the United States, this process is governed by an *Institutional Review Board*, or IRB. Further approval may be needed in cases involving the use of drugs or radioactive materials.

Participant Selection

Before a study can begin, subjects must be enrolled and consent to performing in a study. Typically, eligible participants are selected for a study based on a set of eligibility criterion to ensure that subjects come from a population of interest. For example, in a study designed to test the efficacy of a new medication for lowering cholesterol, clinicians may only select subjects who have a demonstrated risk for heart disease. Similarly, it is often important to have a set of criteria that can *exclude* certain subjects, such as those with confounding issues that may make it difficult to limit the amount of variability in a study.

Of importance here is the fact that the results of study are only valid for populations from which the subjects in the study were drawn. If participants in a study are limited to a certain demographic, such as age or sex, then the findings may not hold for a more general population. This may result in a form of extrapolation bias, defined below, and is often a concern when extending treatment to minors, who are often not included in clinical trials.

Control Groups and Randomization

In order to correctly estimate the efficacy of a treatment, it is vital to have a *control group* with which to evaluate it. In some cases, the control group may receive an existing method of treatment; in others, it may be a *placebo*, or inactive form of a drug or intervention. By collecting a study sample prior to assigning subjects to either a treatment or control group, researchers can minimize the risk that groups we are comparing differ in significant ways. To further minimize this risk, it is also necessary that a subject's assignment to one group or another be conducted through a process known as *randomization*. If, for example, a study was to assign all participants at one clinic to the treatment group, and all participants from another clinic to the placebo group, there may be underlying differences between these groups that impact the results of the study.

For many clinical trials, researchers may choose to create sub-groups before initiating randomization to help control for any potential confounding effects. This process helps ensure that both treatment and control groups are similar in composition, for example, balancing the number of men and women or subjects within particular age groups. This is known as *stratified randomization*.

Blinding

The efficiacy of many treatments can be influenced by perceptual or psychological factors of either the patient or clinician, especially in the evaluation of subjective measures such as pain or nausea. Further, the knowledge of which treatment a subject has been assigned may unintentionally bias the way a physician treats them, further complicating the study. To mitigate this, clinical trials often engage in *blinding*, in which either the subject, the clinician, or both, are unaware of the treatment assigned until the completion of the study. The case in which both the subject and the researcher are blinded is known as a *double blind* and has become standard practice for clinical trials.

Compliance and Intent-To-Treat (ITT)

Not infrequently, external circumstances or lack of compliance can alter the treatment that a subject in a clinical trial actually receives. As such, it is important that the study protocol outlines in advance exactly how such cases should be handled. Common practices utilizes what is known as *intent-to-treat* (ITT), whereby a subject is analyzed as being a member of the group to which they were assigned at randomization, regardless of the treatment that is actually received.

3.4 Biases

3.4.1 Pre-trial Bias

Pre-trial biases describe a collection of biases that can be introduced into a study before it even begins. Generally, these are associated with ensuring that our sample is representative of the population in question and that the study is designed in such a way that meaningful comparisons can be made.

Selection Bias

The first major consideration when planning a study is who you will ask to participate. When selecting your sample, you must ensure the sample is representative of the population of interest. Otherwise, your study may be prone to **selection bias**, where the study participants differ systematically from the population of interest. For example, in the first phase of a vaccine trial, the vaccine is often given to young, healthy volunteers to assess its safety profile. However, the vaccine is intended for use for both young and older people, as well as those that may have underlying medical conditions. If the safety of the vaccine was only determined based on the initial studies in healthy participants, it would look safer than it really is. This is why these initial studies are primarily used to stop development of a vaccine that is unsafe. Vaccines that are found to be safe in healthy people proceed to further testing in the general population to accurately determine the safety profile. Selection bias can be avoided by careful consideration of the make up of the population and a sampling method that accounts for various sub-populations that may differ in respect to the study outcome.

Non-response/Participant Bias

We remarked earlier that in an ideal world, all selected participants would be eager to participate in our study. However, this is not always the case. To make matters even more difficult, sometimes those that want to participate in our study differ in meaningful ways from those that do participate. This type of bias, known as **non-response or participant bias**, can occur in many studies, but is often a particular concern for studies that send out surveys or election polls. For example, consider the end-of-semester teaching evaluations that provide students an opportunity to review their instructor anonymously. When students are not required to fill out these evaluations, it is likely that only students with strong opinions about their instructor will take the time to respond. This may cause the final evaluations to be a mix of very negative ratings and very positive ratings, whereas the instructor’s actual performance maybe be somewhere in the middle. Participant bias can be avoided when designing a study by reducing the survey length and offering incentives for participation.

Confirmation Bias/Placebo Effect

To determine if a new treatment is effective, it must be compared to something else. This is addressed by splitting the sample into two groups: the **treatment group** and the **control group**. Subjects in the treatment group are given an active treatment, whereas subjects in the control

group are not treated and are used for comparison. Subjects must be assigned to receive the treatment or control at random to avoid selection bias. In addition, there is the potential for bias to arise if the subjects know whether or not they're being treated. This is known as **confirmation bias**, where the knowledge of treatment allocation can make study participants perceive the treatment benefit differently. To avoid this and ensure that any differences in the outcome are due to the treatment, researchers use a **placebo**, or inactive treatment (e.g. a sugar pill or saline infection).

Definition 3.2

Selection Bias: *A type of bias in which subgroups of the population were more likely to be included than others.*

Non-response/participant Bias: *A type of bias in which non-participants differ in a meaningful way from the participants*

Treatment group: *The part of the sample that receives the treatment*

Control Group: *The part of the sample that is not treated for comparison purposes*

Confirmation Bias: *Bias due to an individuals prior beliefs*

Placebo Effect: *A psychological phenomenon where an inactive treatment can produce a positive response*

3.4.2 During-trial Bias

Confounding

Many epidemiological studies have shown that coffee drinkers have an increased risk of lung cancer. However, upon further investigation, researchers also noticed that smokers are more likely to drink coffee. There is a known association between smoking and lung cancer, with people who smoke cigarettes being 15 to 30 times more likely to develop lung cancer than people who do not. Thus the association that was detected between coffee and lung cancer was not due to the coffee, and instead was impacted by smoking status.

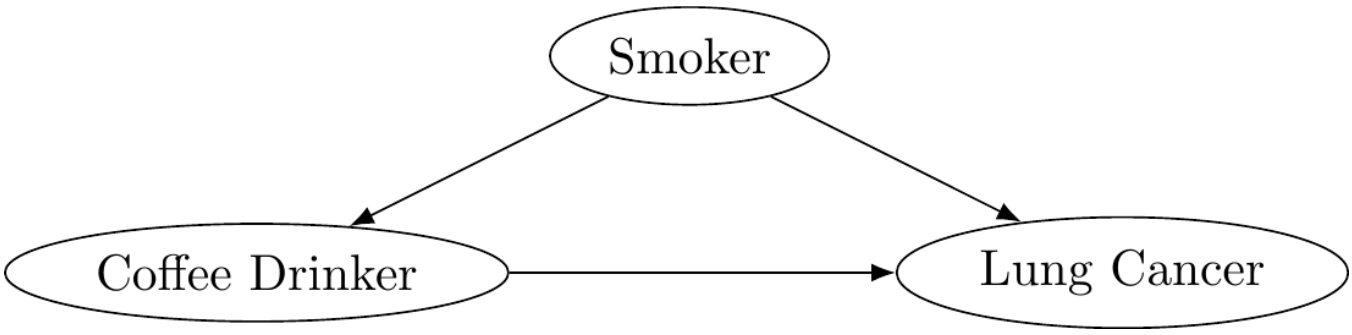


Figure 3.2: Coffee causes cancer?

Once researchers *controlled* for smoking status, they no longer found a change in lung cancer risk due to drinking coffee. Smoking status is said to be a *confounder*, also known as a *lurking variable*. A confounder is a third variable related to both exposure and outcome. Because of this relationship, confounders distort the observed relationship between exposure and outcome.

Confounding can be avoided when designing the study by ensuring the treatment and control groups have similar distributions of each confounding variable. In the coffee and cancer example, this would be analogous to ensuring there were similar proportions of smokers in the coffee drinking and non-coffee drinking groups. If it is not possible to control for a confounder when designing the study, there may be other analytical methods that can be used.

Observer Bias

Researchers often have expectations about how effective their treatment is (otherwise they probably wouldn't be studying it). If the researcher is also responsible for recording subjective assessments of the study participants, there is the potential for their beliefs about the treatment to influence either they perceive a subject's progress during follow-up.

Definition 3.3

Confounding: *A third variable related to both exposure and outcome*

Observer Bias: *Bias arising when observers record subjective data*

3.4.3 Post-trial Bias

> "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."
>
> — R. A. Fisher

Once a study has been completed, there is often very little that can be done to correct potential biases made during or before the study began. That being said, researchers still have a responsibility to present their findings in ways that can promote future research.

Compliance Bias

Despite individuals being willing to participate in a research study, this does not always mean they will comply perfectly with their assigned treatment. Consider an exercise study designed to help people lose weight. Individuals that are heavier and out of shape may tend to skip some of the workouts or drop out of the study at a higher rate than individuals that are in better shape prior to starting the study. Since compliance is directly tied to the outcome of interest (weight loss), bias may occur causing the exercise plan to look less effective than it truly is. Bias due to compliant participants differing from non-compliant participants is called **compliance bias**. Consider a 1980 study of the drug clofibrate, which is designed to reduce blood cholesterol levels to prevent coronary heart disease. Participants were randomly assigned to take up to three capsules of either clofibrate or placebo three times per day. At follow-up visits, the remaining number of capsules were counted to estimate how many each participant actually took each day. When comparing those that adhered (took at least 80% of their required capsules) to those that did not the study results were:

	# of Clofibrate Patients	Deaths
Adherers	708	15%
Non-adherers	357	25%
Total	1,103	20%

When comparing adherers and non-adherers in the clofibrate group, we see that non-adherers were much more likely to die. This seems to provide strong evidence that clofibrate is effective, however we have ignored the results of the patients in the placebo group.

	# of Clofibrate Patients	Deaths	# of Placebo Patients	Deaths
Adherers	708	15%	1,813	15%
Non-adherers	357	25%	882	28%
Total	1,103	20%	2,789	21%

When comparing patients in the clofibrate group to those that were on the placebo, we see the same trend of reduced mortality in adherers. Additionally, the death rate for adherers taking clofibrate is exactly the same as that for patients taking the placebo. This would indicate that clofibrate is not the reason for reduced mortality; instead, it is driven by characteristics of those that adhere to their medication. One possibility is that adherers are more concerned with their health in general, and are thus more likely to take better care of themselves. Compliance bias can be avoided by comparing subjects only by the groups to which they were randomized, despite their adherence to the treatment. This is also called **intent-to-treat (ITT) analysis**.

Extrapolation Bias

The previous two examples demonstrated ways in which we might incorrectly move from a target population to a non-representative sample. This final case describes movement in the opposite direction: from a specified sample to a more general population. Nonetheless, the cause of the bias is the same.

The motivation here can be most readily illustrated by considering the issue of pharmaceutical trails and the use of children. For practical, ethical, and economic reasons, clinical trials usually only involve adults – indeed, only about 25% of drugs are subjected to pediatric studies. Physicians, however, are allowed to use any FDA-approved drug in any way that they think is beneficial and are not required to inform parents if the therapy has not been tested on children.

Definition 3.4

Compliance Bias: *Bias arising when those complying to study protocol differ from those that do not comply*

Intent-to-treat (ITT) analysis: *A method of analysis which includes all participants as they were randomized despite adherence*

4 Introduction to Probability and Simulation

> “All life is an experiment. The more experiments you make, the better.”
>
> — Ralph Waldo Emerson

Learning objectives

1. Conceptual understanding of randomness and simulation
2. Learn the scientific definition of probability
3. Methods for calculating probabilities

4.1 Randomness and Simulation

In this section, we seek to understand randomness and how we can use the computer to quickly simulate an outcome. To do this, we will start with a dice game in which two dies are rolled. Rolling dice is an example of a **random process**, because we cannot predict the outcome with certainty. By contrast, a process that is not random is known as **deterministic**, as the outcome is determined beforehand. An example of a deterministic process would be computing the area of a triangle from its base and height - the area is always $\frac{1}{2} * base * height$.

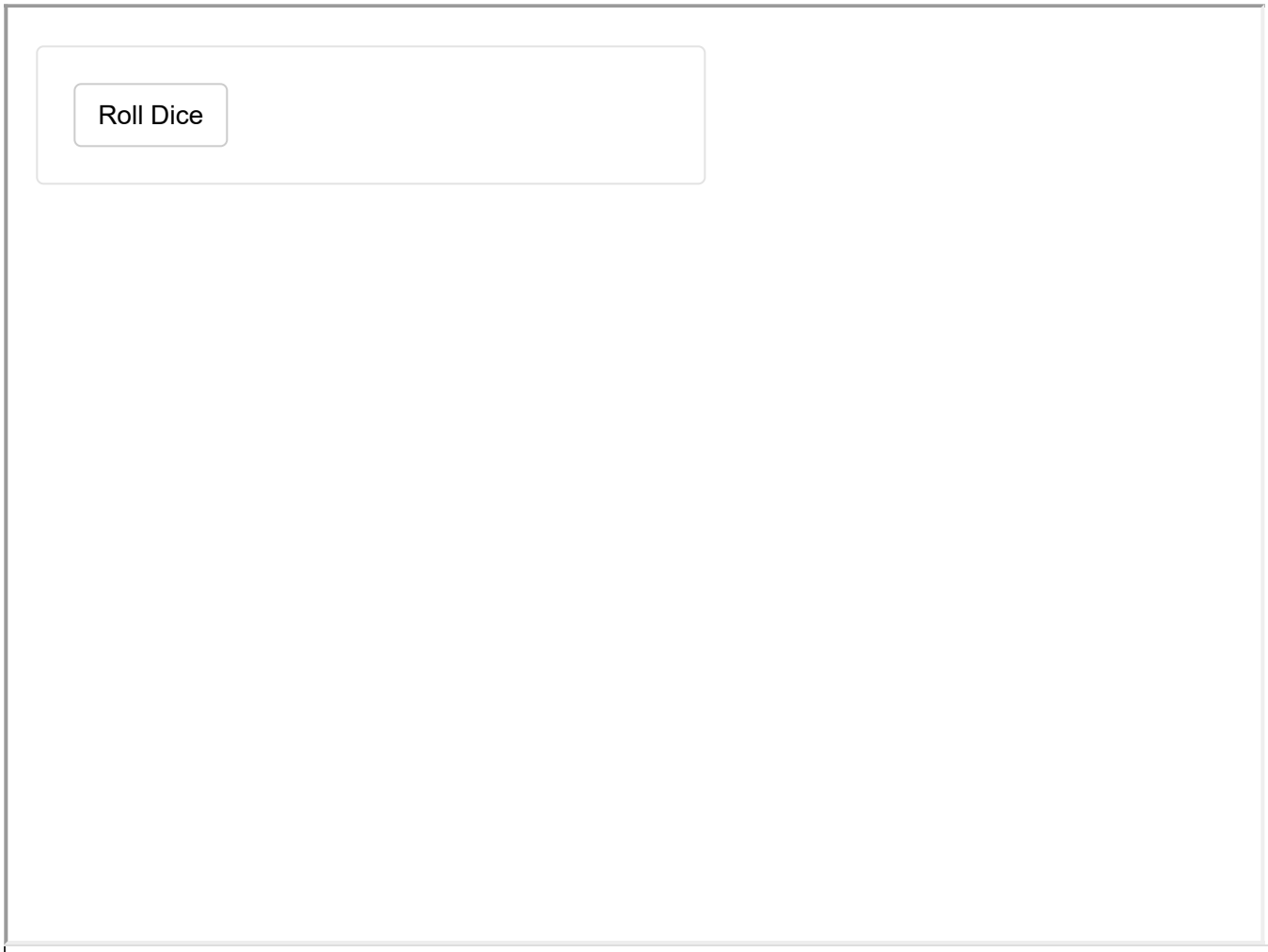
In our dice game, we are interested in the sum total of the roll. For example, if a 2 and a 6 are rolled, the sum total is 8. Each die has 6 outcomes, so there are 36 possible ways to roll the two die:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Exercise 4.1

There are 11 possible sum totals you can achieve when rolling two six-sided dies, {2, 3, 4, …, 11, 12}. In this game, you win if your roll adds up to 7.

