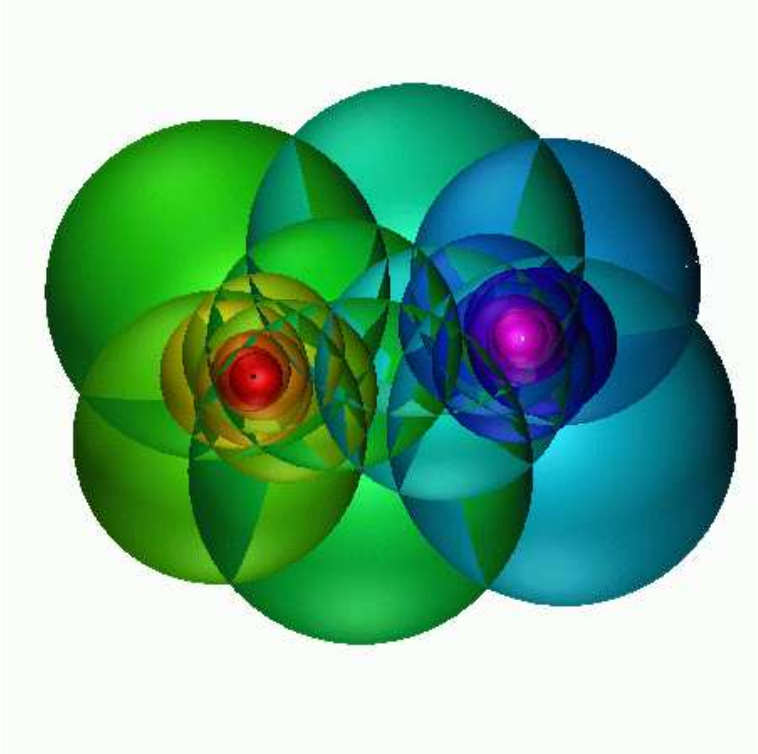


# Lies, Damned Lies, or Statistics:

## *How to Tell the Truth with Statistics*

**Jonathan A. Poritz**

**Department of Mathematics and Physics**
**Colorado State University, Pueblo**
**2200 Bonforte Blvd.**
**Pueblo, CO 81001, USA**
**E-mail: jonathan@poritz.net**
**Web: poritz.net/jonathan**

13 May 2017 23:04MDT

# Release Notes

This is a first draft of a free (as in speech, not as in beer, [**Sta02**]) (although it is free as in beer as well) textbook for a one-semester, undergraduate statistics course. It was used for Math 156 at Colorado State University–Pueblo in the spring semester of 2017.

Thanks are hereby offered to the students in that class who offered many useful suggestions and found numerous typos. In particular, Julie Berogan has an eagle eye, and found a nearly uncountably infinite number of mistakes, both small and large – thank you!



This version: 13 May 2017 23:04MDT.

Jonathan A. Poritz
Spring Semester, 2017
Pueblo, CO, USA

# Contents

# Preface

Mark Twain's autobiography [**TNA10**] modestly questions his own reporting of the numbers of hours per day he sat down to write, and of the number of words he wrote in that time, saying

> *Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force:*
>
> *"**There are three kinds of lies: lies, damned lies, and statistics.**"*

[emphasis added]

Here Twain gives credit for this pithy tripartite classification of lies to Benjamin Disraeli, who was Prime Minister of the United Kingdom in 1868 (under Queen Victoria), although modern scholars find no evidence that Disraeli was the actual originator of the phrase. But whoever actually deserves credit for the phrase, it does seem that statistics are often used to conceal the truth, rather than to reveal it. So much so, for example, that the wonderful book **How to Lie with Statistics** [**Huf93**], by Darrell Huff, gives many, many examples of misused statistics, and yet merely scratches the surface.

We contend, however, that statistics are not a type of lie, but rather, when used carefully, are an *alternative* to lying. For this reason, we use "or" in the title of this book, where Twain/Disraeli used "and," to underline how we are thinking of statistics, correctly applied, as standing in opposition to lies and damned lies.

But why use such a complicated method of telling the truth as statistics, rather than, say, telling a good story or painting a moving picture? The answer, we believe, is simply that there are many concrete, specific questions that humans have about the world which are best answered by carefully collecting some data and using a modest amount of mathematics and a fair bit of logic to analyze them. The thing about the Scientific Method is that it just seems to work. So why not learn how to use it?

Learning better techniques of critical thinking seems particularly important at this moment of history when our politics in the United States (and elsewhere) are so divisive, and different parties cannot agree about the most basic facts. A lot of commentators from all parts of the political spectrum have speculated about the impact of so-called *fake news* on the outcomes of recent recent elections and other political debates. It is therefore the goal

of this book to help you learn **How to Tell the Truth with Statistics** and, therefore, how to tell when others are telling the truth ... or are faking their “news.”

# Part 1

# Descriptive Statistics

The first instinct of the scientist should be to organize carefully a question of interest, and to collect some data about this question. How to collect good data is a real and important issue, but one we discuss later. Let us instead assume for the moment that we have some data, good or bad, and first consider what to do with them[1]. In particular, we want to describe them, both graphically and with numbers that summarize some of their features.

We will start by making some basic definitions of terminology – words like **individual**, **population**, **variable**, **mean**, **median**, *etc.* – which it will be important for the student to understand carefully and completely. So let's briefly discuss what a definition *is*, in mathematics.

Mathematical definitions should be perfectly precise because they do not *describe* something which is observed out there in the world, since such descriptive definitions might have fuzzy edges. In biology, for example, whether a virus is considered "alive" could be subject to some debate: viruses have some of the characteristics of life, but not others. This makes a mathematician nervous.

When we look at math, however, we should always know exactly which objects satisfy some definition and which do not. For example, an *even number* is a whole number which is two times some other whole number. We can always tell whether some number $n$ is even, then, by simply checking if there is some other number $k$ for which the arithmetic statement $n = 2k$ is true: if so, $n$ is even, if not, $n$ is not even. If you claim a number $n$ is even, you need just state what is the corresponding $k$; if claim it is not even, you have to somehow give a convincing, detailed explanation (dare we call it a "proof") that such a $k$ simply does not exist.

So it is important to learn mathematical definitions carefully, to know what the criteria are for a definition, to know examples that satisfy some definition and other examples which do not.

Note, finally, that in statistics, since we are using mathematics in the real world, there will be some terms (like **individual** and **population**) which will not be exclusively in the mathematical realm and will therefore have less perfectly mathematical definitions. Nevertheless, students should try to be as clear and precise as possible.

The material in this Part is naturally broken into two cases, depending upon whether we measure a single thing about a collection of individuals or we make several measurements. The first case is called **one-variable statistics**, and will be our first major topic. The second case could potentially go as far as **multi-variable statistics**, but we will mostly talk about situations where we make *two* measurements, our second major topic. In this case of **bivariate statistics**, we will not only describe each variable separately (both graphically

[1]The word "data" is really a plural, corresponding to the singular "datum." We will try to remember to use plural forms when we talk about "data," but there will be no penalty for (purely grammatical) failure to do so.

and numerically), but we will also describe their relationship, graphically and numerically as well.

CHAPTER 1

# One-Variable Statistics: Basics

## 1.1. Terminology: Individuals/Population/Variables/Samples

Oddly enough, it is often a lack of clarity about *who* [or *what*] *you are looking at* which makes a lie out of statistics. Here are the terms, then, to keep straight:

DEFINITION 1.1.1. The units which are the objects of a statistical study are called the **individuals** in that study, while the collection of all such individuals is called the **population** of the study.

Note that while the term "individuals" sounds like it is talking about people, the individuals in a study could be things, even abstract things like events.

EXAMPLE 1.1.2. The individuals in a study about a democratic election might be *the voters*. But if you are going to make an accurate prediction of who will win the election, it is important to be more precise about what exactly is the population of all of those individuals [voters] that you intend to study, but it *all eligible voters*, *all registered voters*, *the people who actually voted*, *etc.*

EXAMPLE 1.1.3. If you want to study if a coin is "fair" or not, you would flip it repeatedly. The individuals would then be *flips of that coin*, and the population might be something like *all the flips ever done in the past and all that will every be done in the future*. These individuals are quite abstract, and in fact it is impossible ever to get your hands on all of them (the ones in the future, for example).

EXAMPLE 1.1.4. Suppose we're interested in studying whether doing more homework helps students do better in their studies. So shouldn't the individuals be the students? Well, which students? How about we look only at college students. Which college students? OK, how about students at 4-year colleges and universities in the United States, over the last five years – after all, things might be different in other countries and other historical periods.

Wait, a particular student might sometimes do a lot of homework and sometimes do very little. And what exactly does "do better in their studies" mean? So maybe we should look at each student in each class they take, then we can look at the homework they did for that class and the success they had in it.

Therefore, the individuals in this study would be *individual experiences that students in US 4-year colleges and universities had in the last five years*, and population of the study

would essentially be the collection of all the names on all class rosters of courses in the last five years at all US 4-year colleges and universities.

When doing an actual scientific study, we are usually not interested so much in the individuals themselves, but rather in

DEFINITION 1.1.5. A **variable** in a statistical study is the answer of a question the researcher is asking about each individual. There are two types:

- A **categorical variable** is one whose values have a finite number of possibilities.
- A **quantitative variable** is one whose values are numbers (so, potentially an infinite number of possibilities).

The variable is something which (as the name says) *varies*, in the sense that it can have a different value for each individual in the population (although that is not necessary).

EXAMPLE 1.1.6. In Example 1.1.2, the variable most likely would be *who they voted for*, a categorical variable with only possible values "Mickey Mouse" or "Daffy Duck" (or whoever the names on the ballot were).

EXAMPLE 1.1.7. In Example 1.1.3, the variable most likely would be *what face of the coin was facing up after the flip*, a categorical variable with values "heads" and "tails."

EXAMPLE 1.1.8. There are several variables we might use in Example 1.1.4. One might be *how many homework problems did the student do in that course*. Another could be *how many hours total did the student spend doing homework over that whole semester, for that course*. Both of those would be quantitative variables.

A categorical variable for the same population would be *what letter grade did the student get in the course*, which has possible values **A**, **A-**, **B+**, ..., **D-**, **F**.

In many [most?] interesting studies, the population is too large for it to be practical to go observe the values of some interesting variable. Sometimes it is not just impractical, but actually impossible – think of the example we gave of all the flips of the coin, even in the ones in the future. So instead, we often work with

DEFINITION 1.1.9. A **sample** is a subset of a population under study.

Often we use the variable $N$ to indicate the size of a whole population and the variable $n$ for the size of a sample; as we have said, usually $n < N$.

Later we shall discuss how to pick a good sample, and how much we can learn about a population from looking at the values of a variable of interest only for the individuals in a sample. For the rest of this chapter, however, let's just consider what to do with these sample values.

## 1.2. Visual Representation of Data, I: Categorical Variables

Suppose we have a population and variable in which we are interested. We get a sample, which could be large or small, and look at the values of the our variable for the individuals in that sample. We shall informally refer to this collection of values as a *dataset*.

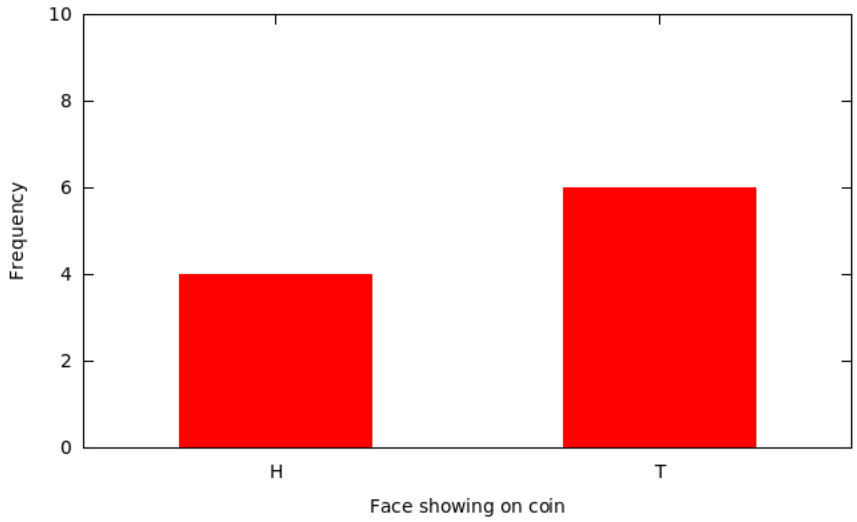
In this section, we suppose also that the variable we are looking at is categorical. Then we can summarize the dataset by telling which categorical values did we see for the individuals in the sample, and how often we saw those values.