1. Out of the 36 possible rolls listed above, how many rolls result in a win? What does this tell you about the chance of winning the game?
2. Using the applet, play the game 30 times and fill out the table below with whether you won and rolled a 7 (W) or lost and rolled any other total (L).



Game	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Result															
Game	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Result															

3. What proportion of times did you win the game? Was this close to what you expected? Why or why not?

While this was a relatively simple dice game, it illustrates several important concepts. Going into the game, we are able to determine how likely it is that we win. We did this by first enumerating all possible outcomes of two dies being rolled. Any such collection of all possible outcomes of a random process is called the **sample space**. Then, we considered the ways we could observe the **event** that the two rolls result in a sum total of 7. In general, an event is an outcome or collection of outcomes from a random process and will always be contained in the sample space.

Another important observation is that we have not physically rolled any dice ourselves. Instead, we have used the computer to perform a **simulation**. Simulations are a powerful tool which allow us to quickly and consistently repeat a random process. There are two components to simulation:

1. The conditions and behavior of the experiment are determined in advance through the **simulation parameters**.
2. The experiment is performed with the aid of a computer.

In the dice rolling simulation, the parameters specified that each side of the die were equally likely and that two dies were rolled each time the game was played. These parameters were specified internally, such that the user cannot alter the specifications. Later in this chapter, we will work with simulations where the parameters can be changed. Lastly, the use of the computer may seem trivial, but it has two important implications:we were able to exactly specify the simulation parameters, and we are able to repeat *this exact same experiment* knowing all conditions will remain exactly the same.

Definition 4.1

Random Process: *An act or process that results in an outcome that cannot be predicted with certainty*

Deterministic Process: *An act or process that results in an outcome that is not random*

Sample Space: *The set of all possible outcomes from a random process*

Event: *An outcome or collection of outcomes from a random process*

Simulation: *A tool to replicate random processes an arbitrary number of times*

Simulation Parameters: *Values which change the behavior of the simulation*

4.2 Probability

Statistical inference is founded in the ability to quantify uncertainty. Probability is the mechanism that allows us to do so, by telling us how likely something is to happen. People talk loosely about probability all the time. For example, “What’s the chance of rain tomorrow?” or “How likely is it

that drug A is better than drug B?" In order for probability to be used for statistical inference, we must be precise about our definition of probability.

We would all agree that the probability of heads when flipping a fair coin is 50% and the probability of rolling a 2 on a 6-sided die is 1/6, but why is that true? Well, if we were to flip a coin many, many times, we would expect half of the flips to result in heads. Similarly, if we roll a 6-sided die over and over, 1/6 of the rolls should result in a value of 2. The big idea is that when we talk about probability, we are thinking about a *long-run frequency* or what will happen if the random process is repeated over and over again under the same conditions. If we think about probabilities in terms of the long-run frequencies, we can define and quantify **probability** as the fraction of time an event occurs if a random process is repeated indefinitely. This means that probabilities are always between 0 and 1, since we can never observed more or less events than the number of times the process is repeated, e.g. we can never observed 12 heads on 10 coin flips.

This leads us to some important properties of probabilities:

Properties of Probability

1. The sum of probabilities for all outcomes in the sample space, $\mathcal{S}$, must equal 1
2. For any event, the probability of that event is the sum of the probabilities for all the outcomes in that event

Consider flipping a fair coin three times. Each act of flipping the coin is random process - the coin might land on heads and it might land on tails. Letting H be shorthand for the flip resulting in heads and T be shorthand for the flip resulting in tails, the sample space can be enumerated as

$$\mathcal{S} = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\},$$

giving us eight possible results from the three coin flips. Since each outcome is equally likely (why?) The first property tells us that the probability of any specific outcome (say, HHH) is 1/8. The second tells us that the probability of heads on the first toss is $4/8 = 1/2$, since four of the eight possible outcomes have heads on the first toss ($\{HHH, HHT, HTH, HTT\}$). These properties underlie a lot of the more complicated formulas and concepts we will cover in this chapter, although we don't always think about them explicitly.

Exercise 4.2

We will now investigate the long-run frequency definition of probability using an applet which simulates coin flipping. Internally, the simulation parameters specify that each coin flip has a 50% chance of heads. As a user, you are able to determine how many coin flips you would like to perform. The simulation results are summarized in three figures:

1. (Top right) the flip results are shown with a blue “T” indicating the flip resulted in tails and a pink “H” indicating the flip resulted in heads.
2. (Bottom left) the total number of heads and tails across all flips is tallied.
3. (Bottom right) the running total of the proportion of heads is plotted. For example, if we flip a coin three times and observe THH, the running proportion of heads is 0/1=0 after the first flip (T), 1/2=0.5 after the second flip (TH), and 2/3=0.67 after the third flip (THH). The dotted red line on this plot falls at 0.5, which translates to half of the flips resulting in heads.

Number of coins to flip:

3

Flip Coins

1. Set the number of flips to 10 and click the “Flip Coins” button 15 times. Each time you perform an experiment, record the number of heads you observed in the following table:

Trial	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# Heads															

   i. On any given experiment where the coin is flipped 10 times, how many flips would you expect to result in heads? Why?
   ii. How many times did you see *exactly* four heads? What about at least two heads?

2. Now set the number of flips to 100 and click the “Flip Coins” button.
   i. What happens to the running proportion of heads as the coin is continually flipped?
   ii. Focusing your attention on the plot of the running proportion of heads, click the “Flip Coins” button several times. What do you notice about the plot? What characteristics of the plot stay the same and what differ?
3. For each of the following number of coin flips, perform one experiment and record the final *proportion* of heads you observed (e.g., 12 heads out of 20 flips = 12/20 =0.6).

3:	5:	10:
20:	100:	500:
1,000:	5,000:	10,000:

   i. What happens to the proportion of heads you observed as the number of flips increased? How does this relate to the concept of long-run frequency?
   ii. How does the final proportion of heads observed relate to the probability of observing heads?

Definition 4.2

Probability: *The fraction of times an event occurs when a random process is repeated indefinitely*

4.3 Methods for Computing Probabilities

There are several methods that can be used to compute probabilities. We will introduce these methods in the context of coin flipping and seeking to answer the question: if a coin is flipped three times, what is the probability that exactly two of the coins will be heads?

1. **Enumeration Method:** The enumeration method proceeds by listing all of the possible outcomes of the experiment and counting the total number of ways the event can be observed. Then, the probability is calculated as:

$$\text{Probability} = \frac{\text{Number of ways event can occur}}{\text{Total number of outcomes}}$$

In our coin flipping example, we know the sample space includes eight outcomes:

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, TTH, THT, TTT\}$$

Next, we count how many of those have exactly two heads:

$$\mathcal{S} = \{HHH, \color{red}{HHT}, \color{red}{HTH}, \color{red}{THH}, TTH, THT, HTT, TTT\}$$

Dividing the events of interest by the total number of outcomes, we see that the probability of getting exactly two heads is $P(\#\text{ Heads} = 2) = \frac{\#\text{ Heads} = 2}{\#\text{ Possible Outcomes}} = \frac{3}{8}$.

That is, by enumerating the eight possible outcomes, we identified three of them in which the number of heads was two.

2. **Probability function method:** Often, a random process has an associated *probability function* that allows us to determine the probability of a set of outcomes. In this case, our coin flipping example follows what is known as a *binomial distribution*, which has the probability function:

$$P(\#\text{ Heads} = \text{k}) = f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

For our experiment, the probability of heads is $p = 0.5$, the total number of flips is $n = 3$, and for our event, $k = 2$ heads. Substituting these numbers, we have"

$$
\begin{aligned}
P(\#\text{ Heads} = 2) &= \binom{3}{2}(0.5)^2(0.5)^{3-2} \\
&= 3 \times (0.5)(0.5)(0.5) \\
&= 0.375.
\end{aligned}
$$

We see, then, that a probability function is a function that takes an event as an argument and returns an associated probability. Just as in the enumeration method, we find that $P(\#\text{ Heads} = 2) = 0.375 = \frac{3}{8}$.

3. **Simulation method:** The previous methods are valid, but can quickly become impractical or even impossible once the problem grows becomes more complex. Even for an example as simple as coin flipping, the enumeration method can quickly become infeasible. As the number of flips increases, the size of the sample space increases exponentially, making it challenging (or lengthy) to count all possible outcomes and all the ways the event can occur. If we seek to use the probability function method, we must know what the function is; for complex processes, these functions can be nearly impossible to construct, leaving statisticians and researchers with no clear way to specify the probability of events.

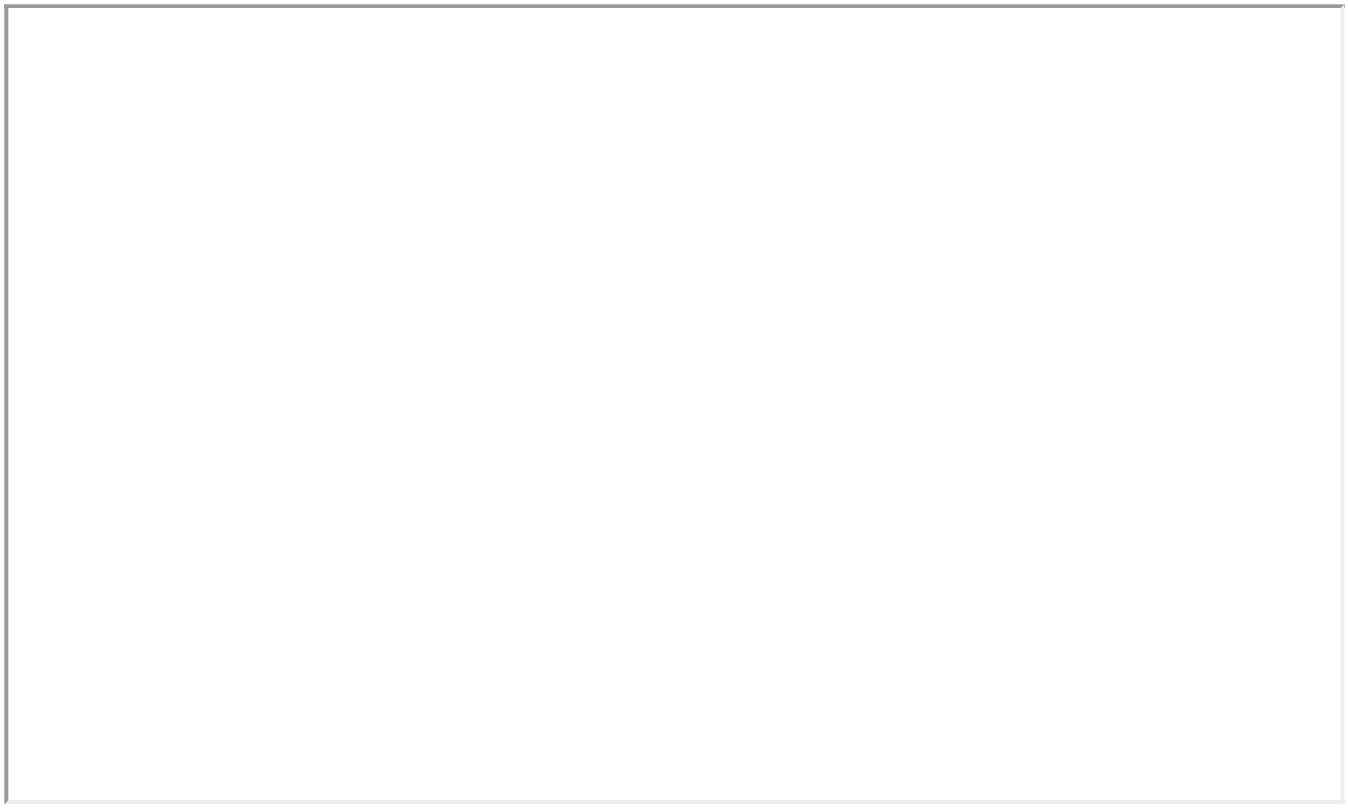
   The simulation method allows us to mitigate both problems by using the computer to repeatedly perform an experiment (say 10,000 times), and computing the proportion of experiments in which the event was observed. We have previously defined probability as the fraction of times the event occurs if the random process is repeated many times. This is exactly what the simulation method does.

   Mathematically, this looks like

$$
P(\#\text{ Heads} = 2) \approx \frac{\text{Number of times } \{\#\text{ Heads} = 2\}}{\#\text{ Experiments}}
$$

   A careful reader might note that for a given random experiment (such as flipping a coin three times) and a specified event (exactly two heads), there is only a single* correct answer to the question, “What is the probability of getting exactly two heads when flipping a coin three times?” and this single correct answer is precisely what was found using the enumeration and probability function methods described above. contrast, The simulation method provides an approximation to the probability, as opposed to an exact calculation. However, as the number of repeated experiments increases, the simulation method will give a result closer and closer to the true answer.

Exercise 4.3



Definition 4.3

Enumeration method: *A method to exactly compute probabilities by dividing the number of ways an event can occur by the total number of possible outcomes*

Probability function method: *A method to exactly compute probabilities by using a known function*

Simulation method: *A method to approximately compute probabilities by repeatedly simulating experiments and taking the proportion of simulations where the event occurs*

5 Probability Distributions

> "Statistics is the grammar of science."
>
> — Karl Pearson

Learning Objectives

1. Learn the definitions of random variables, probability distributions, and expected values and the connections between them.
2. Understand how the parameters of a distribution govern the shape of the distribution
3. Relate probability distributions to data generated according to a distribution
4. Learn about the binomial and normal distributions, their parameters, and how their distribution shape is related to those parameters.

5.1 Introduction to Probability Distributions

In Chapter 4, we introduced the idea of random processes, i.e. situations in which the outcome can not be determined perfectly in advance. Random processes are defined in terms of the *collection of possible events* (sample space) and their *associated probabilities*. In that chapter, we saw three methods for calculating probabilities - the enumeration method, the probability function method, and the simulation method. In this chapter we will expand on the probability function method, which uses a known function called a **probability distribution** to determine the probability of each event.

Probability distributions are closely related to **random variables**, a numeric variable that can take on different values depending on the outcome of a random process. In previous mathematics courses you may have seen variables such as x or y used as placeholder values which are then solved for. For example, you can solve for the variable x in $4x + 5 = 25$, to determine $x = 5$. By contrast, the outcome of a *random* variable cannot be predetermined. Instead, we talk probabilistically about the likelihood of observing each possible outcome. Random variables are typically denoted with capital letters, e.g., X or Y, whereas the observed outcome of the random

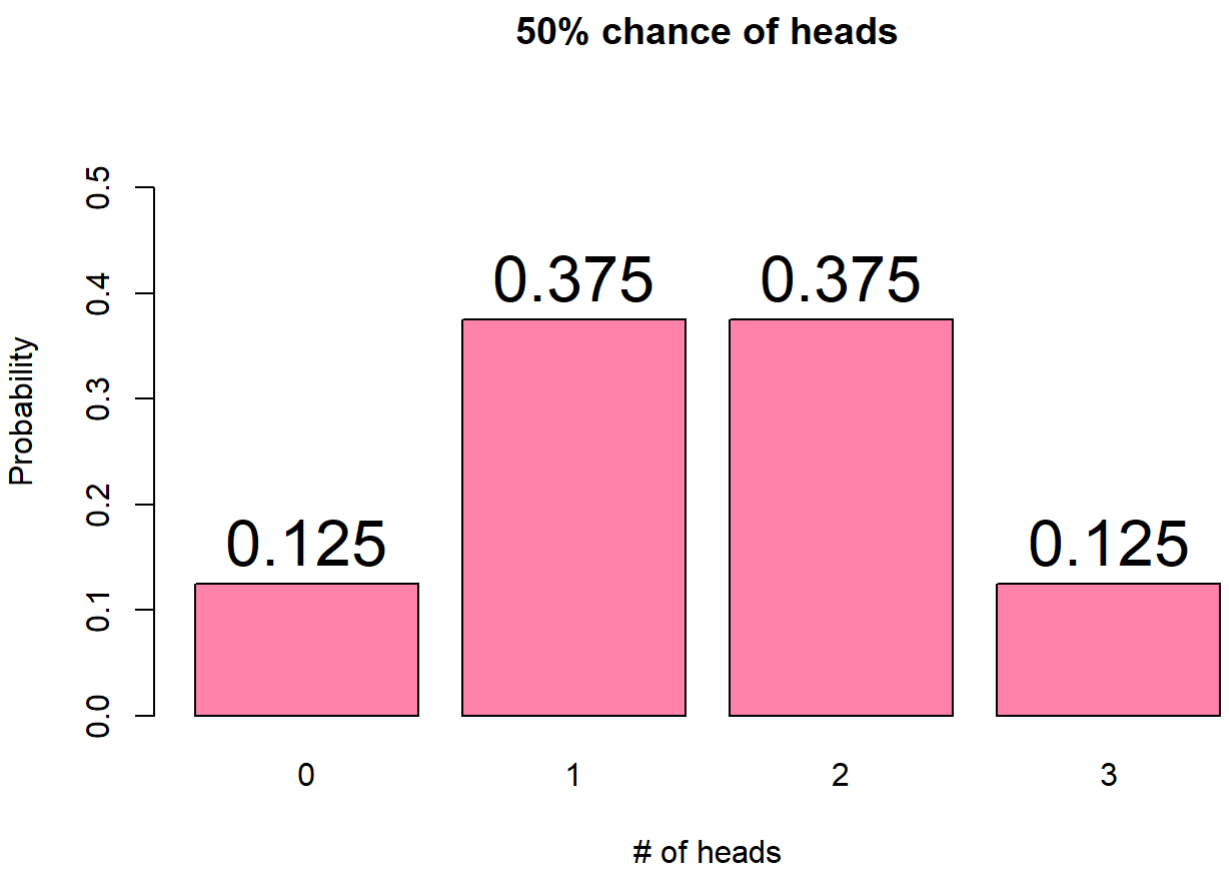
process is denoted with lowercase letters x or y. For example, flipping a coin three times is a random process. We can define the random variable X to represent the number of heads we observe between the three flips. X can take on four possible values: 0, 1, 2, or 3. If we observe 2 heads, we have $x = 2$.

Most simply, a probability distribution (often just called a distribution) is a method for taking a possible event as input, and giving us the corresponding probability as output; the corresponding probability tells us how likely it is that the specific event will occur, out of all of the possible events. We often say random variables have or follow a probability distribution, as the distribution quantifies the probability of observing the possible values a random variable can take on. We can denote a probability distribution as $P(X = x)$, or the probability that the random variable, X, takes on a generic value, x.

There are many useful probability distributions that have been defined by mathematicians and statisticians to describe a variety of scenarios:

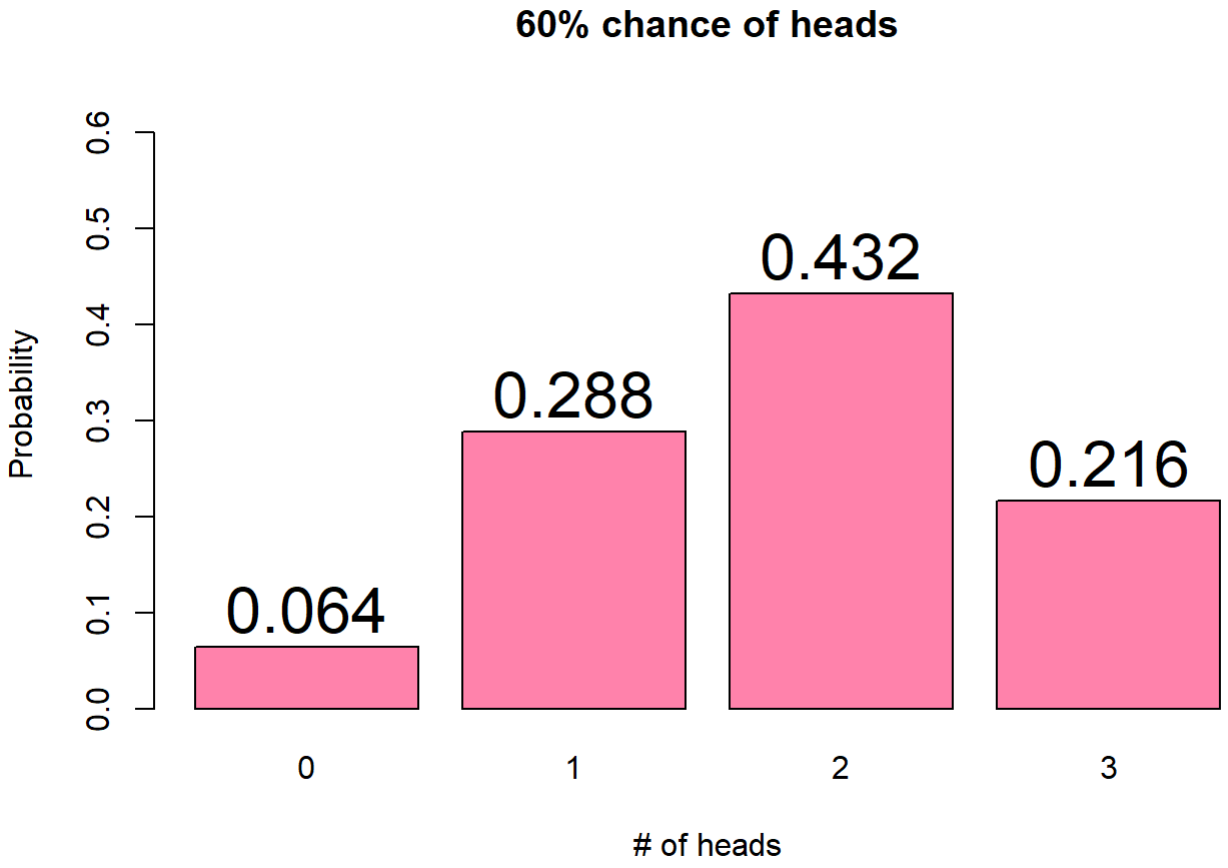
- Counting the number of successes in a fixed number of trials that can results in either success or failure, such as counting the number of heads when flipping a coin three times
- Counting the number of successes before the first failure in a series of success/failure trials, such as counting the number of days before a lightbulb burns out
- Describing the length of time between events that occur at a constant rate, such as the time between phone calls at a help desk
- Describing range of continues values, such as the blood pressure in adults

One can represent a probability distribution visually using a **probability histogram**. On the x-axis, we have the possible outcomes of the random process – the values the random variable could take. For each outcome, the bar height represents the probability of observing that value. For the coin flipping example, the probabilities of observing each possible number of heads can be represented as:



When looking at a probability histogram, we can characterize the shape of the distribution analogously to how we talk about a histogram of observed data. We often use *unimodal* distributions, which may or may not be *skewed*. For example, the above probability histogram shows a unimodal and symmetric distribution.

The beauty of using probability distributions to describe the likelihood of all outcomes of a random process is its simplicity. Probability distributions rely on a small number of **parameters** which determine the distribution's shape. In the coin flipping example, we could consider one of the parameters to be the probability of obtaining heads. We assume that we have a fair coin and that this probability is 50%. If, instead, we had a weighted coin with a 60% chance of landing on heads, the probability distribution would change. That is, the probability associated with each possible outcome would be different.



With a higher chance of the coin resulting in heads, we now see that the three flips are more likely to result in 2 or 3 heads and much less likely to result in 0 heads. In other words, the distribution is now slightly skewed left. Because these probabilities can be described by a distribution function, we can
easily compute and compare the probabilities of each possible outcome as a function of the distribution parameter, in this case, the probability of flipping heads.

Another key concept related to probability distributions and random variables is the idea of the **expected value**, often denoted $E(X)$. The expected value of a random variable, often denoted as $E(X)$, is a weighted average which provides a measure of the central mass of the probability distribution. The expected value averages over all possible outcomes of the random variable with each outcome weighted according to its probability.

Most simply, we can think of the expected value of a random variable as its average value. In more technical terms, we might say that the expected value is a weighted average which provides a measure of the central mass of the probability distribution, averaging over all possible outcomes of the random variable with each outcome weighted according to its probability. Whew!

Returning to the coin flipping example with a fair coin, we have seen that the probability distribution is:

x	P(X = x)
0	0.125
1	0.375
2	0.375
3	0.125

Multiplying each possible outcome (x) by its associated probability ($P(X = x)$) and adding them together (which happens to be the definition of a weighted average) gives us an expected value of

$$(0 \times 0.125) + (1 \times 0.375) + (2 \times 0.375) + (3 \times 0.125) = 1.5.$$

Looking at the probability histogram, this value should make sense as it falls right in the center of the distribution.

Expected values are more easily conceptualized in terms of a game or bet. For example, consider the following game: you flip a fair coin; if the coin lands on heads, you win $20, and if the coin lands on tails, you lose $1. Would you play this game? Assuming you have a spare dollar, the answer is probably yes. Since you have equal chances of the coin landing on heads or tails, you are just as likely to win $20 as you are to lose $1. In terms of the expected value, it would be

$$(20 \times 0.5) + (-1 \times 0.5) = 9.5.$$

The expected value tells us that if you were to play this game over and over, you would be expected to win $9.5 per game. Critically, it is worth reiterating here that *for any single instance of this game, you can only win $20 or lose $1*. The $9.50 indicates the *long term average over many games*.

Finally, let's discuss how probability distributions relate back to our general statistical framework. In our framework, we are interested in learning about the *population*, which we assume follows some probability distribution. This distribution is governed by a set of *unknown parameters*. When we collect a sample, we expect that the distribution of the sample will be similar to that of the true

population. As we will see in the following chapters, our sample will be used to compute *statistics*, or estimates of the true parameters. Indeed, much of statistical inference involves estimating the values of these parameters, with associated measures of certainty.

We end this chapter with a discussion of two of the most commonly used distributions, the binomial and the normal.

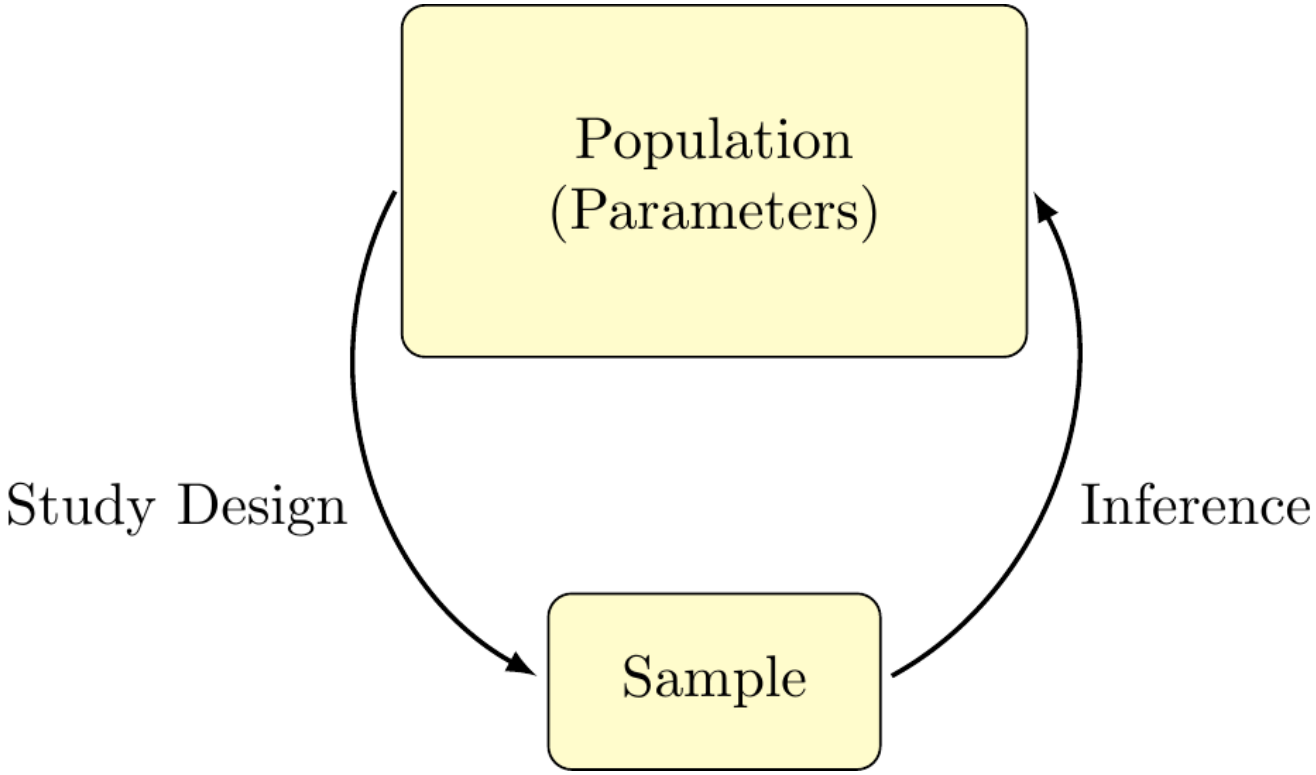


Figure 5.1: Statistical Framework (again)

Definition 5.1

Probability distribution: *A method for assigning probabilities to all possible events*

Random variable: *A numeric variable whose value depends on the outcome of a random process*

Probability histogram: *A graphical display of a probability distribution*

Parameters: *Values associated with a probability distribution that determine the distributions shape*

Expected value: *The weighted average of the outcomes of a random variable, with weights determined by their probability*

Statistics: *Values computed from a sample that serve as estimates of the population parameters.*

5.2 Binomial Distribution

The first distribution we will examine in depth is the **binomial distribution**, which describes the number of successes in a fixed number of independent trials that can result in one of two outcomes (success or failure), where the probability of success is the same between trials. Notably, as the data from a binomial distribution fall into distinct categories, the binomial distribution is also a member of a category of distributions known as **discrete distributions**.

We have already seen one example of the binomial distribution in detail – flipping a coin some number of times and counting the number of flips result in heads (success). Each flip has two possible outcomes (heads or tails), the same probability of heads on each flip (50%), and a predetermined the number of trials (three flips).

To illustrate another examplea of the binomial distribution, we might consider rolling a six-sided die a number of times, counting the number of rolls that result in either a 5 or 6. That is, we would consider a success to be rolling a 5 or 6, and a failure to be rolling a 1, 2, 3, or 5. As there are six equally likely outcomes, with two of these being considered a success, we note that for any particular roll (trial), the probability of success is 2/6, or 1/3. The takeaway here is that, even though there are six possible values for a given roll, we can still posit a binary outcome and determine the associated probabilities.