There are two ways we can make pictures of this information: *bar charts* and *pie charts*.

**1.2.1. Bar Charts I: Frequency Charts.** We can take the values which we saw for individuals in the sample along the $x$-axis of a graph, and over each such label make a box whose height indicates how many individuals had that value – the **frequency** of occurrence of that value.

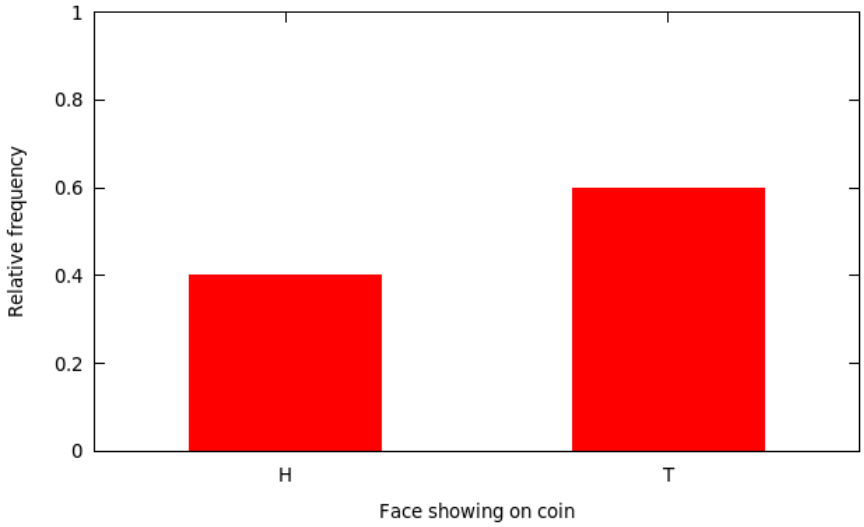
This is called a **bar chart**. As with all graphs, you should *always label all axes.* The $x$-axis will be labeled with some description of the variable in question, the $y$-axis label will always be "frequency" (or some synonym like "count" or "number of times").

EXAMPLE 1.2.1. In Example 1.1.7, suppose we took a sample of consisting of the next 10 flips of our coin. Suppose further that 4 of the flips came up heads – write it as "H" – and 6 came up tails, T. Then the corresponding bar chart would look like




**1.2.2. Bar Charts II: Relative Frequency Charts.** There is a variant of the above kind of bar chart which actually looks nearly the same but changes the labels on the $y$-axis. That is, instead of making the height of each bar be how many times each categorical value occurred, we could make it be *what fraction of the sample had that categorical value* – the **relative frequency**. This fraction is often displayed as a percentage.

EXAMPLE 1.2.2. The relative frequency version of the above bar chart in Example 1.2.1 would look like

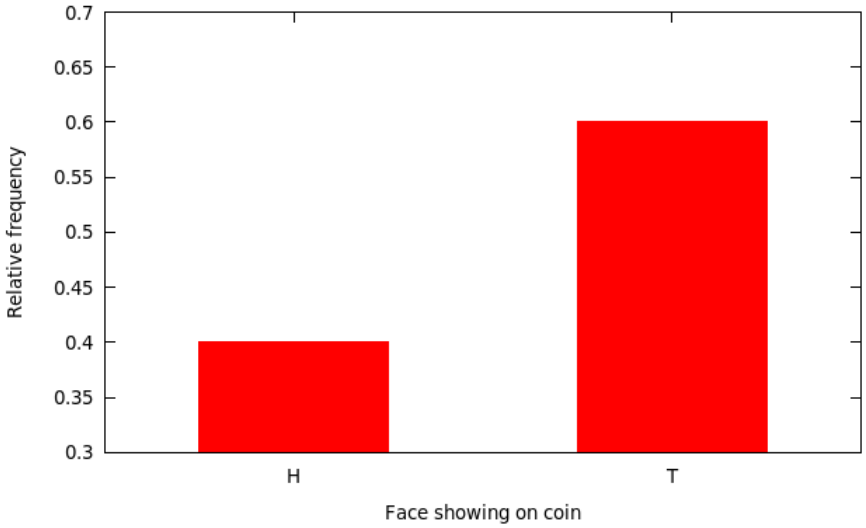



**1.2.3. Bar Charts III: Cautions.** Notice that with bar charts (of either frequency or relative frequency) the variable values along the $x$-axis *can appear in any order whatsoever*. This means that any conclusion you draw from looking at the bar chart must not depend upon that order. For example, it would be foolish to say that the graph in the above Example 1.2.1 "shows and increasing trend," since it would make just as much sense to put the bars in the other order and then "show a decreasing trend" – both are meaningless.

For relative frequency bar charts, in particular, note that the total of the heights of all the bars must be $1$ (or $100\%$). If it is more, something is weird; if it is less, some data has been lost.

Finally, it makes sense for both kinds of bar charts for the $y$-axis to run from the logical minimum to maximum. For frequency charts, this means it should go from $0$ to $n$ (the sample size). For relative frequency charts, it should go from $0$ to $1$ (or $100\%$). Skimping on how much of this appropriate $y$-axis is used is a common trick to lie with statistics.

EXAMPLE 1.2.3. The coin we looked at in Example 1.2.1 and Example 1.2.2 could well be a fair coin – it didn't show exactly half heads and half tails, but it was pretty close. Someone who was trying to claim, deceptively, that the coin was not fair might have shown only a portion of the $y$ axis, as

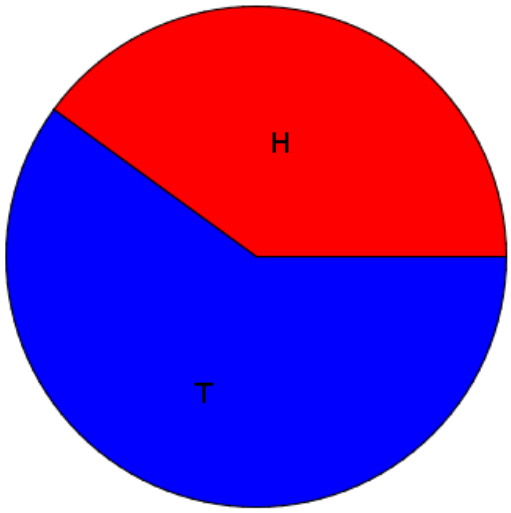



This is actually, in a strictly technical sense, a correct graph. But, looking at it, one might conclude that T seems to occur more than twice as often as H, so the coin is probably not fair ... until a careful examination of the $y$-axis shows that even though the bar for T is more than twice as high as the bar for H, that is an artifact of how much of the $y$-axis is being shown.

In summary, bar charts actually don't have all that much use in sophisticated statistics, but are extremely common in the popular press (and on web sites and so on).

**1.2.4. Pie Charts.** Another way to make a picture with categorical data is to use the fractions from a relative frequency bar chart, but not for the heights of bars, instead for the sizes of wedges of a pie.

EXAMPLE 1.2.4. Here's a pie chart with the relative frequency data from Example 1.2.2.



Pie charts are widely used, but actually they are almost never a good choice. In fact, do an Internet search for the phrase "pie charts are bad" and there will be nearly 3000 hits. Many of the arguments are quite insightful.

When you see a pie chart, it is either an attempt (misguided, though) by someone to be folksy and friendly, or it is a sign that the author is quite unsophisticated with data visualization, or, worst of all, it might be a sign that the author is trying to use mathematical methods in a deceptive way.

In addition, all of the cautions we mentioned above for bar charts of categorical data apply, mostly in exactly the same way, for pie charts.

## 1.3. Visual Representation of Data, II: Quantitative Variables

Now suppose we have a population and *quantitative* variable in which we are interested. We get a sample, which could be large or small, and look at the values of the our variable for the individuals in that sample. There are two ways we tend to make pictures of datasets like this: *stem-and-leaf plots* and *histograms*.

**1.3.1. Stem-and-leaf Plots.** One somewhat old-fashioned way to handle a modest amount of quantitative data produces something between simply a list of all the data values and a graph. It's not a bad technique to know about in case one has to write down a dataset by hand, but very tedious – and quite unnecessary, if one uses modern electronic tools instead – if the dataset has more than a couple dozen values. The easiest case of this technique is where the data are all whole numbers in the range $0 - 99$. In that case, one can take off the tens place of each number – call it the **stem** – and put it on the left side of a vertical bar, and then line up all the ones places – each is a **leaf** – to the right of that stem. The whole thing is called a **stem-and-leaf plot** or, sometimes, just a **stemplot**.

It's important not to skip any stems which are in the middle of the dataset, even if there are no corresponding leaves. It is also a good idea to allow repeated leaves, if there are repeated numbers in the dataset, so that the length of the row of leaves will give a good representation of how much data is in that general group of data values.

EXAMPLE 1.3.1. Here is a list of the scores of 30 students on a statistics test:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 86 | 80 | 25 | 77 | 73 | 76 | 88 | 90 | 69 | 93 |
| 90 | 83 | 70 | 73 | 73 | 70 | 90 | 83 | 71 | 95 |
| 40 | 58 | 68 | 69 | 100 | 78 | 87 | 25 | 92 | 74 |

As we said, using the tens place (and the hundreds place as well, for the data value 100) as the stem and the ones place as the leaf, we get

TABLE 1.3.1.1. Stem-and-leaf plot of students' scores, Key: $1|7 = 17$

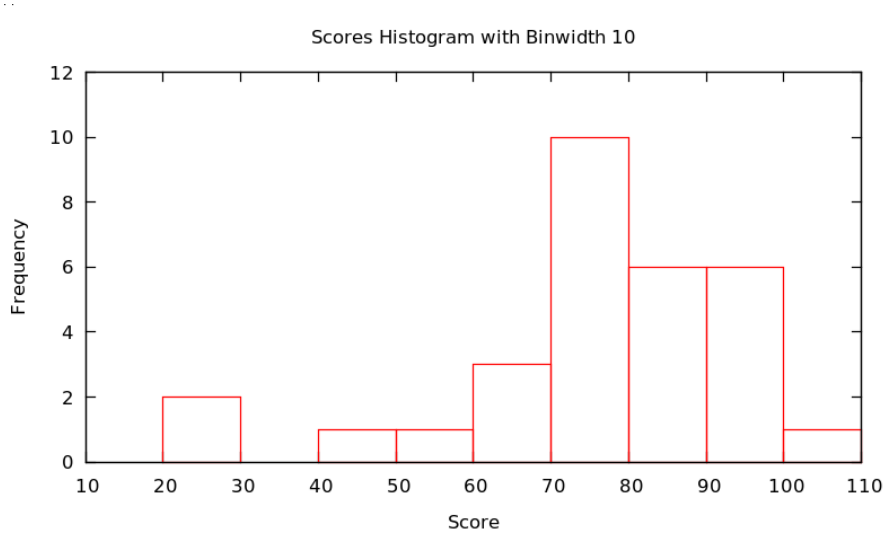
| Stem | Leaf |
|---|---|
| 10 | 0 |
| 9 | 0 0 0 2 3 5 |
| 8 | 0 3 3 6 7 8 |
| 7 | 0 0 1 3 3 3 4 6 7 8 |
| 6 | 8 9 9 |
| 5 | 8 |
| 4 | 0 |
| 3 | |
| 2 | 5 5 |

One nice feature stem-and-leaf plots have is that *they contain all of the data values*, they do not lose anything (unlike our next visualization method, for example).