As you may have noticed from these two examples, the binomial distribution can be used for various success probabilities and numbers of trials. In fact, these quantities define the two parameters of the binomial distribution. These parameters are typically denoted as n = the number of trials and p = the probability of success. The binomial distribution can be written as a function of these parameters

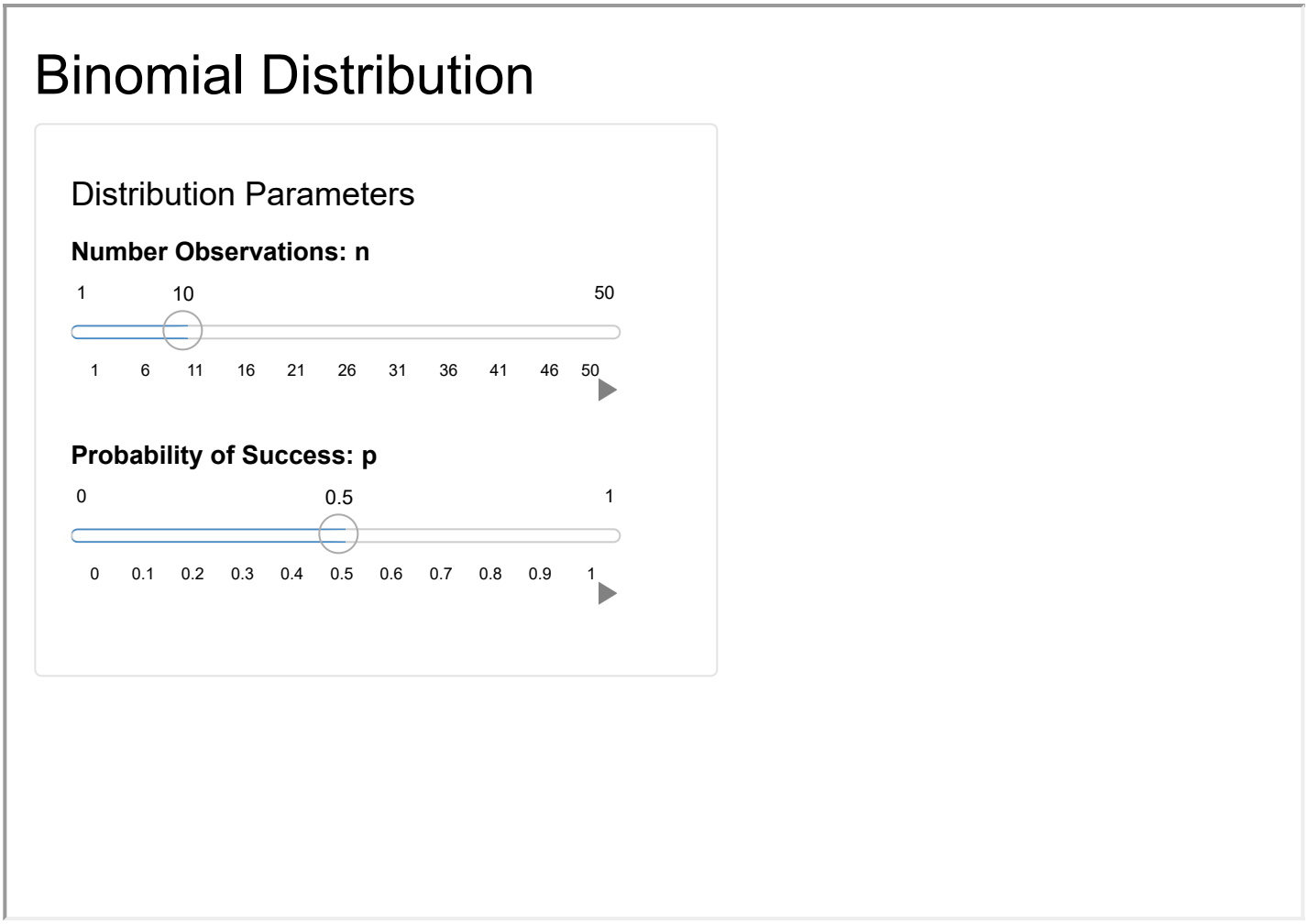
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

While this may look like a nasty formula, don't be afraid! Probabilities following a binomial distribution can be easily computed by any statistical software. For this reason, we will focus on the distribution properties rather than performing calculations.

For any valid value of n and p, we can use the binomial distribution to compute the probability of observing any possible outcome. Valid values of n and p simply mean that the number of trials conducted must be a positive integer (e.g., 1, 2, 3, …) and that p must be between 0 and 1 (it is a probability after all). The expected value of the binomial distribution is given by $E(X) = n \times p$, or the total number of trials times the probability of success for each trial. To gain a better understanding of how these parameters impact the shape of the distribution, we will use the following applet.

Exercise 5.1

The applet below is designed to help you get familiar with the parameters of the binomial distribution, and how they impact the probability distribution. You can change the values of the number of trials n, and the probability of success p, and the app will display the associated distribution in a probability histogram.



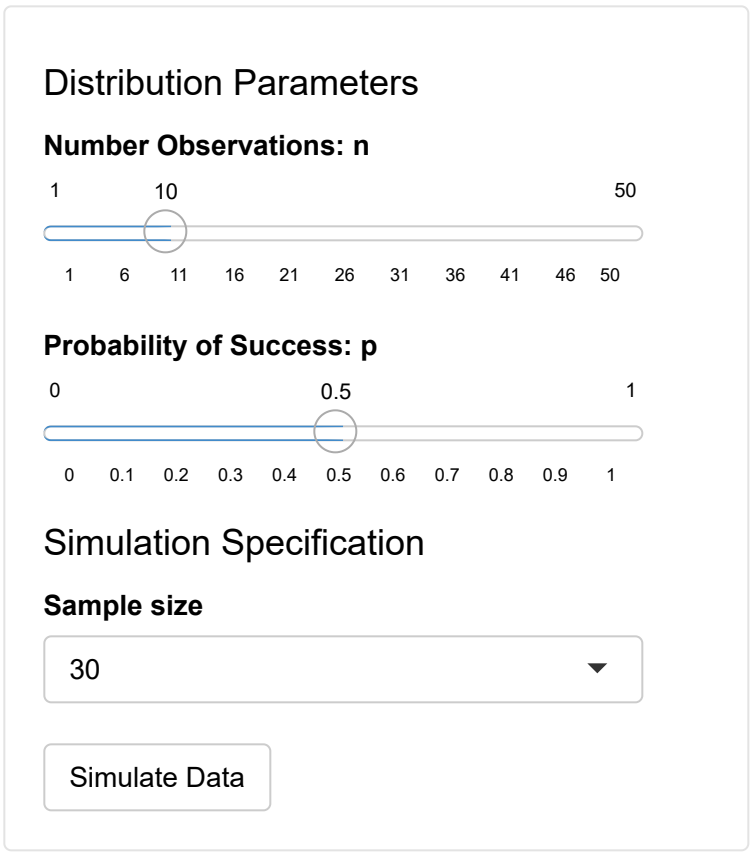
Use the applet to answer the following questions:

1. Set $n = 10$, and change the value of p (note: you can press the triangular “play” button to have the app vary p automatically). What happens to the shape of the distribution as p gets closer to 0? What about when p gets closer to 1?
2. Now, set $p = 0.4$ and vary n over the range of possible inputs. What do you notice about the x-axis as n is changing? Explain this trend by referring back to what n represents.
3. Keeping p constant and varying n between 20 - 50, does the shape of the distribution change? What about the location of the distribution?

Exercise 5.2

The applet below is designed to familiarize you with data generated from the binomial distribution. You can change the values of the number of trials n, and the probability of success p to specify the parameters of the population distribution. Then, you can take a random sample from the distribution.

Binomial Distribution



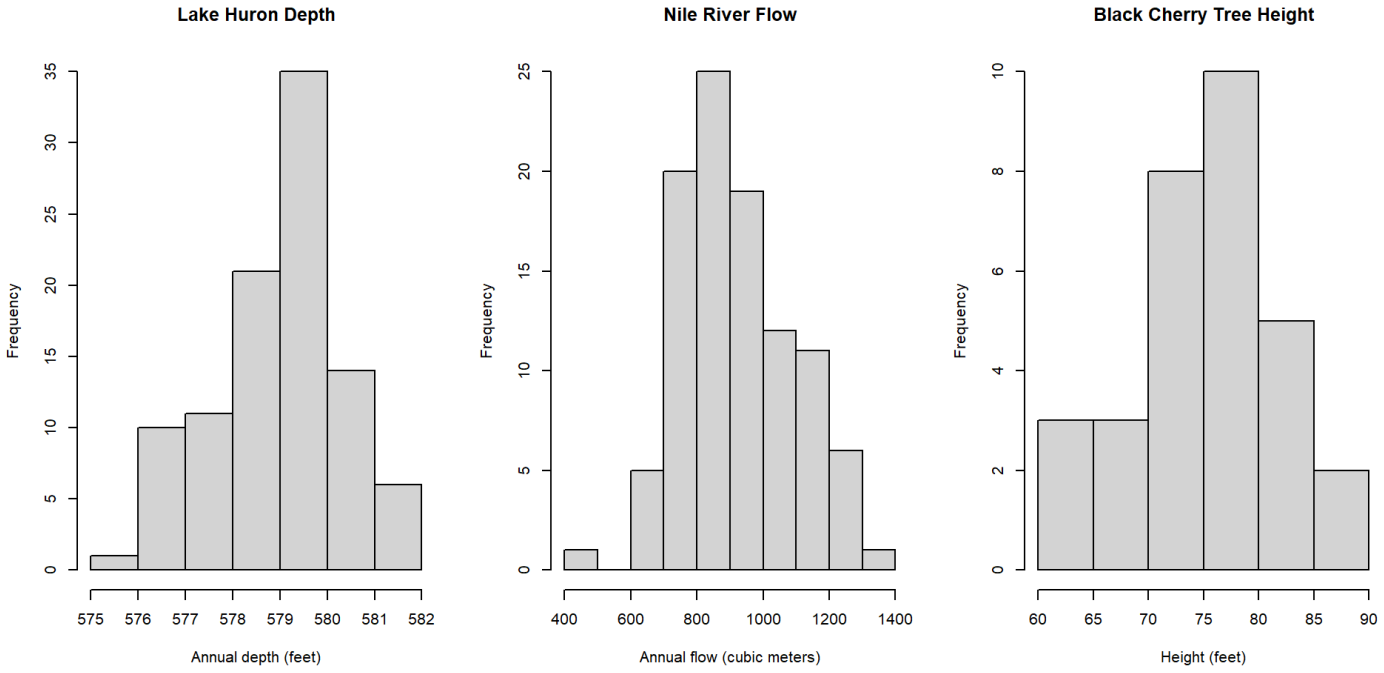
Definition 5.2

Binomial Distribution: *A probability distribution that characterizes the probabilities of observing some number of successes in a fixed number of trials, each with two possible outcomes and the same probability of success*

Discrete Distribution: *Any probability distribution that depicts the occurrence of distinct, countable values*

5.3 Normal Distribution

Let's begin by considering three histograms, each describing a set of continuous data, collected from various sources. First, we have the annual depth of Lake Huron from 1875-1972 in feet. Next, the annual flow of the river Nile from 1871-1970 measured in cubic meters. Lastly, the recorded height, in feet, of 31 black cherry trees. What do all of these histograms seem to have in common?

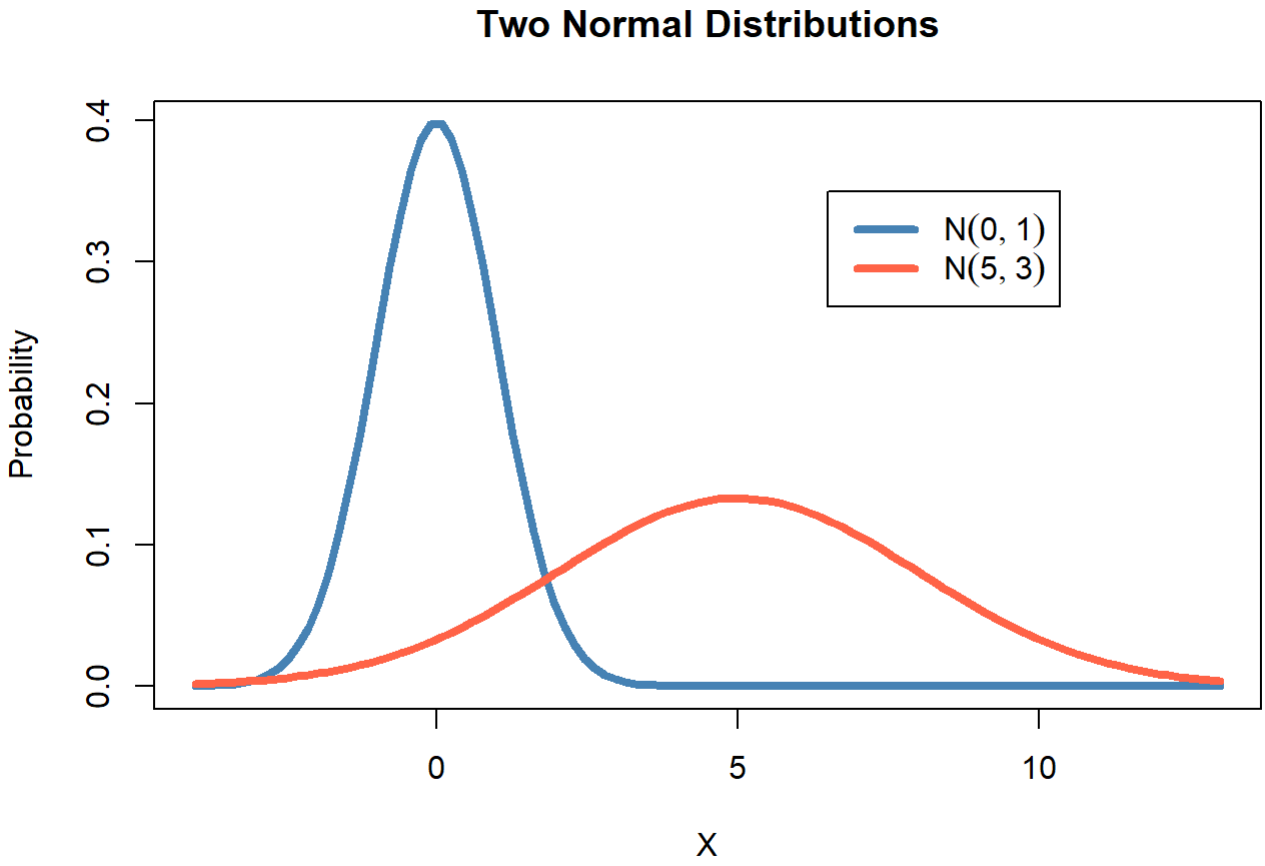


What we see here are examples of a **normal distribution** (also known as a bell curve), one of the most ubiquitous distributions in all of statistics. The normal distribution is characterized by the "bell shape" that is symmetric about it's center.

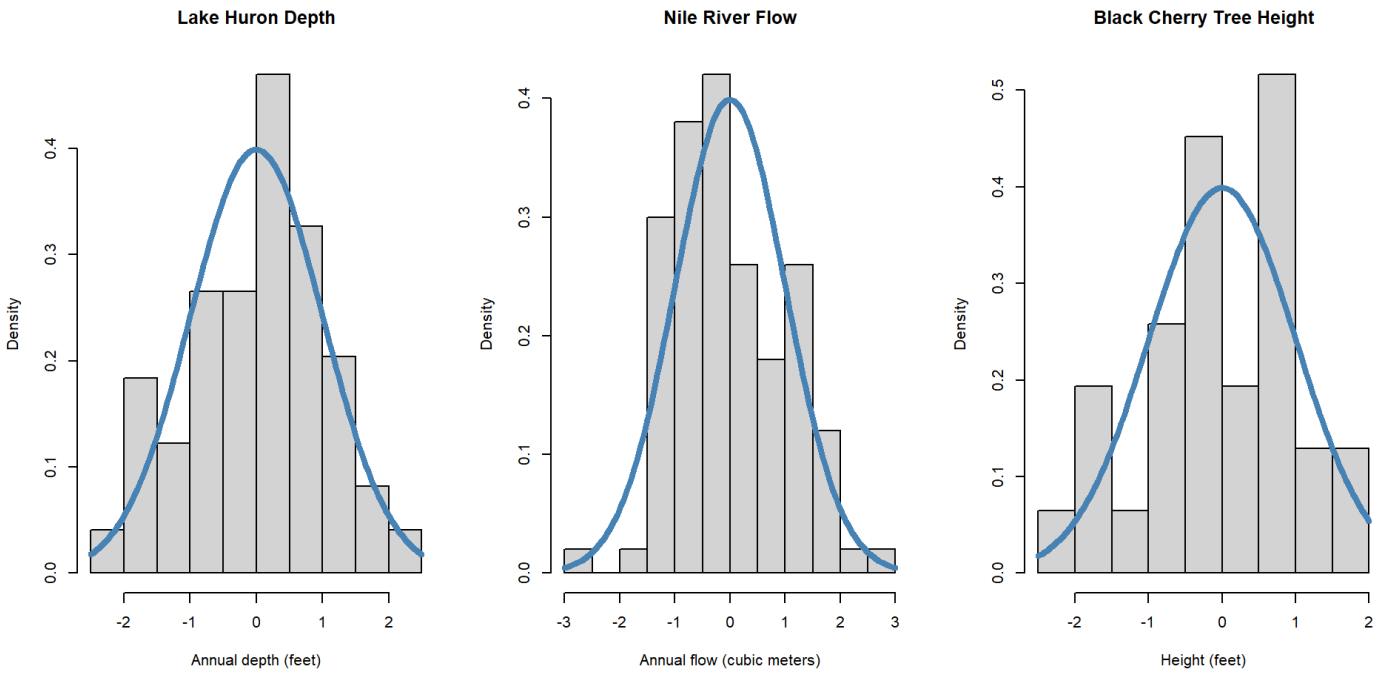
Like the binomial, the normal distribution is characterized by two parameters, μ and σ^2, representing the mean and the variance, respectively. The mean value, μ, indicates the location of the peak on the x-axis, whereas the variance, σ^2, indicates the amount of dispersion about the mean. A random variable X that follows a normal distribution can be expressed $X \sim N(\mu, \sigma^2)$, or, "The random variable X follows a normal distribution with mean μ and variance σ^2." The formula for the normal distribution is given as

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

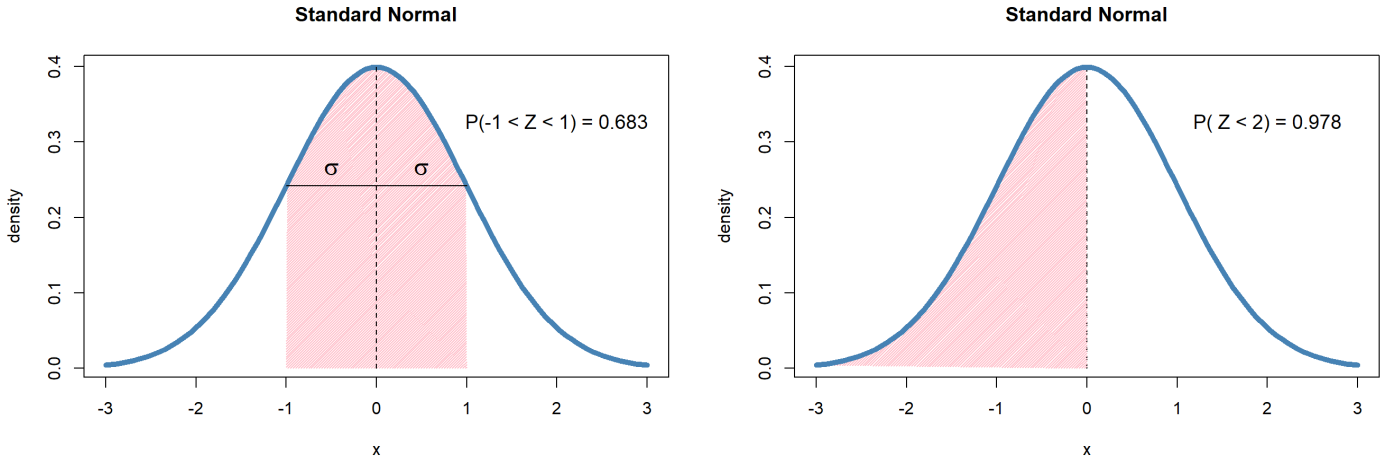
Consider the two normal distributions below, with different values for μ and σ^2. Although they are centered at different locations and have different amounts of dispersion around the mean, they are both bell-shaped curves, characteristic of the normal distribution:



Given that the normal distribution appears so frequently in statistics, it is common practice to *standardize* a normal distribution so that it has a mean value of $\mu = 0$ and variance $\sigma^2 = 1$ (note also that the standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{1} = 1$). A normal random variable that has been standardized is called a **standard normal distribution** and is often written $Z \sim N(0, 1)$. We can consider again the histograms above, once they've been standardized:



Unlike the binomial distribution, in which there are n possible values that our random variable can take, the normal distribution represents a random variable that is **continuous** over a range of values, making the normal distribution a member of the family of **continuous distribtuions**. Instead of asking the probability of a specific value, say, $Z = 0$, probabilities are given as the area under the curve for a certain interval. We might ask, "What is the probability that Z is one standard deviation (σ) away from 0?" or perhaps, "What is the probability that $Z < 0$?"

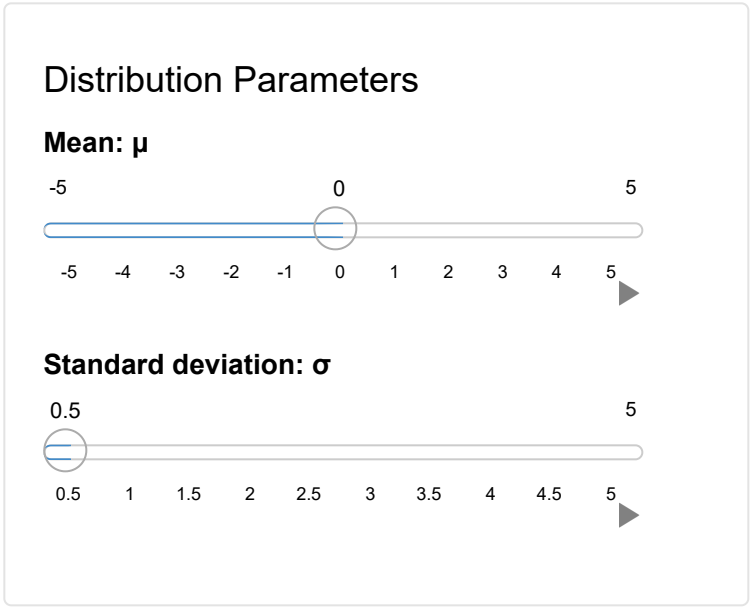


We will return to the normal distribution in the following chapters, demonstrating both how it arises naturally in the study of statistics and the powerful toolkit it provides in the field of statistical inference.

Exercise 5.3

The applet below is designed to help you get familiar with the parameters of the normal distribution, and how they impact the probability distribution. You can change the values of the mean μ, and the standard deviation σ, and the app will display the associated distribution in a probability histogram.

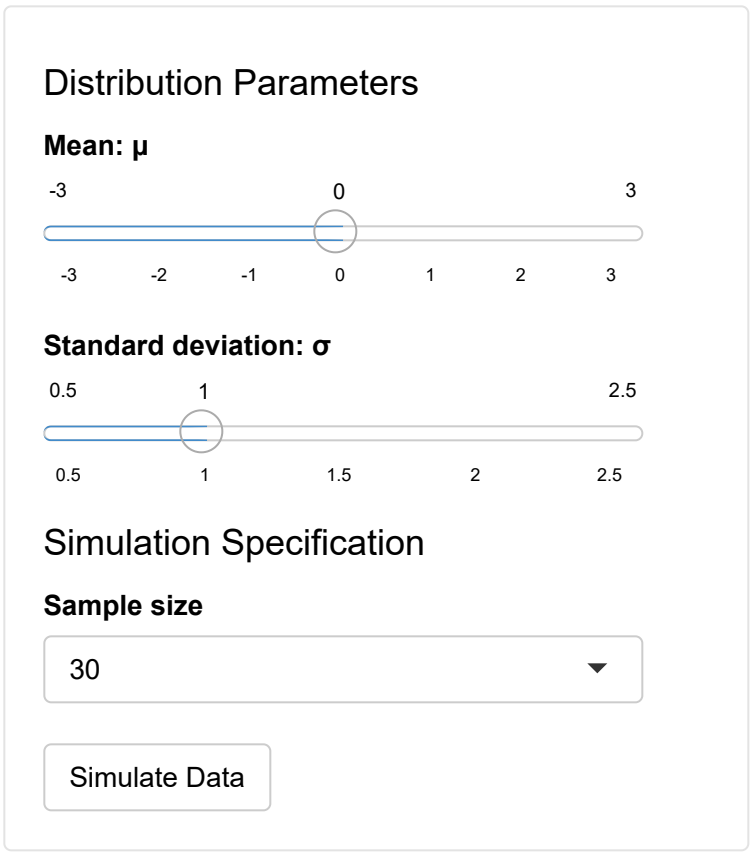
Normal Distribution



Exercise 5.4

The applet below is designed to familiarize you with data generated from the normal distribution. You can change the values of the mean μ, and the standard deviation σ to specify the parameters of the population distribution. Then, you can take a random sample from the distribution.

Normal Distribution



Definition 5.3

Normal Distribution: *A continuous bell-shaped distribution with two parameters that are the mean value, μ, with a variance σ^2*

Standard Normal Distribution: *A special case of the normal distribution, $Z \sim N(0, 1)$*

Continuous Distribution: *Any probability distribution that depicts the occurrence of a random variable that can take an infinite set of possible values within a given (potentially infinite) interval*

6 Sampling Distributions and the Central Limit Theorem

> “While nothing is more uncertain than a single life, nothing is more certain than the average duration of a thousand lives.” - Elizur Wright

> “Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.” - Albert Einstein

Learning objectives

1. Learn to differentiate between statistics and parameters
2. Understand the concept of sampling distributions
3. Conceptual understanding of Central Limit Theorem and how it applies to sample means

6.1 Introduction to Sampling

Suppose policy makers and public health experts in Iowa are concerned with Iowans’ fast food intake. To understand this further, officials want to determine the average monthly spending on fast food for all Iowans. It would be very expensive, however, to have all 3.2 million Iowans track and report their total monthly fast food expenses. Whats more, many Iowans would not be willing to share personal financial details with public health researchers or government officials. These are but a few of the challenges making it practically impossible for officials to ever determine the true average spent by Iowans each month on fast food.

A true numeric quantity about the population, here assumed to be the mean, is known as a **population parameter**. While we may never know the exact value of a population parameter, we have tools at our disposal to estimate it. For example, officials can determine a *representative sample* of Iowans and have those consenting to share their information report their monthly fast food spending. The average spending of this group is the *sample mean*, which can then be used

approximate the population mean. The sample mean is an example of a **sample statistic**, which is a numerical quantity about the sample. In other words, statistics are what investigators know, and parameters are what investigators want to know.

The statistical framework (pictured below) demonstrates the process of making **inference** about population parameters with the use of sample statistics, as was done in the example above. The numerical quantities of interest can be many different things - means, proportions, standard deviations, etc.,. While the methods of estimation are similar in each case, the present chapter will focus specifically on estimations of the mean. (maybe mention here that this is general discussion with mathematical derivations saved for a later chapter)

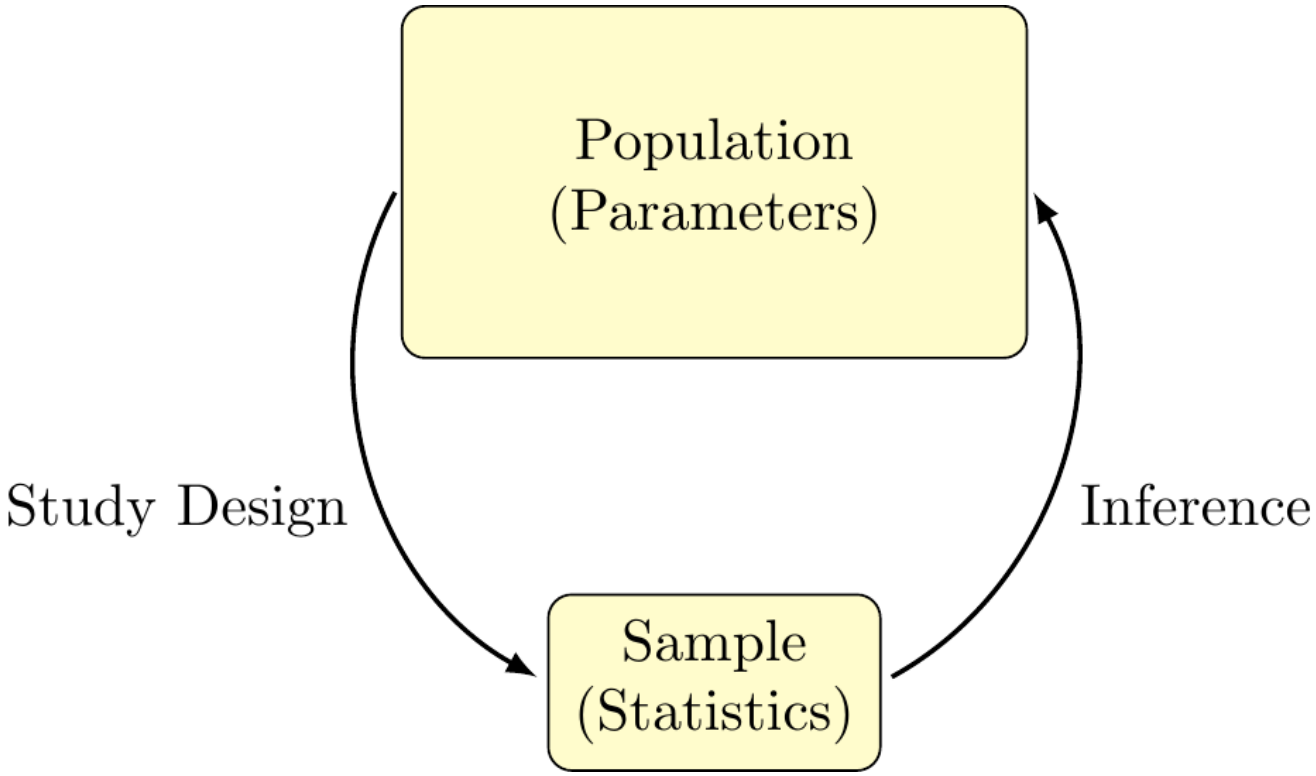


Figure 6.1: Statistical Framework

The statistical framework outlines the general process of making inference using sample statistics to estimate population parameters. In trying to make proper inference, we want to ensure our that our calculated statistic is a valid estimate of the population parameter of interest. To that end, there are two major statistical issues we concern ourselves with:

- On average, does our estimate tend to be centered around the true answer, or is it *biased*?
- How much *variability* is there likely to be in our sample?

The difference between our estimate and our parameter is called **bias**. **Variance** (or “noise”) is a description of the spread of our data. Sometimes we talk about spread in terms of **precision**, which is the reciprocal of variance. The more precision we have, the less variance, and vice-

versa. A good statistic will have little or no bias and would have minimal variability. We can visualize both bias and variability using a dart board analogy:

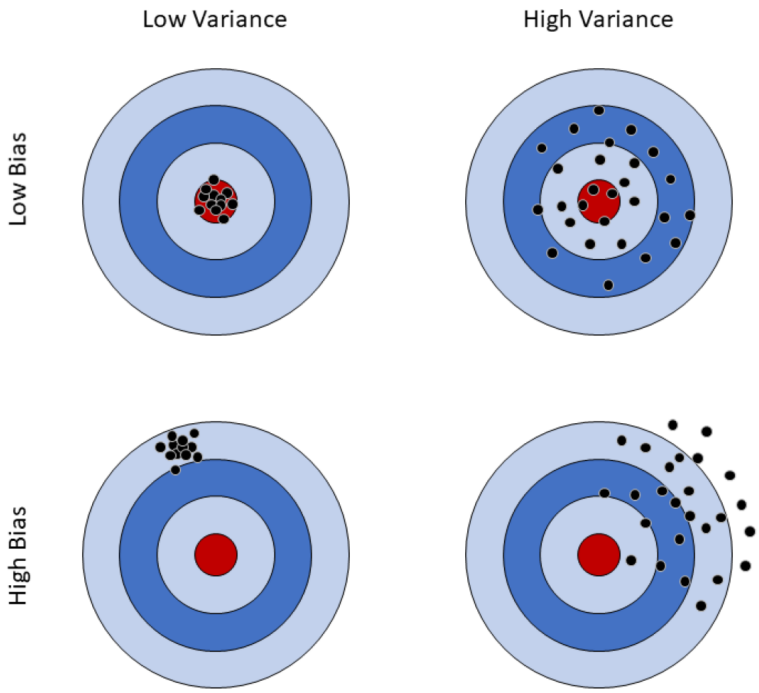


Figure 6.2: Bias and Variability Illustration

In terms of priority, having a low bias is generally more desirable than low variance, and this is represented in the top row above. Having low bias indicates that our sample statistic is correctly estimating the population parameter of interest, even if our statistic is considered "noisy." Variance, illustrated in the columns, gives an idea of how "consistent" our estimate is. One way to think of it is to consider the throwing of a dart to the collection of a sample; comparing the top right and bottom left corners, we might say it is better throw generally close to the target than to consistently miss it entirely. Statisticians work hard to develop sampling and inferential procedures that allow us to estimate population parameters with little low variability and little to no bias.