**1.3.2. [Frequency] Histograms.** The most important visual representation of quantitative data is a **histogram**. Histograms actually look a lot like a stem-and-leaf plot, except turned on its side and with the row of numbers turned into a vertical bar, like a bar graph. The height of each of these bars would be how many

Another way of saying that is that we would be making bars whose heights were determined by how many scores were in each group of ten. Note there is still a question of into which bar a value right on the edge would count: *e.g.,* does the data value $50$ count in the bar to the left of that number, or the bar to the right? It doesn't actually matter which side, but it is important to state which choice is being made.

EXAMPLE 1.3.2. Continuing with the score data in Example 1.3.1 and putting all data values $x$ satisfying $20 \leq x < 30$ in the first bar, values $x$ satisfying $30 \leq x < 40$ in the second, values $x$ satisfying $40 \leq x < 50$ in the second, *etc.* – that is, put data values on the edges in the bar to the right – we get the figure

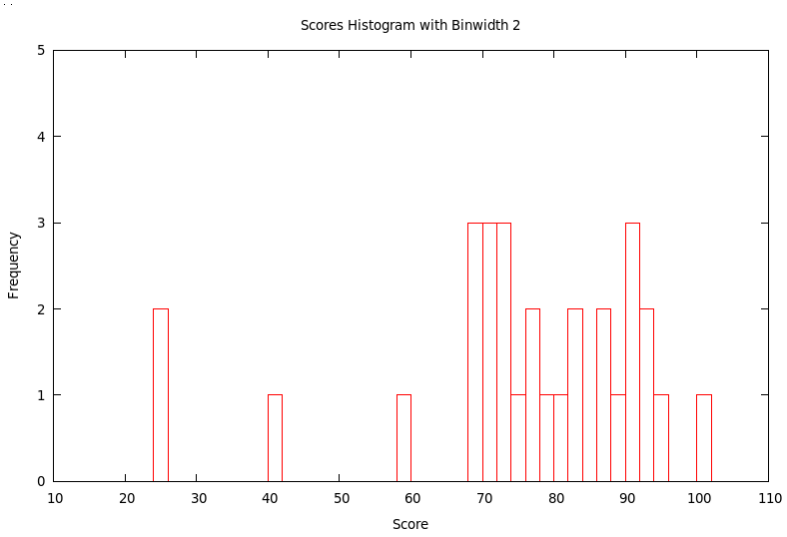



Actually, there is no reason that the bars always have to be ten units wide: it is important that they are all the same size and that how they handle the edge cases (whether the left or right bar gets a data value on edge), but they could be any size. We call the successive ranges of the $x$ coordinates which get put together for each bar the called **bins** or **classes**, and it is up to the statistician to chose whichever bins – where they start and how wide they are – shows the data best.

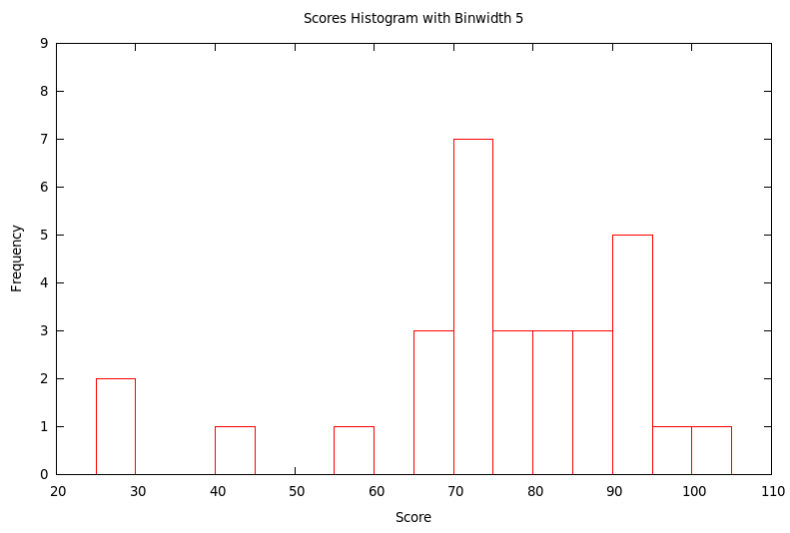
Typically, the smaller the bin size, the more variation (precision) can be seen in the bars ... but sometimes there is so much variation that the result seems to have a lot of random jumps up and down, like static on the radio. On the other hand, using a large bin size makes

the picture smoother ... but sometimes, it is so smooth that very little information is left. Some of this is shown in the following

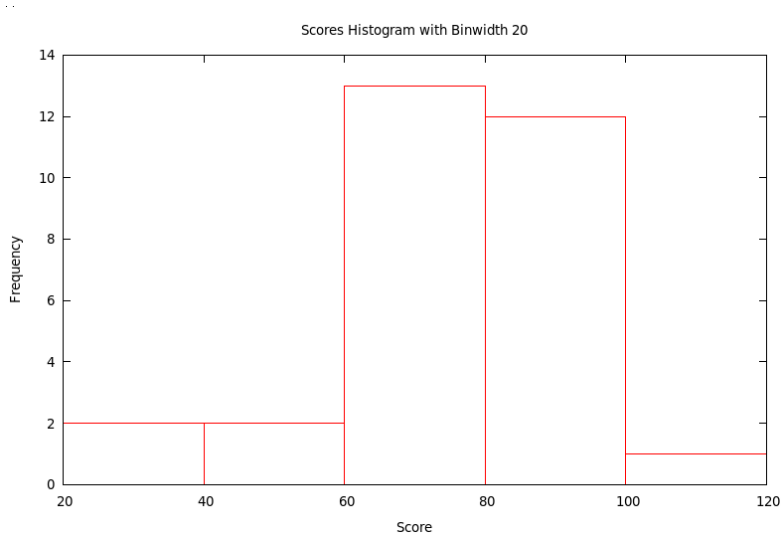
EXAMPLE 1.3.3. Continuing with the score data in Example 1.3.1 and now using the bins with $x$ satisfying $10 \leq x < 12$, then $12 \leq x < 14$, *etc.*, we get the histogram with bins of width 2:



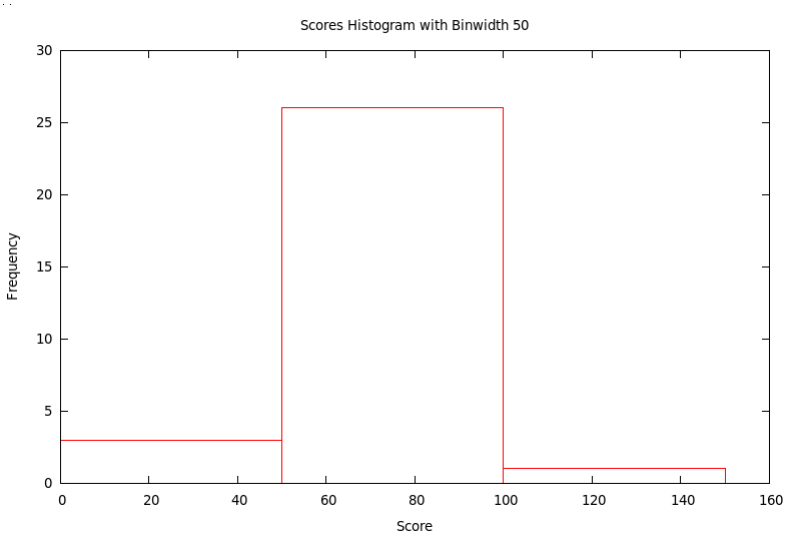

If we use the bins with $x$ satisfying $10 \leq x < 15$, then $15 \leq x < 20$, *etc.*, we get the histogram with bins of width 5:




If we use the bins with $x$ satisfying $20 \leq x < 40$, then $40 \leq x < 60$, *etc.*, we get the histogram with bins of width 20:



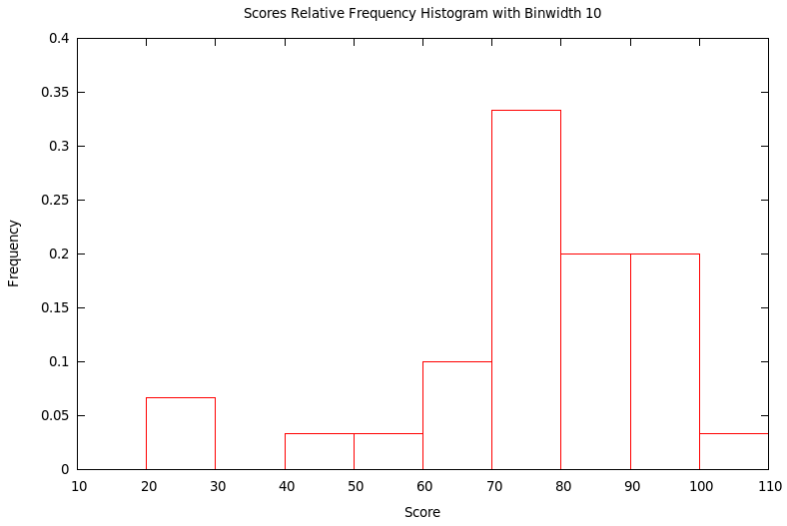

Finally, if we use the bins with $x$ satisfying $0 \leq x < 50$, then $50 \leq x < 100$, and then $100 \leq x < 150$, we get the histogram with bins of width 50:




**1.3.3. [Relative Frequency] Histograms.** Just as we could have bar charts with absolute (§1.2.1) or relative (§1.2.2) frequencies, we can do the same for histograms. Above, in §1.3.2, we made absolute frequency histograms. If, instead, we divide each of the counts used to determine the heights of the bars by the total sample size, we will get fractions or percents – *relative* frequencies. We should then change the label on the $y$-axis and the tick-marks numbers on the $y$-axis, but otherwise the graph will look exactly the same (as it did with relative frequency bar charts compared with absolute frequency bar chars).

EXAMPLE 1.3.4. Let's make the relative frequency histogram corresponding to the absolute frequency histogram in Example 1.3.2, based on the data from Example 1.3.1 – all we have to do is change the numbers used to make heights of the bars in the graph by

dividing them by the sample size, 30, and then also change the $y$-axis label and tick mark numbers.




**1.3.4. How to Talk About Histograms.** Histograms of course tell us what the data values are – the location along the $x$ value of a bar is the value of the variable – and how many of them have each particular value – the height of the bar tells how many data values are in that bin. This is also given a technical name

DEFINITION 1.3.5. Given a variable defined on a population, or at least on a sample, the **distribution** of that variable is a list of all the values the variable actually takes on and how many times it takes on these values.

The reason we like the visual version of a distribution, its histogram, is that our visual intuition can then help us answer general, qualitative questions about what those data must be telling us. The first questions we usually want to answer quickly about the data are

- What is the *shape* of the histogram?
- Where is its *center*?
- How much *variability* [also called *spread*] does it show?

When we talk about the general shape of a histogram, we often use the terms

DEFINITION 1.3.6. A histogram is **symmetric** if the left half is (approximately) the mirror image of the right half.

We say a histogram is **skewed left** if the tail on the left side is longer than on the right. In other words, left skew is when the left half of the histogram – half in the sense that the total of the bars in this left part is half of the size of the dataset – extends farther to the left than the right does to the right. Conversely, the histogram is **skewed right** if the right half extends farther to the right than the left does to the left.

If the shape of the histogram has one significant peak, then we say it is **unimodal**, while if it has several such, we say it is **multimodal**.

It is often easy to point to where the center of a distribution *looks like* it lies, but it is hard to be precise. It is particularly difficult if the histogram is "noisy," maybe multimodal. Similarly, looking at a histogram, it is often easy to say it is "quite spread out" or "very concentrated in the center," but it is then hard to go beyond this general sense.

Precision in our discussion of the center and spread of a dataset will only be possible in the next section, when we work with numerical measures of these features.

## 1.4. Numerical Descriptions of Data, I: Measures of the Center

Oddly enough, there are several measures of central tendency, as ways to define the middle of a dataset are called. There is different work to be done to calculate each of them, and they have different uses, strengths, and weaknesses.

For this whole section we will assume we have collected $n$ numerical values, the values of our quantitative variable for the sample we were able to study. When we write formulæ with these values, we can't give them variable names that look like $a, b, c, \ldots$, because we don't know where to stop (and what would we do if $n$ were more than 26?). Instead, we'll use the variables $x_1, x_2, \ldots, x_n$ to represent the data values.

One more very convenient bit of notation, once we have started writing an unknown number ($n$) of numbers $x_1, x_2, \ldots, x_n$, is a way of writing their sum:

DEFINITION 1.4.1. If we have $n$ numbers which we write $x_1, \ldots, x_n$, then we use the shorthand **summation notation** $\sum x_i$ to represent the sum $\sum x_i = x_1 + \cdots + x_n$. [1]

EXAMPLE 1.4.2. If our dataset were $\{1, 2, 17, -3.1415, 3/4\}$, then $n$ would be 5 and the variables $x_1, \ldots, x_5$ would be defined with values $x_1 = 1$, $x_2 = 2$, $x_3 = 17$, $x_4 = -3.1415$, and $x_5 = 3/4$.

In addition[2], we would have $\sum x_i = x_1+x_2+x_3+x_4+x_5 = 1+2+17-3.1415+3/4 = 17.6085$.

**1.4.1. Mode.** Let's first discuss probably the simplest measure of central tendency, and in fact one which was foreshadowed by terms like "unimodal."

DEFINITION 1.4.3. A **mode** of a dataset $x_1, \ldots, x_n$ of $n$ numbers is one of the values $x_i$ which occurs at least as often in the dataset as any other value.

It would be nice to say this in a simpler way, something like "the mode is the value which occurs the most often in the dataset," but there may not be a single such number.

EXAMPLE 1.4.4. Continuing with the data from Example 1.3.1, it is easy to see, looking at the stem-and-leaf plot, that both 73 and 90 are modes.

Note that in some of the histograms we made using these data and different bin widths, the bins containing 73 and 90 were of the same height, while in others they were of different heights. This is an example of how it can be quite hard to see on a histogram where the mode is... or where the mode**s are**.

[1] Sometimes you will see this written instead $\sum_{i=1}^{n} x_i$ . Think of the "$\sum_{i=1}^{n}$" as a little computer program which with $i = 1$, increases it one step at a time until it gets all the way to $i = n$, and adds up whatever is to the right. So, for example, $\sum_{i=1}^{3} 2i$ would be $(2 * 1) + (2 * 2) + (2 * 3)$, and so has the value 12.

[2] no pun intended

**1.4.2. Mean.** The next measure of central tendency, and certainly the one heard most often in the press, is simply the average. However, in statistics, this is given a different name.

DEFINITION 1.4.5. The **mean** of a dataset $x_1, \ldots, x_n$ of $n$ numbers is given by the formula $\left(\sum x_i\right)/n$.

If the data come from a sample, we use the notation $\overline{x}$ for the **sample mean**.

If $\{x_1, \ldots, x_n\}$ is all of the data from an entire population, we use the notation $\mu_X$ [this is the Greek letter "mu," pronounced "mew," to rhyme with "new."] for the **population mean**.

EXAMPLE 1.4.6. Since we've already computed the sum of the data in Example 1.4.2 to be $17.6085$ and there were $5$ values in the dataset, the mean is $\overline{x} = 17.6085/5 = 3.5217$.

EXAMPLE 1.4.7. Again using the data from Example 1.3.1, we can calculate the mean $\overline{x} = \left(\sum x_i\right)/n = 2246/30 = 74.8667$.

Notice that the mean in the two examples above was not one of the data values. This is true quite often. What that means is that the phrase "the average *whatever*," as in "the average American family has $X$" or "the average student does $Y$," is not talking about any particular family, and we should not expect any particular family or student to have or do that thing. Someone with a statistical education should mentally edit every phrase like that they hear to be instead something like "the mean of the variable $X$ on the population of all American families is ...," or "the mean of the variable $Y$ on the population of all students is ...," or whatever.

**1.4.3. Median.** Our third measure of central tendency is not the result of arithmetic, but instead of putting the data values in increasing order.

DEFINITION 1.4.8. Imagine that we have put the values of a dataset $\{x_1, \ldots, x_n\}$ of $n$ numbers in increasing (or at least non-decreasing) order, so that $x_1 \leq x_2 \leq \cdots \leq x_n$. Then if $n$ is odd, the **median** of the dataset is the middle value, $x_{(n+1)/2}$, while if $n$ is even, the median is the mean of the two middle numbers, $\frac{x_{n/2}+x_{(n/2)+1}}{2}$.

EXAMPLE 1.4.9. Working with the data in Example 1.4.2, we must first put them in order, as $\{-3.1415, 3/4, 1, 2, 17\}$, so the median of this dataset is the middle value, $1$.

EXAMPLE 1.4.10. Now let us find the median of the data from Example 1.3.1. Fortunately, in that example, we made a stem-and-leaf plot and even put the leaves in order, so that starting at the bottom and going along the rows of leaves and then up to the next row, will give us all the values in order! Since there are 30 values, we count up to the $15^{th}$ and $16^{th}$ values, being 76 and 77, and from this we find that the median of the dataset is $\frac{76+77}{2} = 76.5$.

**1.4.4. Strengths and Weaknesses of These Measures of Central Tendency.** The weakest of the three measures above is the mode. Yes, it is nice to know which value happened most often in a dataset (or which values all happened equally often and more often then all other values). But this often does not necessarily tell us much about the over-all structure of the data.

EXAMPLE 1.4.11. Suppose we had the data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 86 | 80 | 25 | 77 | 73 | 76 | 100 | 90 | 67 | 93 |
| 94 | 83 | 72 | 75 | 79 | 70 | 91 | 82 | 71 | 95 |
| 40 | 58 | 68 | 69 | 100 | 78 | 87 | 25 | 92 | 74 |

with corresponding stem-and-leaf plot

| Stem | Leaf |
|---|---|
| 10 | 0 |
| 9 | 0 1 2 3 4 5 |
| 8 | 0 2 3 6 7 8 |
| 7 | 0 1 2 3 4 5 6 7 8 9 |
| 6 | 7 8 9 |
| 5 | 8 |
| 4 | 0 |
| 3 | |
| 2 | 5 5 |

This would have a histogram with bins of width 10 that looks exactly like the one in Example 1.3.2 – so the center of the histogram would seem, visually, still to be around the bar over the 80s – but now there is a unique mode of 25.

What this example shows is that a small change in some of the data values, small enough not to change the histogram at all, can change the mode(s) drastically. It also shows that the location of the mode says very little about the data in general or its shape, the mode is based entirely on a possibly accidental coincidence of some values in the dataset, no matter if those values are in the "center" of the histogram or not.

The mean has a similar problem: a small change in the data, in the sense of adding only one new data value, but one which is very far away from the others, can change the mean quite a bit. Here is an example.

EXAMPLE 1.4.12. Suppose we take the data from Example 1.3.1 but change only one value – such as by changing the 100 to a 1000, perhaps by a simple typo of the data entry. Then if we calculate the mean, we get $\overline{x} = \left(\sum x_i\right)/n = 3146/30 = 104.8667$, which is quite different from the mean of original dataset.

A data value which seems to be quite different from all (or the great majority of) the rest is called an *outlier*[3] What we have just seen is that **the mean is very sensitive to outliers**. This is a serious defect, although otherwise it is easy to compute, to work with, and to prove theorems about.

Finally, the median is somewhat tedious to compute, because the first step is to put all the data values in order, which can be very time-consuming. But, once that is done, throwing in an outlier tends to move the median only a little bit. Here is an example.

EXAMPLE 1.4.13. If we do as in Example 1.4.12 and change the data value of 100 in the dataset of Example 1.3.1 to 1000, but leave all of the other data values unchanged, it does not change the median at all since the 1000 is the new largest value, and that does not change the two middle values at all.

If instead we take the data of Example 1.3.1 and simply add another value, 1000, without taking away the 100, that does change the media: there are now an odd number of data values, so the median is the middle one after they are put in order, which is 78. So the median has changed by only half a point, from 77.5 to 78. And his would even be true if the value we were adding to the dataset were 1000000 and not just 1000!

In other words, **the median is very insensitive to outliers**. Since, in practice, it is very easy for datasets to have a few random, bad values (typos, mechanical errors, *etc.*), which are often outliers, it is usually smarter to use the median than the mean.

As one final point, note that as we mentioned in §1.4.2, the word "average," the unsophisticated version of "mean," is often incorrectly used as a modifier of the individuals in some population being studied (as in "the average American ..."), rather than as a modifier of the variable in the study ("the average income..."), indicating a fundamental misunderstanding of what the mean *means*. If you look a little harder at this misunderstanding, though, perhaps it is based on the idea that we are looking for the center, the "typical" value of the variable.

The mode might seem like a good way – it's the most frequently occurring value. But we have seen how that is somewhat flawed.

The mean might also seem like a good way – it's the "average," literally. But we've also seen problems with the mean.

In fact, the median is probably closest to the intuitive idea of "the center of the data." It is, after all, a value with the property that both above and below that value lie half of the data values.

One last example to underline this idea:

EXAMPLE 1.4.14. The period of economic difficulty for world markets in the late 2000s and early 2010s is sometimes called the **Great Recession**. Suppose a politician says that

[3]This is a very informal definition of an outlier. Below we will have an extremely precise one.

we have come out of that time of troubles, and gives as proof the fact that the average family income has increased from the low value it had during the Great Recession back to the values it had before then, and perhaps is even higher than it was in 2005.

It is possible that in fact people are better off, as the increase in this average – mean – seems to imply. But it is also possible that while the mean income has gone up, the *median* income is still low. This would happen if the histogram of incomes recently still has most of the tall bars down where the variable (family income) is low, but has a few, very high outliers. In short, if the super-rich have gotten even super-richer, that will make the mean (average) go up, even if most of the population has experienced stagnant or decreasing wages – but the median will tell what is happening to most of the population.

So when a politician uses the evidence of the average (mean) as suggested here, it is possible they are trying to hide from the pubic the reality of what is happening to the rich and the not-so-rich. It is also possible that this politician is simply poorly educated in statistics and doesn't realize what is going on. You be the judge ... but pay attention so you know what to ask about.

The last thing we need to say about the strengths and weaknesses of our different measures of central tendency is a way to use the weaknesses of the mean and median to our advantage. That is, since the mean is sensitive to outliers, and pulled in the direction of those outliers, while the median is not, we can use the difference between the two to tell us which way a histogram is skewed.

FACT 1.4.15. If the mean of a dataset is larger than the median, then histograms of that dataset will be right-skewed. Similarly, if the mean is less than the median, histograms will be left-skewed.

## 1.5. Numerical Descriptions of Data, II: Measures of Spread

**1.5.1. Range.** The simplest – and least useful – measure of the spread of some data is literally how much space on the $x$-axis the histogram takes up. To define this, first a bit of convenient notation:

DEFINITION 1.5.1. Suppose $x_1, \ldots, x_n$ is some quantitative dataset. We shall write $x_{min}$ for the smallest and $x_{max}$ for the largest values in the dataset.

With this, we can define our first measure of spread

DEFINITION 1.5.2. Suppose $x_1, \ldots, x_n$ is some quantitative dataset. The **range** of this data is the number $x_{max} - x_{min}$.

EXAMPLE 1.5.3. Using again the statistics test scores data from Example 1.3.1, we can read off from the stem-and-leaf plot that $x_{min} = 25$ and $x_{max} = 100$, so the range is $75 (= 100 - 25)$.

EXAMPLE 1.5.4. Working now with the made-up data in Example 1.4.2, which was put into increasing order in Example 1.4.9, we can see that $x_{min} = -3.1415$ and $x_{max} = 17$, so the range is $20.1415 (= 17 - (-3.1415))$.

The thing to notice here is that since the idea of outliers is that they are outside of the normal behavior of the dataset, if there are any outliers they will definitely be what value gets called $x_{min}$ or $x_{max}$ (or both). So **the range is supremely sensitive to outliers**: if there are any outliers, the range will be determined exactly by them, and not by what the typical data is doing.