Definition 6.1

Population Parameter: *The true numeric quantity describing a population (i.e., the population mean)*

Sample Statistic: *A numerical quantity about the sample, often used to estimate a population parameter*

Inference: *The process of making generalizations about the population based on sampling statistics*

Bias: *The systematic difference between our estimate and our parameter*

Variability: *The spread or noise in the data*

Precision: *Inverse of variance, how narrow the data is*

6.2 Sampling Distributions

In order to make inference, we must be able to quantify our uncertainty about the population parameter based on our sample data. To illustrate this uncertainty, consider again our example above in estimating the average expenditures of Iowans on fast food, and suppose that researchers decide to sample 500 individuals from the population and calculate the sample mean, $\overline{x}$. Now suppose that those same researchers go out and collect a *second* sample of 500 individuals from the population and calculate the sample mean again, which me might call $\overline{x}_2$. Should we expect $\overline{x}$ and $\overline{x}_2$ to be exactly the same? What if they did this again and calculated $\overline{x}_3$? Or $\overline{x}_4$?

As we might expect, it is very unlikely that any of these will be exactly equal to any of the others, though they might be close. But if none of these sample means are equal to each other, how accurate might we expect any of them to be with regards to estimating our population parameter? Herein lies the question of quantifying our uncertainty. Fortunately, statisticians have created powerful methods for being able to do so accurately, even when only a single sample statistic has been obtained.

We might first begin by recognizing that the process of sampling from a population is itself a *random experiment*, with the sample mean $\overline{x}$ being a numeric quantity which assumes a value based on the outcome of that experiment. This should sound familiar – this is precisely the definition of a random variable! Just as with other random variables, sample statistics themselves also follow a probability distribution. When the random variable itself is a statistic, we call the probability distribution a **sampling distribution**. Sampling distributions reflect which values of the statistic are likely if we were to repeat the sampling process (that is, likely values for $\overline{x}_2$, $\overline{x}_3$, etc.,), and which values are improbable. We might then expect that the value of the true

population parameter will fall somewhere within the range of likely values. [[As we will see later in this chapter, this uncertainty depends on the number of individuals sampled and the variability of the data we measure.]]

Exercise 6.1

The applet below is designed to help you get familiar with the concept of random sampling. The plot on the left side gives a population distribution consisting of 25 people, where each subject in the population is represented by one box. The x-axis represents the value of the outcome variable for each individual, with stacked boxes indicating that multiple people have the same value. For example, we have one subject in the population with a value of zero, three subjects with values of one, and so on.

The plot on the right displays values for the subjects collected in our random sample, with the same subjects being highlighted in green on the left. We can take samples from our population of various sizes by changing the value of the slider at the top and clicking on the “Sample” button.

Sampling

Sample Size:

Sample

1. Set the sample size to 2 and click the “Sample” button five times. Fill out the following table with the sample means observed in your five samples.

Sample Number	Sample mean
1	
2	
3	
4	
5	

a. Did you get the same sample mean for any of your five samples?
b. Was the sample mean ever equal to the population mean in any of your five samples? If not, how close were your sample means to the true population mean?

2. Now set the sample size to 15 and click the “Sample” button five times. Fill out the following table with the sample means observed in your five samples.

Sample Number	Sample mean
1	
2	
3	
4	
5	

a. Did any of your five simulations give you a sample mean equal to the population mean?
b. How close were your sample means to the true population mean?

3. Compare your results from problems 1 and 2.
   a. In either case were you able to obtain a sample mean equal to the population mean?
   b. Which one had sample means with more variability? What does this indicate about the relationship between the sample size and the distribution of sample means in the context of repeated sampling?

Following the conventions used previously, we will use upper-case letters to denote random variables and lower-case letters to denote values that have been observed (that is, they are no longer random). Regarding the sampling distribution of the mean, $\overline{X}$ represents the random

variable which arises from the repeated random sampling from the population, while $\overline{x}$ represents an *observed* sample mean from an already collected sample.

In describing the sampling distribution, we are often interested in the mean and the standard deviation which, together, describe the size of the interval of probable values for our population parameter. When describing statistics with a sampling distribution, we refer to the sample mean as the **expected value** of our statistic, while the standard deviation is referred to as the **standard error**. For the random variable describing the sample mean, these are denoted as $E(\overline{X})$ and $SE(\overline{X}) = \sqrt{Var(\overline{X})}$, respectively.

Let's return to the fast food scenario to motivate our development, but for now let's assume there are only five individuals in the entire population. The monthly spending for those five individuals (rounded to the nearest dollar) is as follows:

Person	Monthly Spending
A	8
B	22
C	22
D	36
E	50

The population mean can be found by taking the average monthly spending of these five individuals, as they are the only people in this population. This means the population mean is 27.6. Now suppose we are taking samples of three individuals from this population. There are $\binom{5}{3}$ = 10 ways we can sample three people from this population. Since this is a small population, we can enumerate all possible samples, the values we would obtain for the monthly spending in each sample, and then the sample mean monthly fast food spending for each sample:

Sample	Values	$\overline{x}$
(A, B, C)	(8, 22, 22)	17.33
(A, B, D), (A, C, D)	(8, 22, 36), (8, 22, 36)	22.00
(A, B, E), (A, C, E), (B, C, D)	(8, 22, 50), (8, 22, 50), (22, 22, 36)	22.00
(A, D, E), (B, C, E)	(8, 36, 50), (22, 22, 50)	26.67
(B, D, E), (C, D, E)	(22, 36, 50), (22, 36, 50)	36.00

One thing to note is that *none* of the sample means are actually equal to the population mean. This is often the case! Based on this table, we can construct the sampling distribution of $\overline{X}$ for a sample of size three.

Why is our sample mean often not equal to the population mean?

	Probability
P($\overline{x}$ = 17.33)	0.1
P($\overline{x}$ = 22)	0.2
P($\overline{x}$ = 26.67)	0.3
P($\overline{x}$ = 31.33)	0.2
P($\overline{x}$ = 36)	0.2

With this probability distribution, we can get the expected value of the sampling distribution. Whenever we think about probability, we are thinking about long-run frequencies, so when we think about the expected value, we are thinking about the average sample mean if we were to take repeated samples of size three from this population. The probability distribution indicates that if we took samples of size three from this population over and over again (say 1,000 times),

we would end up with 10% with a mean of 17.33, 20% with a mean of 22, 30% with a mean of 26.67; and so on. So the expected value (average) we would observed if we kept taking samples over and over would be:

(0.1 $\times$ 17.33) + (0.2 $\times$ 22) + (0.3 $\times$ 26.67) + (0.2 $\times$ 31.33) + (0.2 $\times$ 36) = 27.6

But wait! That *is* the population mean. This illustrates an extremely powerful property of the sample mean - the expected value of the sample mean $\overline{X}$ is equal to the population parameter μ! Because of this, we say that the sample mean is an **unbiased estimator** of the population mean. While we illustrated the property with a small example, this holds regardless of the population or sample size.

Definition 6.2

Sampling Distribution: *A probability distribution associated with a sample statistic*

Expected Value: *The mean value of a statistic in repeated sampling*

Standard Error: *The standard deviation of a statistic (including the mean) in repeated sampling*

Unbiased Estimator: *The statistic with an expected value equal to the true population parameter*

6.3 Central Limit Theorem

In the last section, we used a toy example to conceptualize repeated sampling and the property of the sample mean being an unbiased estimator. Now we will formalize these properties. For *any* population distribution which can be described by a random variable X with mean μ and variance σ^2, the sample distribution of the mean based on sample of size n, denoted $\overline{X}_n$, has the following properties:

1. The expected value of the sampling distribution is μ

$$E(\overline{X}_n) = \mu$$

2. The variance of the sampling distribution is

$$Var(\overline{X}_n) = \frac{\sigma^2}{n}$$

Together, these expressions are used to describe the typical value of the sample mean as well as the expected amount of variability in the sample mean under repeated sampling. From this, there are two further observations that are worth noting:

1. The expected value of the sample mean is always equal to μ, regardless of the size of the sample: this property makes the sample mean an *unbiased* estimator of the population mean.

2. Whereas the expected value of the sample mean does not depend on n, the variance of the sampling distribution does. Indeed, this should match our intuition; as the size of our sample increases, the variability of $\overline{X}_n$ decreases, giving us a more precise estimate of the population mean.

As we've seen already, it is more common to discuss the standard deviation of a statistic than the variance. In the context of sampling distributions, this quantity is called the *standard error* (to emphasize that we are discussing a sampling distribution) and is denoted

$$SE(\overline{X}_n) = \sqrt{Var(\overline{X}_n)} = \sigma/\sqrt{n}$$

Note that while all standard errors could be referred to as standard deviations, the converse does not hold – standard deviations not associated with a sampling distribution are never referred to as standard errors.

Until now, we have discussed the mean and standard error of a random statistic without having described the sampling *distribution* itself. Why might having the distribution of a statistic be useful? Recall from the previous chapter that we defined a distribution as a method of associating an event or outcome with its corresponding probability. As we shall see, the sample mean, along with its standard error, will help us in constructing a range of "likely" values that may contain the true parameter. With the use of a sampling distribution, we will now be able to assign a probability to the range of these likely values.

As it turns out, the process of randomly collecting a sample from a population and computing its sample means has a truly amazing property – the sampling distribution will always be normally distributed. This brings us to what is largely considered the "fundamental theorem of statistics."

Central Limit Theorem

What the Central Limit Theorem (CLT) states is this: so long as the size of our sample is “large enough,” the sampling distribution of $\overline{X}$ will be approximately normal, regardless of the characteristics of the underlying distribution. That is, we have

$$\overline{X}_n \sim N(\mu, \sigma^2/n).$$

While the term “large enough” will inevitably depend on the context of the problem at hand, a typical rule of thumb puts this value at $n \geq 30$. Because the variance of $\overline{X}$ is inversely related to the sample size, as the n increases, the variance decreases and the distribution becomes more concentrated around its mean value. This is one of the most important, remarkable, and powerful results in all of statistics. In the real world, we rarely know the underlying population distribution of our data, yet the CLT assures us that this is not an issue. In the next chapter, we will investigate how the consequences of the CLT create an immensely powerful tool in performing statistical inference.

Exercise 6.2

The applet below is designed to help your conceptual understanding of the CLT. You can change the underlying population distribution, the size of the samples taken (n), and the number of experiments performed. Each experiment consists of taking a sample of the specified size from the population and calculating the sample mean. The top left panel (blue) shows the underlying population distribution. The top right panel (red) shows the data from the most recent sample, with the sample mean from that specific sample indicated by the dashed black line. The bottom panel (gold) shows us the distribution of sample means from all the experiments, with the observed mean of the sample means indicated by the dashed black line and the true underlying population mean indicated by the solid blue line. Once you select the parameters to the values you want, click “Run Simulation” to tabulate the results.

Central Limit Theorem for Means

Population Distribution:

Normal

Sample Size:

10

Run Simulation

Set the population distribution and sample size.

Each population distribution has an associated population mean.

Click 'Run Simulation' to repeat 10,000 experiments sampling from the population.

1. Set the population distribution to Normal, the sample size to 30 and perform 100 experiments to collect 100 sample means.
    a. Describe the histogram of the the data from the last experiment. Comment on the modality and skew.
    b. Describe the histogram of the distribution of sample means from all experiments.

c. What is the range of values observed for the distribution of sample means? How does this compare the the range of values observed in the data from the last experiment?

2. Now adjust the sample size to 100. Perform 100 experiments.
   a. What is the range of values observed for the distribution of sample means? How does this compare to the range observed when the sample size was 30?
   b. What about the range of values observed in the data from the last experiment – did this change much when the sample size was increased?
   c. Play around with various sample sizes. How does changing the sample size effect the spread of the distribution of sample means?
3. Return to a sample size of 30, but change the population distribution to be right skewed. Perform 1,000 experiments.
   a. Describe the histogram of the the data from the last experiment. Comment on the modality and skew.
   b. Describe the histogram of the distribution of sample means from all experiments. Does it resemble the population distribution?
   c. Re-run the simulation using a sample size of 10. How does this effect the distribution of the sample means? How does this compare to what you observed when taking samples of size 10 with a normal population distribution?
4. Change the population distribution to be left skewed, set the sample size to 80, and perform 1,000 experiments.
   a. Describe the histogram of the the data from the last experiment. Comment on the modality and skew.
   b. Describe the histogram of the distribution of sample means from all experiments. Does it resemble the population distribution?
   c. Run the simulation with these parameters a few times and pay attention to how the mean of the distribution of sample means compares to the population mean. Is the population mean similar or different from the mean of the sample means? What property does this illustrate?
5. Change the parameters of the simulations as needed to answer the following true/false questions. Explain your answers.
   a. CLT tells us that the distribution of data from any experiment will be normally distributed.
   b. With a larger sample size, the data from the last experiment will resemble the population distribution.
   c. Performing 10 experiments is enough to see the effect of CLT.
   d. Regardless of the underlying population distribution and the sample size, the distribution of the sample means will be normally distributed.

e. Increasing the sample size causes the data from a single experiment to look more and more normally distributed.

Definition 6.3

Central Limit Theorem: *A mathematical theorem that provides us with the sampling distribution of the mean*

7 Introduction to Inference

> " A useful property of a test of significance is that it exerts a sobering influence on the type of experimenter who jumps to conclusions on scanty data, and who might otherwise try to make everyone excited about some sensational treatment effect that can well be ascribed to the ordinary variation in his experiment."
>
> — Gertrude Mary Cox

Learning objectives

1. Understand conceptually how confidence intervals are constructed
2. Know correct interpretation of confidence intervals
3. Hypothesis testing null and alternative hypothesis
4. Type 1 and type 2 error
5. Definition of p-value and how to interpret

7.1 Statistical Inference

It may be helpful at this point to quickly review our goals as statisticians, as well as some of the key concepts that we have covered so far:

1. First, let's consider again the primary goal of statistical analysis. We often begin with some population of interest. In particular, we are generally interested in determining some numerical quantity that describes this population. This may include things such as average monthly spending or the proportion of individuals who respond positively to some treatment. As it is typically impractical to measure some quantity within an entire population, we are often limited to collecting a *random sample* of the population from which we hope to make generalizations that apply to the population as a whole.

2. The data that we collect is usually *random*, and we can quantify the amount of uncertainty associated with a particular outcome with a numerical quantity known as a probability. The relationship between possible values of our data and their associated probabilities is known as a *probability distribution*, two of the most widely used being the normal and binomial distributions, covered in chapter 5.

3. The act of collecting a sample and computing a statistic is itself a random process and, as such, also follows a probability distribution known as a *sampling distribution*. As we saw with the Central Limit Theorem introduced in the last chapter, as the number of individuals collected in our sample increases, the computed statistic will increasingly approximate a normal distribution. As with the normal distribution, the sampling distribution will depend on the value of the sample mean, as well as the size of the standard error

Our goal now is to determine how we might use these ideas to go from information in our random sample to statements about the broader population. This process is known as statistical inference and is the focus of the present chapter.

Let's start by playing a simple "game" in the exercise below.