**1.5.2. Quartiles and the $IQR$.** Let's try to find a substitute for the range which is not so sensitive to outliers. We want to see how far apart not the maximum and minimum of the whole dataset are, but instead how far apart are the typical larger values in the dataset and the typical smaller values. How can we measure these typical larger and smaller? One way is to define these in terms of the typical – central – value of the upper half of the data and the typical value of the lower half of the data. Here is the definition we shall use for that concept:

DEFINITION 1.5.5. Imagine that we have put the values of a dataset $\{x_1, \ldots, x_n\}$ of $n$ numbers in increasing (or at least non-decreasing) order, so that $x_1 \leq x_2 \leq \cdots \leq x_n$. If $n$ is odd, call the **lower half data** all the values $\{x_1, \ldots, x_{(n-1)/2}\}$ and the **upper half data** all the values $\{x_{(n+3)/2}, \ldots, x_n\}$; if $n$ is even, the **lower half data** will be the values $\{x_1, \ldots, x_{n/2}\}$ and the **upper half data** all the values $\{x_{(n/2)+1}, \ldots, x_n\}$.

Then the **first quartile**, written $Q_1$, is the median of the lower half data, and the **third quartile**, written $Q_3$, is the median of the upper half data.

Note that the first quartile is halfway through the lower half of the data. In other words, it is a value such that one quarter of the data is smaller. Similarly, the third quartile is halfway through the upper half of the data, so it is a value such that three quarters of the data is small. Hence the names "first" and "third quartiles."

We can build a outlier-insensitive measure of spread out of the quartiles.

DEFINITION 1.5.6. Given a quantitative dataset, its **inter-quartile range** or $IQR$ is defined by $IQR = Q_3 - Q_1$.

EXAMPLE 1.5.7. Yet again working with the statistics test scores data from Example 1.3.1, we can count off the lower and upper half datasets from the stem-and-leaf plot, being respectively

$$\text{Lower} = \{25, 25, 40, 58, 68, 69, 69, 70, 70, 71, 73, 73, 73, 74, 76\}$$

and

$$\text{Upper} = \{77, 78, 80, 83, 83, 86, 87, 88, 90, 90, 90, 92, 93, 95, 100\}\ .$$

It follows that, for these data, $Q_1 = 70$ and $Q_3 = 88$, so $IQR = 18 (= 88 - 70)$.

EXAMPLE 1.5.8. Working again with the made-up data in Example 1.4.2, which was put into increasing order in Example 1.4.9, we can see that the lower half data is $\{-3.1415, .75\}$, the upper half is $\{2, 17\}$, $Q_1 = -1.19575 (= \frac{-3.1415+.75}{2})$, $Q_3 = 9.5 (= \frac{2+17}{2})$, and $IQR = 10.69575 (= 9.5 - (-1.19575))$.

**1.5.3. Variance and Standard Deviation.** We've seen a crude measure of spread, like the crude measure "mode" of central tendency. We've also seen a better measure of spread, the $IQR$, which is insensitive to outliers like the median (and built out of medians). It seems that, to fill out the parallel triple of measures, there should be a measure of spread which is similar to the mean. Let's try to build one.

Suppose the data is sample data. Then how far a particular data value $x_i$ is from the sample mean $\overline{x}$ is just $x_i - \overline{x}$. So the mean displacement from the mean, the mean of $x_i - \overline{x}$, should be a good measure of variability, shouldn't it?

Unfortunately, it turns out that the mean of $x_i - \overline{x}$ is always 0. This is because when $x_i > \overline{x}$, $x_i - \overline{x}$ is positive, while when $x_i < \overline{x}$, $x_i - \overline{x}$ is negative, and it turns out that the positives always exactly cancel the negatives (see if you can prove this algebraically, it's not hard).

We therefore need to make the numbers $x_i - \overline{x}$ positive before taking their mean. One way to do this is to square them all. Then we take something which is almost the mean of these squared numbers to get another measure of spread or variability:

DEFINITION 1.5.9. Given sample data $x_1, \ldots, x_n$ from a sample of size $n$, the **sample variance** is defined as

$$S_x^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1}.$$

Out of this, we then define the **sample standard deviation**

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}.$$

Why do we take the square root in that sample standard deviation? The answer is that the measure we build should have the property that if all the numbers are made twice as big, then the measure of spread should also be twice as big. Or, for example, if we first started working with data measured in feet and then at some point decided to work in inches, the numbers would all be 12 times as big, and it would make sense if the measure of spread were also 12 times as big.

The variance does not have this property: if the data are all doubled, the variance increases by a factor of 4. Or if the data are all multiplied by 12, the variance is multiplied by a factor of 144.

If we take the square root of the variance, though, we get back to the nice property of doubling data doubles the measure of spread, *etc.* For this reason, while we have defined the variance on its own and some calculators, computers, and on-line tools will tell the variance whenever you ask them to computer 1-variable statistics, we will in this class only consider the variance a stepping stone on the way to the real measure of spread of data, the standard deviation.

One last thing we should define in this section. For technical reasons that we shall not go into now, the definition of standard deviation is slightly different if we are working with population data and not sample data:

DEFINITION 1.5.10. Given data $x_1, \ldots, x_n$ from an entire population of size $n$, the **population variance** is defined as

$${\sigma_X}^2 = \frac{\sum (x_i - \mu_X)^2}{n}.$$

Out of this, we then define the **population standard deviation**

$$\sigma_X = \sqrt{{\sigma_X}^2} = \sqrt{\frac{\sum (x_i - \mu_X)^2}{n}}.$$

[This letter $\sigma$ is the lower-case Greek letter sigma, whose upper case $\Sigma$ you've seen elsewhere.]

Now for some examples. Notice that to calculate these values, we shall always use an electronic tool like a calculator or a spreadsheet that has a built-in variance and standard deviation program – experience shows that it is nearly impossible to get all the calculations entered correctly into a non-statistical calculator, so we shall not even try.

EXAMPLE 1.5.11. For the statistics test scores data from Example 1.3.1, entering them into a spreadsheet and using `VAR.S` and `STDEV.S` for the sample variance and standard deviation and `VAR.P` and `STDEV.P` for population variance and population standard deviation, we get

$$S_x^2 = 331.98$$
$$S_x = 18.22$$
$${\sigma_X}^2 = 330.92$$
$$\sigma_X = 17.91$$

EXAMPLE 1.5.12. Similarly, for the data in Example 1.4.2, we find in the same way that

$$S_x^2 = 60.60$$
$$S_x = 7.78$$
$${\sigma_X}^2 = 48.48$$
$$\sigma_X = 6.96$$

**1.5.4. Strengths and Weaknesses of These Measures of Spread.** We have already said that **the range is extremely sensitive to outliers**.

The $IQR$, however, is built up out of medians, used in different ways, so **the $IQR$ is insensitive to outliers**.

The variance, both sample and population, is built using a process quite like a mean, and in fact also has the mean itself in the defining formula. Since the standard deviation in both cases is simply the square root of the variance, it follows that **the sample and population variances and standard deviations are all sensitive to outliers**.

This differing sensitivity and insensitivity to outliers is the main difference between the different measures of spread that we have discussed in this section.

One other weakness, in a certain sense, of the $IQR$ is that there are several different definitions in use of the quartiles, based upon whether the median value is included or not when dividing up the data. These are called, for example, `QUARTILE.INC` and `QUARTILE.EXC` on some spreadsheets. It can then be confusing which one to use.

**1.5.5. A Formal Definition of Outliers – the $1.5\,IQR$ Rule.** So far, we have said that outliers are simply data that are *atypical*. We need a precise definition that can be carefully checked. What we will use is a formula (well, actually two formulæ) that describe that idea of an outlier being *far away from the rest of data*.

Actually, since outliers should be far away either in being significantly bigger than the rest of the data or in being significantly smaller, we should take a value on the upper side of the rest of the data, and another on the lower side, as the starting points for this *far away*.

We can't pick the $x_{max}$ and $x_{min}$ as those starting points, since they will be the outliers themselves, as we have noticed. So we will use our earlier idea of a value which is typical for the larger part of the data, the quartile $Q_3$, and $Q_1$ for the corresponding lower part of the data.

Now we need to decide how far is *far enough away* from those quartiles to count as an outlier. If the data already has a lot of variation, then a new data value would have to be quite far in order for us to be sure that it is not out there just because of the variation already in the data. So our measure of *far enough* should be in terms of a measure of spread of the data.

Looking at the last section, we see that only the $IQR$ is a measure of spread which is insensitive to outliers – and we definitely don't want to use a measure which is sensitive to the outliers, one which would have been affected by the very outliers we are trying to define.

All this goes together in the following

DEFINITION 1.5.13. [The $1.5\,IQR$ **Rule for Outliers**] Starting with a quantitative dataset whose first and third quartiles are $Q_1$ and $Q_3$ and whose inter-quartile range is $IQR$, a data value $x$ is [officially, from now on] called an **outlier** if $x < Q_1 - 1.5\,IQR$ or $x > Q_3 + 1.5\,IQR$.

Notice this means that $x$ is not an outlier if it satisfies $Q_1 - 1.5\,IQR \leq x \leq Q_3 + 1.5\,IQR$.

EXAMPLE 1.5.14. Let's see if there were any outliers in the test score dataset from Example 1.3.1. We found the quartiles and $IQR$ in Example 1.5.7, so from the $1.5\,IQR$ Rule, a data value $x$ will be an outlier if

$$x < Q_1 - 1.5\,IQR = 70 - 1.5 \cdot 18 = 43$$

or if

$$x > Q_3 + 1.5\,IQR = 88 + 1.5 \cdot 18 = 115\ .$$

Looking at the stemplot in Table 1.3.1, we conclude that the data values $25$, $25$, and $40$ are the outliers in this dataset.

EXAMPLE 1.5.15. Applying the same method to the data in Example 1.4.2, using the quartiles and $IQR$ from Example 1.5.8, the condition for an outlier $x$ is

$$x < Q_1 - 1.5\,IQR = -1.19575 - 1.5 \cdot 10.69575 = -17.239375$$

or

$$x > Q_3 + 1.5\,IQR = 9.5 + 1.5 \cdot 10.69575 = 25.543625\ .$$

Since none of the data values satisfy either of these conditions, there are no outliers in this dataset.

**1.5.6. The Five-Number Summary and Boxplots.** We have seen that numerical summaries of quantitative data can be very useful for quickly understanding (some things about) the data. It is therefore convenient for a nice package of several of these

DEFINITION 1.5.16. Given a quantitative dataset $\{x_1, \ldots, x_n\}$, the **five-number summary**[4] of this data is the set of values

$$\{x_{min},\ Q_1,\ \text{median},\ Q_3,\ x_{max}\}$$

EXAMPLE 1.5.17. Why not write down the five-number summary for the same test score data we saw in Example 1.3.1? We've already done most of the work, such as calculating the min and max in Example 1.5.3, the quartiles in Example 1.5.7, and the median in Example 1.4.10, so the five-number summary is

$$\begin{aligned} x_{min} &= 25 \\ Q_1 &= 70 \\ \text{median} &= 76.5 \\ Q_3 &= 88 \\ x_{max} &= 100 \end{aligned}$$

EXAMPLE 1.5.18. And, for completeness, the five number summary for the made-up data in Example 1.4.2 is

$$\begin{aligned} x_{min} &= -3.1415 \\ Q_1 &= -1.9575 \\ \text{median} &= 1 \\ Q_3 &= 9.5 \\ x_{max} &= 17 \end{aligned}$$

where we got the min and max from Example 1.5.4, the median from Example 1.4.9, and the quartiles from Example 1.5.8.