Exercise 7.1

This applet is designed to provide some intuition into inferential thinking. The goal of the game is to determine the true population mean. There are six game modes, each differentiated by the size of the sample collected. Each game mode has a different true distribution (different population parameters). Each time the user presses the "Simulate Data" button, a new dataset will be generated from the true underlying distribution and the sample mean and sample standard deviation are computed. AFTER you make your guess at the true mean, you can get the truth by selecting the "Get True Mean" button and subsequently hide the truth by selecting the "Hide True Mean" button. DO NOT look at the true mean until you are done with all of the exercises, or you will ruin the game!

Simulation Specification

Sample size

30

Simulate Data

Get True Mean | Hide True Mean

1. Start with the game mode corresponding to a sample size of 10. Draw at least five samples and record the sample means and sample standard deviations (SD) you observe as in the table below.

Sample	Sample mean	Sample SD
1		
2		
3		
4		
5		

2. Based on what you have observed in your samples, what do you think is a range of values that would have a good chance of containing the true population mean? How did you select this interval?
3. If you could only select one value, what is your best guess for the true population mean? How did you determine your guess?
4. Based on what you observed, do you think it is likely or unlikely that the true population mean is -6? Explain your reasoning.
5. Now look at the true mean, how close was your best guess? Was the true mean contained in your interval?
6. Repeat exercises (1-5) for each game mode, until you have a range and your best guess of the true mean for all six sample sizes. Determine if the given comparison value for each scenario is likely or unlikely. Summarize your work in the following table.

Sample size	Interval	Best guess	Comparison value	Likely/unlikely?	Truth
10			-6		
30			3		
50			0.2		
100			70		
200			0		
500			20		

   a. Were some of the game modes easier than others? Why do you think this was?
   b. What does the Central Limit Theorem tell us about the variation we would expect in the sample mean as the sample size increases?

How good were you at guessing the truth? To play the game, you had to come up with some method to make your best guess, an interval of plausible values, and determine how likely it was that a given value was the truth. These are some of the foundational goals of statistical inference, which help us answer questions like:

- By how much can you expect your blood pressure to be reduced after starting a blood pressure medication?
- How likely is it that a new cancer treatment will lead to fewer deaths than the current treatment standard?

- After getting the flu, for approximately how many days is someone infectious?

In the rest of this chapter, we will further explore these ideas.

7.2 Point Estimation and Confidence Intervals

The idea of coming up with a single best guess at a population parameter is more formally known as **point estimation**. For example, using the mean value of a sample to estimate the mean value of the population is a form of point estimation. In other words, it is a single number computed from the sample data that is used to infer the value of the population parameter.

As we have seen in the probability distribution exercises of chapter 5, the process of obtaining a sample is random, which means that sample means computed from two separate samples taken from the same population will not be identical. To account for both this random variation and the fact that we are generally only able to obtain one sample, we might consider instead determining an interval of likely values that our point estimate may fall within. We might consider this our *margin of error* for the point estimate. A wider interval around a point estimate is associated with a higher margin of error, which may also serve as a measure of our uncertainty This process is known as **interval estimation**. By including a measure of uncertainty, interval estimation is a more informative metric than a point estimation alone, making it a critical part of performing statistical inference.

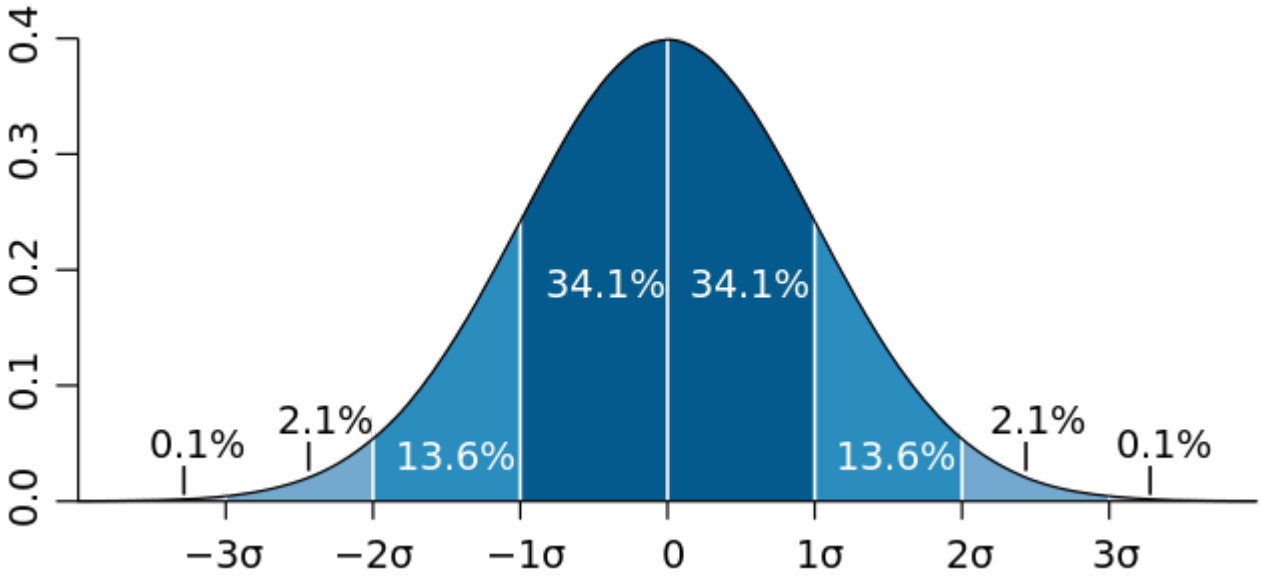
Constructing a desirable interval consists in suitably resolving the tension between two competing values:

1. The interval should be large enough that it contains the *true* parameter value with high probability

2. The interval should be narrow enough to remain useful.

We might consider with an example how these two goals compete: suppose that we are preparing for a trip and trying to anticipate the weather so that we know how to pack. On one hand, there is a clear utility in having an estimate of the temperature range that contains the true value, suggesting the a larger interval may be appropriate. On the other hand, however, having a range of temperatures between 0°F and 100°F hardly tells us if we need to pack a swimming suit or a parka. Fortunately for us, we have statistical tools available to remove much of this guesswork – namely, through an application of the Central Limit Theorem.

How does the Central Limit Theorem help us construct confidence intervals? Well, the CLT tells us that the distribution of the sample mean is normally distributed, and it gives us the parameters (mean and standard deviation) of said distribution. As we saw in Chapter 5, (we need to add this to ch 5) the standard deviation gives us a useful metric in determining how dispersed our data is around the mean. The highlighted sections in the image below detail what proportion of our data lies within a standard deviation of the mean in the case when $\mu = 0$:



For example, knowing that 34.1% of our data will be within one standard deviation (σ) above the mean and 34.1% will be within one standard deviation below. Together, this tells us that the interval generated by $\mu \pm \sigma$ should contain about 68.2% of the total observations. That is,

$$(\mu - \sigma, \mu + \sigma) = \text{interval with } 68.2\% \text{ of data}$$

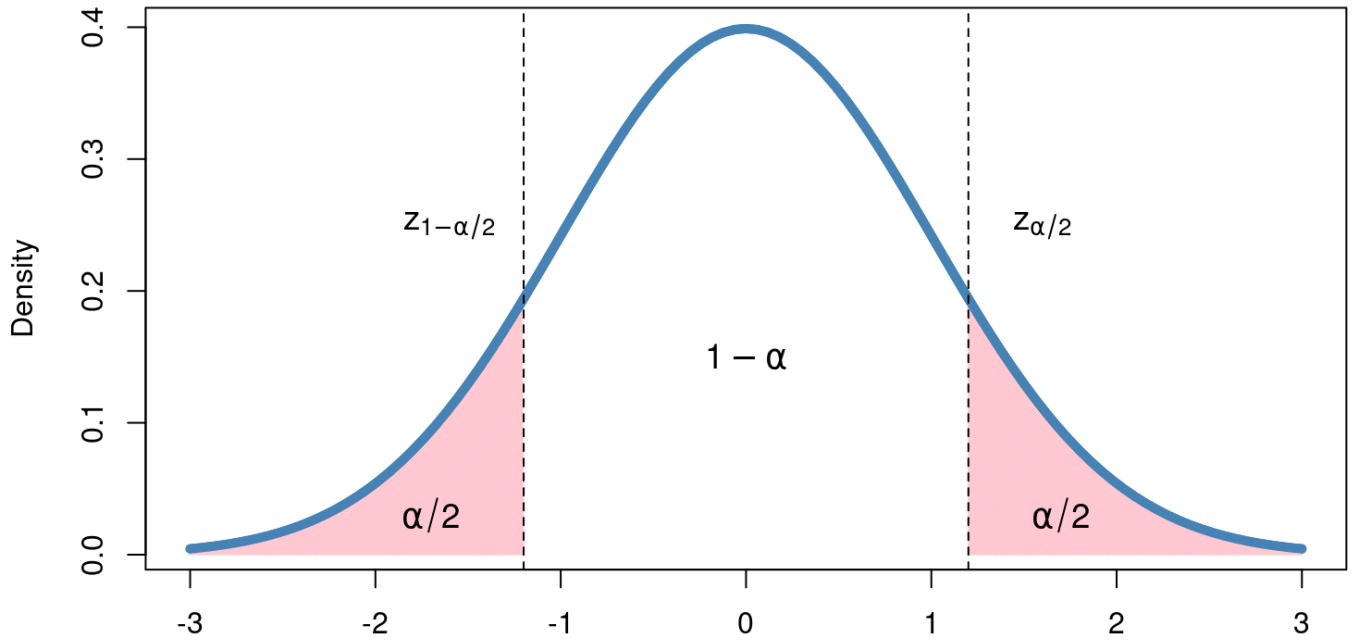
As our sampling distribution is approximately normal, we can approximate this interval with the statistics generated from our sample mean:

$$(\overline{x} - \hat{\sigma}, \overline{x} + \hat{\sigma}) \approx \text{interval with } 68.2\% \text{ of data}$$

Fortunately, this process also works in reverse: once we have determined the probability we want for our interval to contain the true value, we can work backwards to find the appropriate values. A common interval size used in statistics is to capture the true mean with a probability of 95%, also called a 95% **confidence interval**. The probability itself, in this case 95%, is known as the **coverage probability**. For the normal distribution, a 95% coverage probability is given with the interval $\overline{x} \pm 1.96\hat{\sigma}$. It is important to note here that, just like the normal distribution itself, the confidence interval is *also* symmetric around the mean.

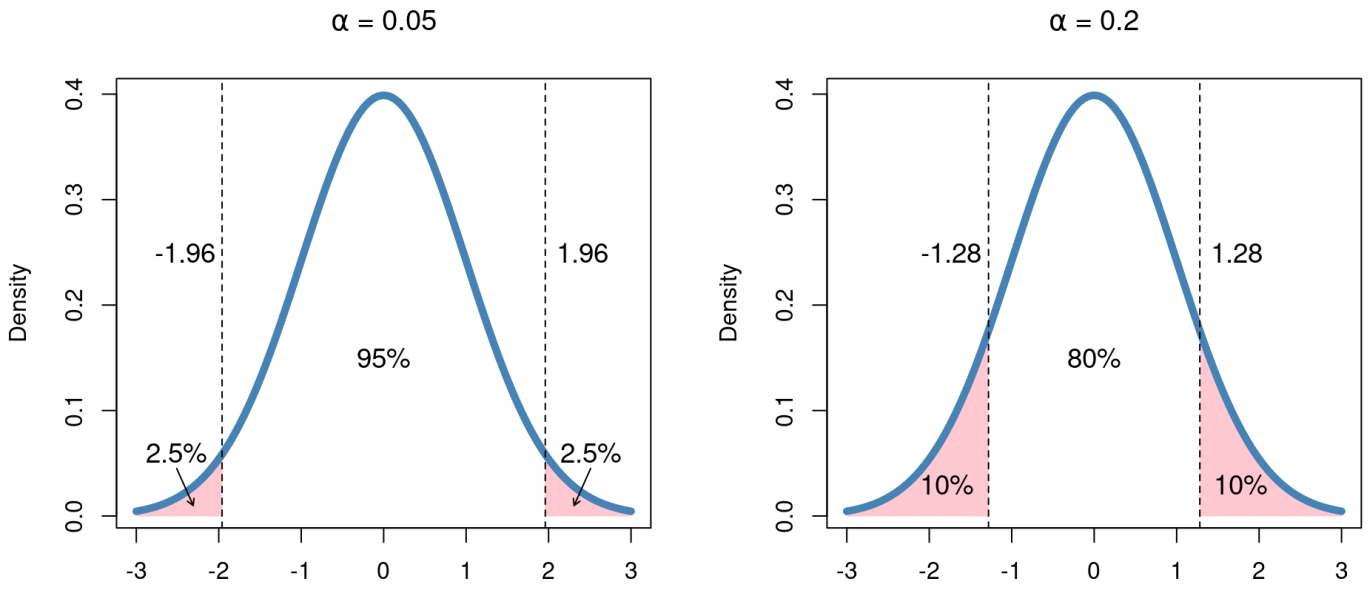
Of course, other values can be used, generating confidence intervals of different sizes. To introduce this more general case (that is, without specifying a value), we need a few variables in the previous example. Let's begin with coverage probability. We use the value $1 - \alpha$ to describe the coverage probability, where α represents the probability that our interval *does not* contain the true mean. This may seem a bit backwards at first, but our reasons for doing so will be clear shortly.

If $1 - \alpha$ represents our coverage probability, then α must represent the probability of our interval not containing the true value. And, as our interval is symmetric about the sample mean, this means that there is a probability of $\alpha/2$ that the true value is *above* the sample mean, and a probability of $\alpha/2$ that it lies *below* the sample mean. This concept is illustrated below:



As you may have guessed, the lines at $z_{1-\alpha/2}$ and $z_{\alpha/2}$ represent the endpoints of our confidence interval. These values are known as **critical values**. In particular, $z_{1-\alpha/2}$ gives us the value for which $\alpha/2\%$ of the distribution is less than $z_{1-\alpha/2}$, and $z_{\alpha/2}$ gives us the value for which $\alpha/2\%$ is greater. In the case of a 95% confidence interval, we have that $\alpha = 0.05$, and our critical values are denoted $z_{0.025}$ and $z_{0.975}$

By changing the values of α, we are able to change the probability that our generated interval contains the true value and, as a consequence, change the size of the interval itself. We illustrate this below by considering two different commonly used values for α and the intervals that are generated:



The mathematical details of constructing a confidence interval depends on the parameter of interest and how the study was designed. Here, however, we will focus on conceptual understanding and interpretation, using sample data and sampling distributions to construct our $100(1-\alpha)\%$ confidence intervals.

Before moving on, let's take a moment to review where we are and what we know:

1. We are interested in constructing an interval that captures a fixed but unknown quantity about our population, in this case, μ. This value is *not* random.
2. What is random, however, are the values used to construct this interval, namely $\overline{X}$ and $\hat{\sigma}$. These values come from our *random sample* and, as such, follow a sampling distribution.
3. In practice, we are generally only able to collect a single random sample, giving us single estimates of $\overline{X}$ and $\hat{\sigma}$ and, consequently, a single confidence interval.

An important question to ask ourselves at this point is: *If there is a single true parameter whose value we do not know, and if we can only construct a single confidence interval from our sample data, what does it mean for a confidence interval to have a coverage probability of 95%?*

The answer to this lies in the sampling process itself. Specifically, what we mean when we say that an interval has a coverage probability of 95% is:

1. If we were to continuously collect samples from our population in the exact same way and
2. for each sample, we compute the sample mean and standard error and
3. use these values to construct our confidence interval then
4. 95% of the time, our computed confidence interval will contain the true parameter value.

In other words, although we will never know for certain if our particular interval contains the true value, we can take confidence (see what we did there?) in knowing that the *process* will give us a valid interval 95% of the time. Let's explore this concept a little further by playing with the following applet.