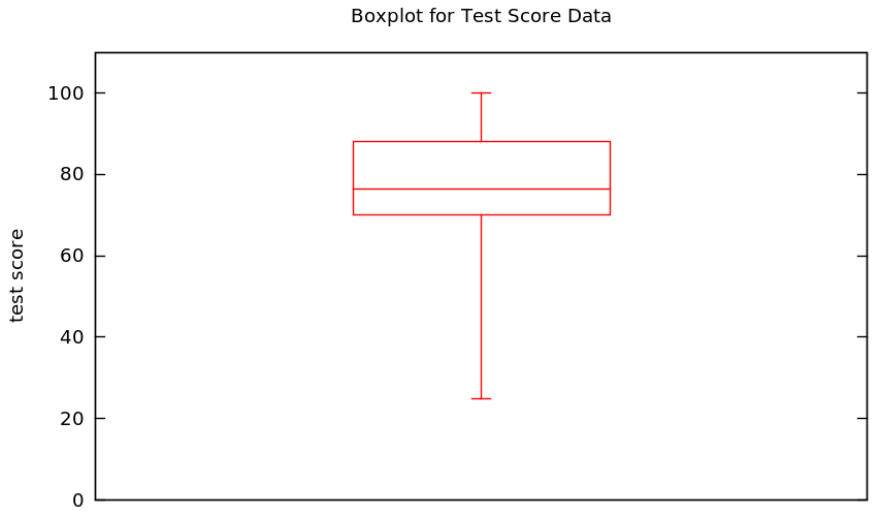
As we have seen already several times, it is nice to have a both a numeric and a graphical/visual version of everything. The graphical equivalent of the five-number summary is

DEFINITION 1.5.19. Given some quantitative data, a **boxplot** [sometimes **box-and-whisker plot**] is a graphical depiction of the five-number summary, as follows:

[4]Which might write 5NΣary for short.

- an axis is drawn, labelled with the variable of the study
- tick marks and numbers are put on the axis, enough to allow the following visual features to be located numerically
- a rectangle (the *box*) is drawn parallel to the axis, stretching from values $Q_1$ to $Q_3$ on the axis
- an addition line is drawn, parallel to the sides of the box at locations $x_{min}$ and $x_{max}$, at the axis coordinate of the median of the data
- lines are drawn parallel to the axis from the middle of sides of the box at the locations $x_{min}$ and $x_{max}$ out to the axis coordinates $x_{min}$ and $x_{max}$, where these *whiskers* terminate in "T"s.

EXAMPLE 1.5.20. A boxplot for the test score data we started using in Example 1.3.1 is easy to make after we found the corresponding five-number summary in Example 1.5.17:

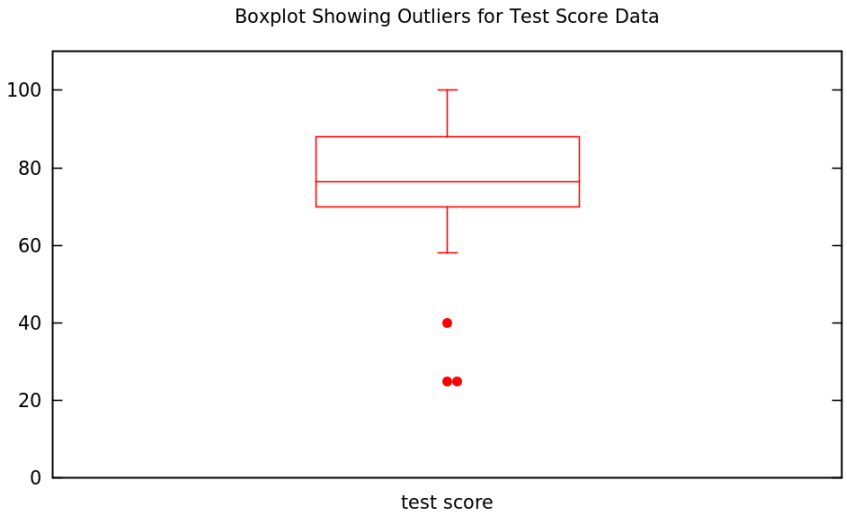



Sometimes it is nice to make a version of the boxplot which is less sensitive to outliers. Since the endpoints of the whiskers are the only parts of the boxplot which are sensitive in this way, they are all we have to change:

DEFINITION 1.5.21. Given some quantitative data, a **boxplot showing outliers** [sometimes **box-and-whisker plot showing outliers**] is minor modification of the regular boxplot, as follows

- the whiskers only extend as far as the largest and smallest non-outlier data values
- dots are put along the lines of the whiskers at the axis coordinates of any outliers in the dataset

EXAMPLE 1.5.22. A boxplot showing outliers for the test score data we started using in Example 1.3.1 is only a small modification of the one we just made in Example 1.5.20



## Exercises

EXERCISE 1.1. A product development manager at the campus bookstore wants to make sure that the backpacks being sold there are strong enough to carry the heavy books students carry around campus. The manager decides she will collect some data on how heavy are the bags/packs/suitcases students are carrying around at the moment, by stopping the next 100 people she meets at the center of campus and measuring.

What are the individuals in this study? What is the population? Is there a sample – what is it? What is the variable? What kind of variable is this?

EXERCISE 1.2. During a blood drive on campus, 300 donated blood. Of these, 136 had blood of type $O$, 120 had blood of type $A$, 32 of type $B$, and the rest of type $AB$.

Answer the same questions as in the previous exercise for this new situation.

Now make at least two visual representations of these data.

EXERCISE 1.3. Go to the **Wikipedia** page for "Heights of Presidents and Presidential Candidates of the United States" and look only at the heights of the presidents themselves, in centimeters (*cm*).

Make a histogram with these data using bins of width 5. Explain how you are handling the edge cases in your histogram.

EXERCISE 1.4. Suppose you go to the supermarket every week for a year and buy a bag of flour, packaged by a major national flour brand, which is labelled as weighing $1kg$. You take the bag home and weigh it on an extremely accurate scale that measures to the nearest $1/100^{th}$ of a gram. After the 52 weeks of the year of flour buying, you make a histogram of the accurate weights of the bags. What do you think that histogram will look like? Will it be symmetric or skewed left or right (which one?), where will its center be, will it show a lot of variation/spread or only a little? Explain why you think each of the things you say.

What about if you buy a $1kg$ loaf of bread from the local artisanal bakery – what would the histogram of the accurate weights of those loaves look like (same questions as for histogram of weights of the bags of flour)?

If you said that those histograms were symmetric, can you think of a measurement you would make in a grocery store or bakery which would be skewed; and if you said the histograms for flour and loaf weights were skewed, can you think of one which would be symmetric? (Explain why, always, of course.) [If you think one of the two above histograms was skewed and one was symmetric (with explanation), you don't need to come up with another one here.]

EXERCISE 1.5. Twenty sacks of grain weigh a total of $1003kg$. What is the mean weight per sack?

Can you determine the median weight per sack from the given information? If so, explain how. If not, give two examples of datasets with the same total weight be different medians.

EXERCISE 1.6. For the dataset $\{6, -2, 6, 14, -3, 0, 1, 4, 3, 2, 5\}$, which we will call $DS_1$, find the mode(s), mean, and median.

Define $DS_2$ by adding 3 to each number in $DS_1$. What are the mode(s), mean, and median of $DS_2$?

Now define $DS_3$ by subtracting 6 from each number in $DS_1$. What are the mode(s), mean, and median of $DS_3$?

Next, define $DS_4$ by multiplying every number in $DS_1$ by 2. What are the mode(s), mean, and median of $DS_4$?

Looking at your answers to the above calculations, how do you think the mode(s), mean, and median of datasets must change when you add, subtract, multiply or divide all the numbers by the same constant? Make a specific conjecture!

EXERCISE 1.7. There is a very hard mathematics competition in which college students in the US and Canada can participate called the **William Lowell Putnam Mathematical Competition**. It consists of a six-hour long test with twelve problems, graded 0 to 10 on each problem, so the total score could be anything from 0 to 120.

The median score last year on the Putnam exam was 0 (as it often is, actually). What does this tell you about the scores of the students who took it? Be as precise as you can. Can you tell what fraction (percentage) of students had a certain score or scores? Can you figure out what the quartiles must be?

EXERCISE 1.8. Find the range, $IQR$, and standard deviation of the following sample dataset:

$$DS_1 = \{0, 0, 0, 0, 0, .5, 1, 1, 1, 1, 1\} \quad .$$

Now find the range, $IQR$, and standard deviation of the following sample data:

$$DS_2 = \{0, .5, 1, 1, 1, 1, 1, 1, 1, 1, 1\} \quad .$$

Next find the range, $IQR$, and standard deviation of the following sample data:

$$DS_3 = \{0, 0, 0, 0, 0, 0, 0, 0, 0, .5, 1\} \quad .$$

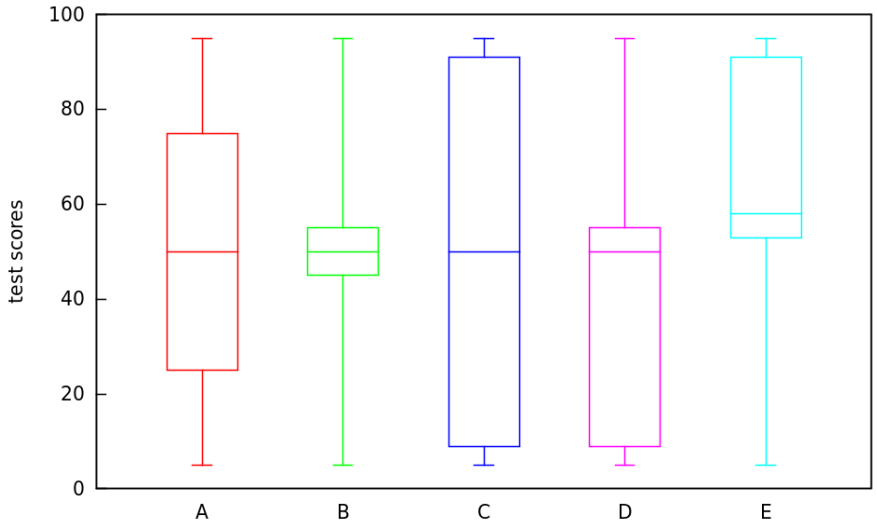
Finally, find the range, $IQR$, and standard deviation of sample data $DS_4$, consisting of 98 0s, one .5, and one 1 (so like $DS_3$ except with 0 occurring 98 times instead of 9 time).

EXERCISE 1.9. What must be true about a dataset if its range is 0? Give the most interesting example of a dataset with range of 0 and the property you just described that you can think of.

What must be true about a dataset if its $IQR$ is 0? Give the most interesting example of a dataset with $IQR$ of 0 and the property you just described that you can think of.

What must be true about a dataset if its standard deviation is 0? Give the most interesting example of a dataset with standard deviation of 0 and the property you just described that you can think of.

EXERCISE 1.10. Here are some boxplots of test scores, out of 100, on a standardized test given in five different classes – the same test, different classes. For each of these plots, $A - E$, describe qualitatively (in the sense of §1.3.4) but in as much detail as you can, what must have been the histogram for the data behind this boxplot. Also sketch a possible such histogram, for each case.



CHAPTER 2

# Bi-variate Statistics: Basics

## 2.1. Terminology: Explanatory/Response or Independent/Dependent

All of the discussion so far has been for studies which have a single variable. We may collect the values of this variable for a large population, or at least the largest sample we can afford to examine, and we may display the resulting data in a variety of graphical ways, and summarize it in a variety of numerical ways. But in the end all this work can only show a single characteristic of the individuals. If, instead, we want to study a *relationship*, we need to collect two (at least) variables and develop methods of descriptive statistics which show the relationships between the values of these variables.

Relationships in data require at least two variables. While more complex relationships can involve more, in this chapter we will start the project of understanding *bivariate data*, data where we make two observations for each individual, where we have exactly two variables.

If there is a relationship between the two variables we are studying, the most that we could hope for would be that that relationship is due to the fact that one of the variables *causes* the other. In this situation, we have special names for these variables

DEFINITION 2.1.1. In a situation with bivariate data, if one variable can take on any value without (significant) constraint it is called the **independent variable**, while the second variable, whose value is (at least partially) controlled by the first, is called the **dependent variable**.

Since the value of the dependent variable depends upon the value of the independent variable, we could also say that it is explained by the independent variable. Therefore the independent variable is also called the **explanatory variable** and the dependent variable is then called the **response variable**

Whenever we have bivariate data and we have made a choice of which variable will be the independent and which the dependent, we write $x$ for the independent and $y$ for the dependent variable.

EXAMPLE 2.1.2. Suppose we have a large warehouse of many different boxes of products ready to ship to clients. Perhaps we have packed all the products in boxes which are perfect cubes, because they are stronger and it is easier to stack them efficiently. We could do a study where

- the *individuals* would be the boxes of product;

- the *population* would be all the boxes in our warehouse;
- the *independent variable* would be, for a particular box, the length of its side in *cm*;
- the *dependent variable* would be, for a particular box, the cost to the customer of buying that item, in US dollars.

We might think that the size *determines* the cost, at least approximately, because the larger boxes contain larger products into which went more raw materials and more labor, so the items would be more expensive. So, at least roughly, the size may be anything, it is a free or *independent* choice, while the cost is (approximately) determined by the size, so the cost is *dependent*. Otherwise said, the size *explains* and the cost is the *response*. Hence the choice of those variables.

EXAMPLE 2.1.3. Suppose we have exactly the same scenario as above, but now we want to make the different choice where

- the *dependent variable* would be, for a particular box, the volume of that box.

There is one quite important difference between the two examples above: in one case (the cost), knowing the length of the side of a box give us a hint about how much it costs (bigger boxes cost more, smaller boxes cost less) but this knowledge is imperfect (sometimes a big box is cheap, sometimes a small box is expensive); while in the other case (the volume), knowing the length of the side of the box perfectly tells us the volume. In fact, there is a simple geometric formula that the volume $V$ of a cube of side length $s$ is given by $V = s^3$.

This motivates a last preliminary definition

DEFINITION 2.1.4. We say that the relationship between two variables is **deterministic** if knowing the value of one variable completely determines the value of the other. If, instead, knowing one value does not completely determine the other, we say the variables have a **non-deterministic relationship**.

## 2.2. Scatterplots

When we have bivariate data, the first thing we should always do is draw a graph of this data, to get some feeling about what the data is showing us and what statistical methods it makes sense to try to use. The way to do this is as follows

DEFINITION 2.2.1. Given bivariate quantitative data, we make the **scatterplot** of this data as follows: Draw an $x$- and a $y$-axis, and label them with descriptions of the independent and dependent variables, respectively. Then, for each individual in the dataset, put a dot on the graph at location $(x, y)$, if $x$ is the value of that individual's independent variable and $y$ the value of its dependent variable.

After making a scatterplot, we usually describe it qualitatively in three respects:

DEFINITION 2.2.2. If the cloud of data points in a scatterplot generally lies near some curve, we say that the scatterplot has [approximately] that **shape**.

A common shape we tend to find in scatterplots is that it is **linear**

If there is no visible shape, we say the scatterplot is **amorphous**, or **has no clear shape**.

DEFINITION 2.2.3. When a scatterplot has some visible shape – so that we do not describe it as amorphous – how close the cloud of data points is to that curve is called the **strength** of that association. In this context, a **strong** [linear, *e.g.,*] association means that the dots are close to the named curve [line, *e.g.,*], while a **weak** association means that the points do not lie particularly close to any of the named curves [line, *e.g.,*].

DEFINITION 2.2.4. In case a scatterplot has a fairly strong linear association, the **direction** of the association described whether the line is increasing or decreasing. We say the association is **positive** if the line is increasing and **negative** if it is decreasing.

[Note that the words *positive* and *negative* here can be thought of as describing the *slope* of the line which we are saying is the underlying relationship in the scatterplot.]

## 2.3. Correlation

As before (in §§1.4 and 1.5), when we moved from describing histograms with words (like *symmetric*) to describing them with numbers (like the *mean*), we now will build a numeric measure of the strength and direction of a linear association in a scatterplot.

DEFINITION 2.3.1. Given bivariate quantitative data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ the **[Pearson] correlation coefficient** of this dataset is

$$r = \frac{1}{n-1} \sum \frac{(x_i - \overline{x})}{s_x} \frac{(y_i - \overline{y})}{s_y}$$

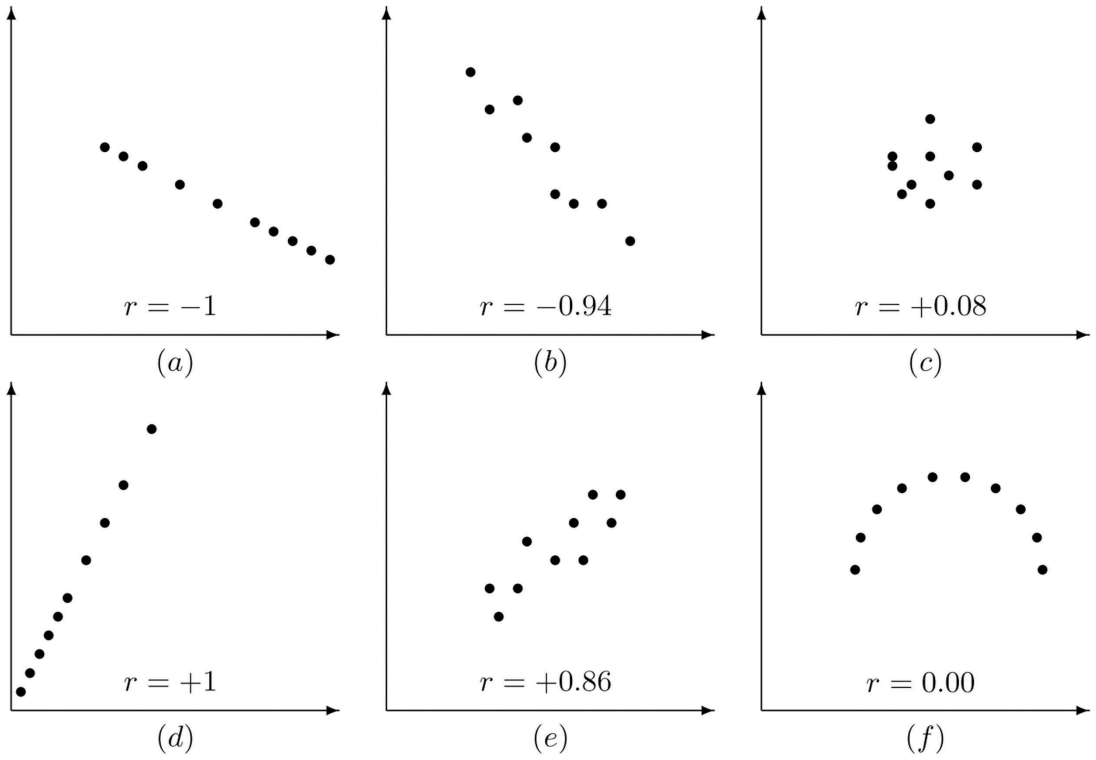
where $s_x$ and $s_y$ are the standard deviations of the $x$ and $y$, respectively, datasets by themselves.

We collect some basic information about the correlation coefficient in the following

FACT 2.3.2. For any bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with correlation coefficient $r$, we have

(1) $-1 \leq r \leq 1$ is always true;
(2) if $|r|$ is near $1$ – meaning that $r$ is near $\pm 1$ – then the linear association between $x$ and $y$ is *strong*
(3) if $r$ is near $0$ – meaning that $r$ is positive or negative, but near $0$ – then the linear association between $x$ and $y$ is *weak*
(4) if $r > 0$ then the linear association between $x$ and $y$ is positive, while if $r < 0$ then the linear association between $x$ and $y$ is negative
(5) $r$ is the same no matter what units are used for the variables $x$ and $y$ – meaning that if we change the units in either variable, $r$ will not change
(6) $r$ is the same no matter which variable is begin used as the explanatory and which as the response variable – meaning that if we switch the roles of the $x$ and the $y$ in our dataset, $r$ will not change.

It is also nice to have some examples of correlation coefficients, such as




Many electronic tools which compute the correlation coefficient $r$ of a dataset also report its square, $r^2$. There reason is explained in the following

FACT 2.3.3. If $r$ is the correlation coefficient between two variables $x$ and $y$ in some quantitative dataset, then its square $r^2$ it the fraction (often described as a percentage) of the variation of $y$ which is associated with variation in $x$.

EXAMPLE 2.3.4. If the square of the correlation coefficient between the independent variable *how many hours a week a student studies statistics* and the dependent variable *how many points the student gets on the statistics final exam* is $.64$, then $64\%$ of the variation in scores for that class is cause by variation in how much the students study. The remaining $36\%$ of the variation in scores is due to other random factors like whether a student was coming down with a cold on the day of the final, or happened to sleep poorly the night before the final because of neighbors having a party, or some other issues different just from studying time.

## Exercises

EXERCISE 2.1. Suppose you pick 50 random adults across the United States in January 2017 and measure how tall they are. For each of them, you also get accurate information about how tall their (biological) parents are. Now, using as your individuals these 50 adults and as the two variables their heights and the average of their parents' heights, make a sketch of what you think the resulting scatterplot would look like. Explain why you made the choice you did of one variable to be the explanatory and the other the response variable. Tell what are the shape, strength, and direction you see in this scatterplot, if it shows a deterministic or non-deterministic association, and why you think those conclusions would be true if you were to do this exercise with real data.

Is there any time or place other than right now in the United States where you think the data you would collect as above would result in a scatterplot that would look fairly different in some significant way? Explain!

EXERCISE 2.2. It actually turns out that it is not true that the more a person works, the more they produce ... at least not always. Data on workers in a wide variety of industries show that working more hours produces more of that business's product for a while, but then after too many hours of work, keeping on working makes for almost no additional production.

Describe how you might collect data to investigate this relationship, by telling what individuals, population, sample, and variables you would use. Then, assuming the truth of the above statement about what other research in this area has found, make an example of a scatterplot that you think might result from your suggested data collection.

EXERCISE 2.3. Make a scatterplot of the dataset consisting of the following pairs of measurements:

$$\{(8,16),(9,9),(10,4),(11,1),(12,0),(13,1),(14,4),(15,9),(16,16)\}.$$

You can do this quite easily by hand (there are only nine points!). Feel free to use an electronic device to make the plot for you, if you have one you know how to use, but copy the resulting picture into the homework you hand in, either by hand or cut-and-paste into an electronic version.

Describe the scatterplot, telling what are the shape, strength, and direction. What do you think would be the correlation coefficient of this dataset? As always, explain all of your reasoning!

CHAPTER 3

# Linear Regression

Quick review of equations for lines:

Recall the equation of a line is usually in the form $y = mx + b$, where $x$ and $y$ are variables and $m$ and $b$ are numbers. Some basic facts about lines:

- If you are given a number for $x$, you can plug it in to the equation $y = mx + b$ to get a number for $y$, which together give you a point with coordinates $(x, y)$ that is on the line.
- $m$ is the *slope*, which tells how much the line goes up (increasing $y$) for every unit you move over to the right (increasing $x$) – we often say that the value of the slope is $m = \frac{rise}{run}$. It can be
  - *positive*, if the line is tilted up,
  - *negative*, if the line is tilted down,
  - *zero*, if the line is horizontal, and
  - *undefined*, if the line is vertical.
- You can calculate the slope by finding the coordinates $(x_1, y_1)$ and $(x_2, y_2)$ of any two points on the line and then $m = \frac{y_2 - y_1}{x_2 - x_1}$.
- In particular, $x_2 - x_1 = 1$, then $m = \frac{y_2 - y_1}{1} = y_2 - y_1$ – so if you look at how much the line goes up in each step of one unit to the right, that number will be the slope $m$ (and if it goes *down*, the slope $m$ will simply be negative). In other words, the slope answers the question "for each step to the right, how much does the line increase (or decrease)?"
- $b$ is the $y$-*intercept*, which tells the $y$-coordinate of the point where the line crosses the $y$-axis. Another way of saying that is that $b$ is the $y$ value of the line when the $x$ is 0.