Exercise 7.2

The applet below is designed to help you get familiar with some of the important concepts related to confidence intervals. The inputs to the simulation are the sample size, number of experiments, and the confidence level. The simulation proceeds by taking a sample of the specified size and constructing a confidence interval based on that sample with the specified level of confidence. This process is one experiment. Each sample is drawn from the same population. For a specified number of experiments, the sampling and confidence interval construction is repeated either 10, 50, 100, or 1000 times. The plot shows all the confidence intervals constructed from each experiment - the number of intervals shown is equal to the "Number of Experiments." The true population parameter in this case is 0 (only for simplicity). Intervals are shaded in blue if they cover the true population parameter and red if they do not contain the true parameter. The applet also reports the "observed coverage" from all experiments, this is the proportion of all the intervals shown that did contain the truth. Clicking "Run Simulation" will redo the specified number of experiments using the inputs provided.

1. Set the sample size to 30, the confidence level to 95%, and simulate running 100 experiments (100 intervals total).
    a. How many intervals did not contain the true population parameter? In other words, how many red intervals are there?
    b. What was the observed coverage? Show how this was computed based on the number of intervals that did/did not contain the truth.
    c. Given the specifications use, What would you expect the coverage to be?
    d. Re-run the simulation under the same conditions (click “Run Simulation” again). What was the observed coverage for the second set of 100 experiments?
    e. Did you get the same observed coverage for both simulations? What does this indicates about the simulation?
    f. Re-run the simulation several times and compare the observed coverage from each run. What do you notice?
2. Keep the confidence level at 95%, but change the sample size to 50. Simulate 100 experiments.

a. How do the intervals differ from those constructed from experiments using a sample size of 30?
b. What is the observed coverage? What would you expect the coverage to be based on the specifications used?
c. Across the 100 confidence intervals shown, how do the widths of the confidence intervals compare? Is there any difference in the confidence interval widths between intervals that do contain the true parameter and those that do not?
d. Now change the sample size to 100. How do these intervals compare to those using sample sizes of 30 or 50?
e. Change the number of experiments to be 1000. Play around with various sample sizes. How does changing the sample size effect the observed coverage?

3. Return to a sample size of 30. Now we will simulate 1000 experiments. Simulating more experiments makes things harder to visualize (we can no longer easily count the intervals that did not contain the truth), but the coverage is more stable when the simulation is re-ran.
   a. With a confidence level of 95%, what is the observed coverage?
   b. Re-run the simulation with a sample size of 100. How do these intervals compare to those using sample sizes of 30 or 50?
   c. Re-run the simulation using a confidence level of 50%. How do these intervals compare to those using a confidence level of 95%? What is the observed coverage?
4. Change the parameters of the simulations as needed to answer the following true/false questions. Explain your answers.
   a. As the sample size increases, the coverage probability increases.
   b. As the confidence level decreases, the width of the confidence intervals decreases,
   c. As the sample size increases, the width of the confidence intervals increases.
   d. If a researcher wants a narrower confidence interval, they should obtain a larger sample.
   e. If a researcher wants a narrower confidence interval, they should decrease the confidence level.

Definition 7.1

Point Estimation: *Using a single numeric quantity to estimate a population parameter*

Interval Estimation: *Using a range of numeric values to estimate a population parameter*

Coverage Probability: *The probability that a constructed confidence interval contains the true population parameter*

Critical Value: *The cutoff value that defines the upper and lower bounds of the interval*

Confidence Interval: *An interval estimate which contains the true population parameter according to a specified confidence level*

7.3 Hypothesis Tests

Often, the motivation for performing a particular experiment is to answer a scientific question about a population. We have just seen that, along with the CLT, the collection of a sample mean along with a standard error allows us to construct an intervals of values representing likely values of the true parameter.

Along with parameter estimation, a cornerstone of statistical inference is significance testing. Confidence intervals allow us to construct a range of plausible values for the parameter, and significance tests allow us to determine the likelihood of a parameter taking a certain value. A **hypothesis test** uses sample data to assess the plausibility of each of two competing hypotheses regarding an unknown parameter (or set of parameters). A **statistical hypothesis** is a statement or claim about an unknown parameter. The **null hypothesis** generally represents what is assumed to be true before the experiment is conducted. This hypothesis is typically denoted H_0. The **alternative hypothesis** represents what the investigator is interested in establishing. This hypothesis is typically denoted H_A. Oftentimes when people refer to the "scientific hypothesis," this in reference to the alternative hypothesis - it is what the investigators think will happen or what they want to show. The goal of a hypothesis test is to determine which hypothesis is the most plausible - the null or the alternative.

As an example, consider researchers that have developed a new drug to treat cancer. In order for the drug to be approved for use, the investigators must prove that it is more effective in treating cancer than the current treatment options. To do this, the investigators gather a sample of cancer patients and randomize half of them to receive their new drug and half to receive the current treatment. Then they determine how many patients improved in both groups. In this scenario, the null hypothesis would be that the new drug and the current drug result in the same improvement. Why? Well, the null hypothesis is what is believed before the data was collected. The key is *whose* beliefs we are talking about. While the scientists that developed the drug most likely believe that their new drug is more effective, the rest of the scientific community remains in a

state of uncertainty. The null hypothesis reflects the general beliefs of the scientific community. The alternative hypothesis in this scenario is that the drugs differ in their effectiveness on treating cancer. This is what the researchers hope to show, specifically, they hope to show that their drug is more effective, however, when the study is conducted there is no evidence in either direction, so we leave the alternative hypothesis to also encompass the possibility of the new drug being worse.

In general, we can think about the null hypothesis as being the "baseline," "boring," "nothing to see here" hypothesis. The exact specification will depend on the study context and the type of data being measured (categorical or continuous):

- H_0: the average cholesterol for hypertensive smokers' is no different than the general population
- H_0: no difference between the treatment and control groups
- H_0: men and women have identical probabilities of colorectal cancer
- H_0: observing a "success" in a population is identical to flipping a coin

The hypothesis testing procedure uses probability to quantify the amount of evidence against the null hypothesis. Since the null hypothesis is the baseline, we start by assuming that it is true. Then, we conduct the study and collect data to quantify the likelihood that the the null is true. The reason for this approach is rooted in the scientific method. As we introduced in Chapter 1, the scientific method has 7 steps:

1. Ask a question
2. Do background research
3. Construct a hypothesis
4. Test your hypothesis with a study or an experiment
5. Analyze data and draw conclusions
6. How do the results align with your hypothesis?
7. Communicate results

We are really focusing on steps 3-5. In step 3, we construct the "scientific hypothesis" and in step 4, we test that hypothesis. In order to produce rigorous scientific results we cannot assume that the scientific hypothesis is true, as that is the goal of our study. We must assume the current state of knowledge (null hypothesis) and then if we are to prove that our hypothesis is correct, we would show that if the current knowledge was true, it would be really unlikely that our experiment would have ended up how it did.

A great analogy to the concept of hypothesis testing is our judicial system. In court, the legal principle is that everyone is “innocent until proven guilty” and the prosecution must prove that the accused is guilty beyond a reasonable doubt. In many cases, there may not be definitive evidence if the defendant is actually innocent or guilty. But, if the prosecution can show that the likelihood of the accused individual being guilty is high (or equivalently, that the likelihood of the accused individual being innocent is low), then the defendant will be convicted. In hypothesis testing researchers are like the prosecution and must use data to prove that the null hypothesis (current state of knowledge) is false beyond a reasonable doubt. The next several chapters will go through how we can use different types of data to quantify our evidence against the null hypothesis.

With this set up in mind, we have two possible outcomes of a hypothesis test. Either we conclude that we do not have a lot of evidence against the null hypothesis, i.e. the null hypothesis looks reasonable, and we *fail to reject the null* or we conclude that we have enough evidence against the null and we *reject the null*. It is extremely important to note here that we NEVER accept the null or accept the alternative. Many people find this annoying because we can never say anything with 100% certainty. But this is exactly the point! Remember that statistics is all about quantifying our uncertainty. Think back to our drug development example. There will never ever ever be a drug that works exactly the same in every person that takes it. People are too variable and many aspects of a person’s life impacts how a drug works in their body. So it would be completely unreasonable to say that a new drug works all the time. However, there can be a drug that improves outcomes for the average person or that this drug is likely to improve outcomes in a randomly selected person who takes it.

We can create a two by two table for the results of any hypothesis test. In the rows we have the two possible outcomes from our test - fail to reject H_0 and reject H_0. In the columns we have the true underlying state of nature - either H_0 is true or false.

	H_0 **true**	H_0 **false**
Fail to reject H_0	Correct (1 - α)	Incorrect (β)
Reject H_0	Incorrect α	Correct (1 - β)

In the upper-left and bottom-right cells of the table we are making the correct decision based on our test. When the null hypothesis is true, failing to reject H_0 is the correct decision and when the null hypothesis is false, rejecting H_0 is the correct decision. However, the other two cells

correspond to a mistake being made. Because statisticians are not creative, these mistakes are referred to as **type 1 error** and **type 2 error**. A type 1 error is equivalent to a false alarm or a false positive - the null hypothesis was rejected, when in fact it was true. A type 2 error can be thought of as a missed opportunity or a false negative - the null hypothesis was false, but it was not rejected. Typically, the type 1 error rate is symbolically denoted with α and the type 2 error rate is denoted by β (again the statisticians of the past were not creative).

If we were looking to create a good hypothesis test, we would want to minimize type 1 and type 2 errors. In other words, if we were to conduct our study over and over again and there was something to be found, we would want to reject the null hypothesis (find something) at a high rate and fail to reject the null hypothesis at a low rate. If there was truly nothing to be found, we wouldn't want to find anything and if there is something to be found we want to find it. However, there is a trade-off between type 1 and type 2 error. Let us illustrate this point with a simulation.

Exercise 7.3

In this experiment, we are looking to determine if a coin is fair, i.e. whether or not the probability of heads is 50%. A type 1 error in this context would be concluding that the coin is not fair, when it actually is. A type 2 error in this context would be concluding that the coin is fair when it is actually not. Each experiment consists of flipping a coin 20 times and observing the proportion of the 20 flips that result in heads. If the coin is fair, we would expect around 10 of the 20 flip to result in heads. However, if we observed 11 or 12 heads in one experiment, would you be convinced the coin isn't fair? What about if you observed 19/20 of the flips resulting in heads?

This applet allows the user to specify the "Rejection Threshold" to determine the cutoff where the app will reject the null hypothesis. This is specified in terms of how far off the proportion of heads in the experiment is from 50%. Since we have no way to determine if the coin is more prone to heads or tails, we consider this distance in both directions. If the observed proportion of the 20 flips that result in heads is beyond the specified threshold, the null hypothesis will be rejected. For example, if you choose a threshold of 10%, then any experiment where there are 60% (12/20) or more flips resulting in heads *or* 40% (8/20) or less flips resulting in heads, then the null hypothesis is rejected and the coin is determined to not be fair.

The simulation involves replicating the 20 flip experiment 10,000 times. For each experiment, a coin is flipped 20 times and the proportion of heads is calculated. If that proportion exceeds the specified threshold it is counted as "Reject." If the proportion does not exceed the threshold, that experiment is considered a "Fail to Reject." The barchart shows the proportion of 10,000 experiments in which the null hypothesis was and was not rejected. Note: Since we are

replicating the experiment 10,000 times, we don't expect the results to change much if the simulation is re-ran. However, you may see small changes in the proportion of rejects/fail to rejects for the same input values.

1. Set the true status of the coin to be fair and start with a threshold of 5%.
    a. In this setting, if we reject the null hypothesis are we making the correct conclusion or the incorrect conclusion? Explain.
    b. At this threshold, how many heads would cause an experiment to reject the null hypothesis
    c. At this threshold, what proportion of the experiments resulted in the null hypothesis being rejected?
    d. In terms of α and β as described above, which of those two can you specify under these circumstances (i.e., with the true status of the coin being fair and the threshold of 5%)? What is it's value?

e. As the threshold increases, what happens to the proportion of experiments in which the null hypothesis is rejected
f. Did we investigate type 1 or type 2 errors in this problem?

2. Now change the true status of the coin to be unfair with a 70% chance of heads and set the rejection threshold to 15%.
    a. In this setting, if we reject the null hypothesis are we making the correct conclusion or the incorrect conclusion? Explain.
    b. At this threshold, how many heads would cause an experiment to reject the null hypothesis
    c. At this threshold, what proportion of the experiments resulted in the null hypothesis *not* being rejected?
    d. In terms of α and β as described above, which of those two can you specify under these circumstances (i.e., with the true status of the coin being fair and the threshold of 5%)? What is it's value?
    e. As the threshold increases, what happens to the proportion of experiments in which the null hypothesis is rejected? Explain why.
    f. Did we investigate type 1 or type 2 errors in this problem?
3. Now change the true status of the coin to be unfair with a 20% chance of heads and set the rejection threshold to 20%
    a. In this setting, if we reject the null hypothesis are we making the correct conclusion or the incorrect conclusion? Explain.
    b. At this threshold, how many heads would cause an experiment to reject the null hypothesis
    c. At this threshold, what proportion of the experiments resulted in the null hypothesis *not* being rejected?
    d. In terms of α and β as described above, which of those two can you specify under these circumstances (i.e., with the true status of the coin being fair and the threshold of 5%)? What is it's value?
    e. As the threshold increases, what happens to the proportion of experiments in which the null hypothesis is rejected?
    f. Did we investigate Type 1 or Type 2 errors in this problem?
4. What can you conclude about the relationship between type 1 and type 2 errors?

Definition 7.2

Hypothesis test: *A decision making technique which assesses the plausibility of two competing hypotheses*

Statistical hypothesis: *A statement about the value of an unknown parameter*

Null hypothesis: *The assumed state of truth prior to running the experiment, denoted* H_0

Alternative hypothesis: *Range of values for the parameter we might believe are true, denoted* H_A

Type I error: *Rejecting a true null hypothesis, i.e., false positive*

Type II error: *Failing to reject a false null hypothesis i.e., false negative*

7.4 P-values

All hypothesis tests are based on quantifying the probability of the study results assuming the null hypothesis is true. This probability is so important that it has a special name, the **p-value**. In technical terms, the p-value gives the probability of obtaining results as extreme or more extreme than the ones observed in the sample, *given* that the null hypothesis is true. A less technical way to describe a p-value is that assuming there is truly nothing going on, what's the chances of obtaining results similar in opposition to the null hypothesis as our study?

If we are thinking about hypothesis testing as a court case, p-values are the way that we can quantify the evidence against the defendant. Recall, the prosecution wants to prove beyond a reasonable doubt that the defendant is not innocent. So what does the scientific community consider sufficient evidence? There is a generally agreed-upon scale for interpreting p-values with regards to the strength of evidence that they represent.

p-value	Evidence against null
0.1	Borderline
0.05	Moderate
0.025	Substantial
0.01	Strong
0.001	Overwhelming

Often, the term “**statistically significant**” is used to describe p-values below 0.05, possibly with a descriptive modifier.

- “Borderline significant” (p < 0.1)
- “Highly significant” (p < 0.01)

However, don’t let these clearly arbitrary cutoffs distract you from the main idea that p-values represent - how far off is the data from what you would expect under the null hypothesis. A p-value of 0.04 and 0.000000001 are not at all the same thing, even though both are “statistically significant.”

In general, a p-value cutoff is chosen and if a p-value below the cutoff is observed, the null hypothesis is rejected. The investigators choose this cutoff, which is equivalent to the type I error rate and thus denoted by α, before analyzing the data. Most of the time α is set to 0.05, which means there is a 5% chance of a type I error (false alarm). The smaller the value of α, the greater the “burden of proof” required to reject the null hypothesis. α is also commonly called the **significance level**.

A fundamental property of p-values is that if we use $p < \alpha$ as cutoff for rejecting the null hypothesis, the type I error rate is guaranteed to be no more than α. However, the $p < \alpha$ cutoff guarantees us nothing about the type II error rate. This is because p-values are calculated assuming the null hypothesis is true, so they don’t give us any information about what to expect when the null hypothesis is false.

While p-values are widely used, have a distinct purpose, and can be informative they also have a number of limitations.

Definition 7.3

p-value: *The probability of observing data as extreme or more extreme than what was observed in the sample, given the null hypothesis is true*

Statistically significant: *When the p-value is < 0.05. Not necessarily related to clinical significance*

Significance level: *The p-value cutoff used to determine if the null hypothesis is rejected Denoted α and often considered to be 0.01, 0.05, or 0.1*