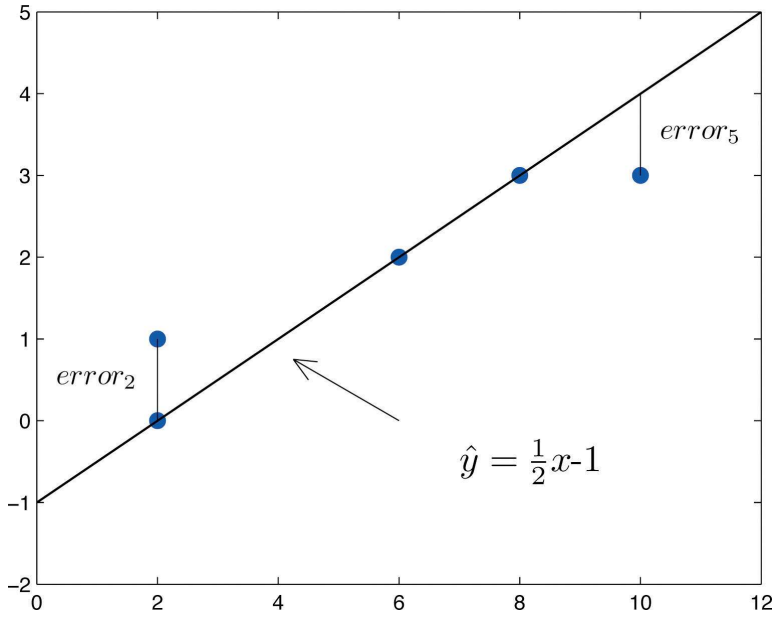
## 3.1. The Least Squares Regression Line

Suppose we have some bivariate quantitative data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ for which the correlation coefficient indicates some linear association. It is natural to want to write down explicitly the equation of the best line through the data – the question is what is this line. The most common meaning given to *best* in this search for the line is *the line whose total square error is the smallest possible.* We make this notion precise in two steps

DEFINITION 3.1.1. Given a bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and a candidate line $\widehat{y} = mx + b$ passing through this dataset, a **residual** is the difference in $y$-coordinates of an actual data point $(x_i, y_i)$ and the line's $y$ value at the same $x$-coordinate. That is, if the $y$-coordinate of the line when $x = x_i$ is $\widehat{y_i} = mx_i + b$, then the residual is the measure of error given by $error_i = y_i - \widehat{y_i}$.

Note we use the convention here and elsewhere of writing $\widehat{y}$ for the $y$-coordinate on an approximating line, while the plain $y$ variable is left for actual data values, like $y_i$.

Here is an example of what residuals look like




Now we are in the position to state the

DEFINITION 3.1.2. Given a bivariate quantitative dataset the **least square regression line**, almost always abbreviated to **LSRL**, is the line for which the sum of the squares of the residuals is the smallest possible.

FACT 3.1.3. If a bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ has LSRL given by $\widehat{y} = mx + b$, then

(1) The slope of the LSRL is given by $m = r\frac{s_y}{s_x}$, where $r$ is the correlation coefficient of the dataset.
(2) The LSRL passes through the point $(\overline{x}, \overline{y})$.
(3) It follows that the $y$-intercept of the LSRL is given by $b = \overline{y} - \overline{x}\, m = \overline{y} - \overline{x}\, r\, \frac{s_y}{s_x}$.

It is possible to find the (coefficients of the) LSRL using the above information, but it is often more convenient to use a calculator or other electronic tool. Such tools also make it very easy to graph the LSRL right on top of the scatterplot – although it is often fairly easy to sketch what the LSRL will likely look like by just making a good guess, using

visual intuition, if the linear association is strong (as will be indicated by the correlation coefficient).

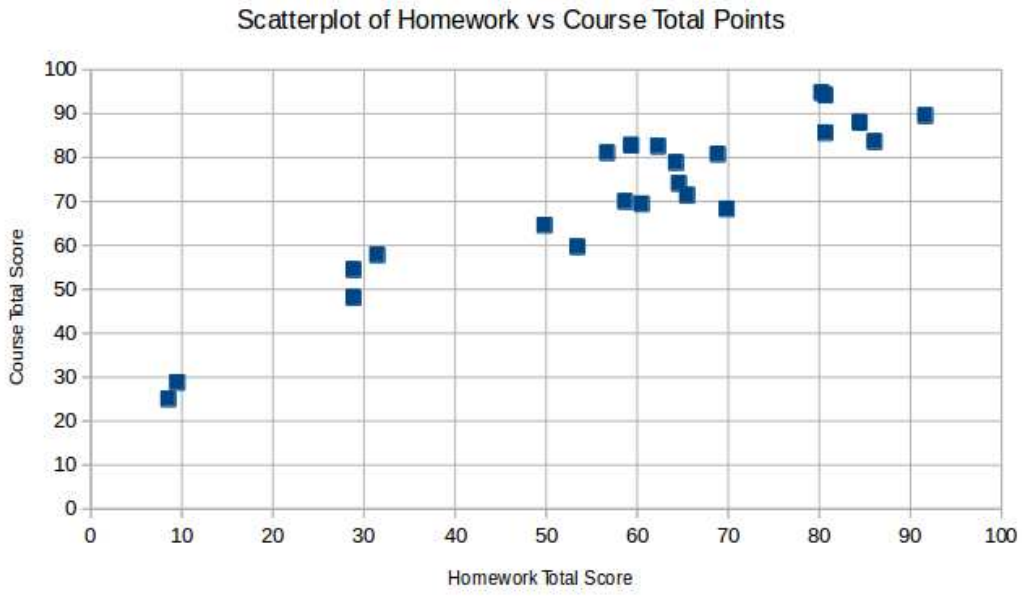
EXAMPLE 3.1.4. Here is some data where the individuals are 23 students in a statistics class, the independent variable is the students' total score on their homeworks, while the dependent variable is their final total course points, both out of 100.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ : | 65 | 65 | 50 | 53 | 59 | 92 | 86 | 84 | 29 |
| $y$ : | 74 | 71 | 65 | 60 | 83 | 90 | 84 | 88 | 48 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ : | 29 | 9 | 64 | 31 | 69 | 10 | 57 | 81 | 81 |
| $y$ : | 54 | 25 | 79 | 58 | 81 | 29 | 81 | 94 | 86 |

| | | | | | |
|---|---|---|---|---|---|
| $x$ : | 80 | 70 | 60 | 62 | 59 |
| $y$ : | 95 | 68 | 69 | 83 | 70 |

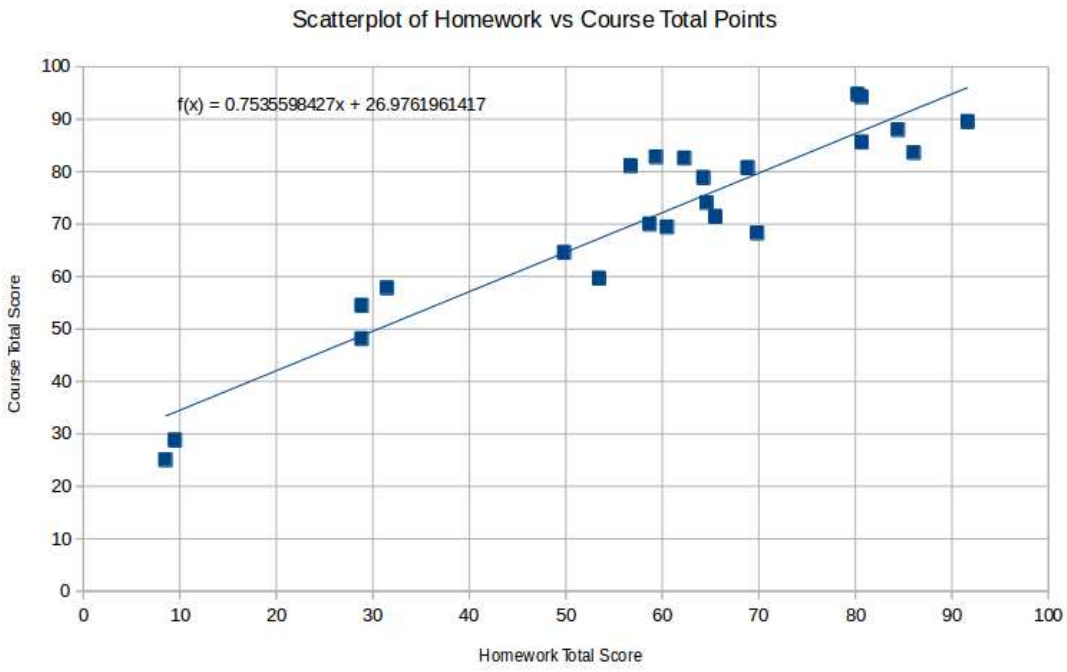
Here is the resulting scatterplot, made with **LibreOffice Calc**(a free equivalent of **Microsoft Excel**)




It seems pretty clear that there is quite a strong linear association between these two variables, as is born out by the correlation coefficient, $r = .935$ (computed with **LibreOffice Calc**'s `CORREL`). Using then `STDEV.S` and `AVERAGE`, we find that the coefficients of the LSRL for this data, $\widehat{y} = mx + b$ are

$$m = r\frac{s_y}{s_x} = .935\frac{18.701}{23.207} = .754 \qquad \text{and} \qquad b = \overline{y} - \overline{x}\,m = 71 - 58 \cdot .754 = 26.976$$

We can also use **LibreOffice Calc**'s `Insert Trend Line`, with `Show Equation`, to get all this done automatically. Note that when **LibreOffice Calc** writes the equation of the LSRL, it uses $f(x)$ in place of $\widehat{y}$, as we would.



## 3.2. Applications and Interpretations of LSRLs

Suppose that we have a bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we have computed its correlation coefficient $r$ and (the coefficients of) its LSRL $\widehat{y} = mx + b$. What is this information good for?

The main use of the LSRL is described in the following

DEFINITION 3.2.1. Given a bivariate quantitative dataset and associated LSRL with equation $\widehat{y} = mx + b$, the process of guessing that the value of the dependent variable in this relationship to have the value $mx_0 + b$, for $x_0$ any value for the independent variable which satisfies $x_{min} \leq x_0 \leq x_{max}$, is called **interpolation**.

The idea of interpolation is that we think the LSRL describes as well as possible the relationship between the independent and dependent variables, so that if we have a new $x$ value, we'll use the LSRL equation to predict what would be our best guess of what would be the corresponding $y$. Note we might have a new value of $x$ because we simply lost part of our dataset and are trying to fill it in as best we can. Another reason might be that a new individual came along whose value of the independent variable, $x_0$, was typical of the rest of the dataset – so the the very least $x_{min} \leq x_0 \leq x_{max}$ – and we want to guess what will be the value of the dependent variable for this individual before we measure it. (Or maybe we cannot measure it for some reason.)

A common (but naive) alternate approach to interpolation for a value $x_0$ as above might be to find two values $x_i$ and $x_j$ in the dataset which were as close to $x_0$ as possible, and on either side of it (so $x_i < x_0 < x_j$), and simply to guess that the $y$-value for $x_0$ would be the average of $y_i$ and $y_j$. This is not a terrible idea, but it is not as effective as using the LSRL as described above, since we use the entire dataset when we build the coefficients of the LSRL. So the LSRL will give, by the process of interpolation, the best guess for what should be that missing $y$-value based on everything we know, while the "average of $y_i$ and $y_j$" method only pays attention to those two nearest data points and thus may give a very bad guess for the corresponding $y$-value if those two points are not perfectly typical, if they have any randomness, any variation in their $y$-values which is not due to the variation of the $x$.

It is thus always best to use interpolation as described above.

EXAMPLE 3.2.2. Working with the statistics students' homework and total course points data from Example 3.1.4, suppose the gradebook of the course instructor was somewhat corrupted and the instructor lost the final course points of the student Janet. If Janet's homework points of 77 were not in the corrupted part of the gradebook, the instructor might use interpolation to guess what Janet's total course point probably were. To do this, the instructor would have plugged in $x = 77$ into the equation of the LSRL, $\widehat{y} = mx + b$ to get the estimated total course points of $.754 \cdot 77 + 26.976 = 85.034$.

Another important use of the (coefficients of the) LSRL is to use the underlying meanings of the slope and $y$-intercept. For this, recall that in the equation $y = mx + b$, the slope $m$ tells us how much the line goes up (or down, if the slope is negative) for each increase of the $x$ by one unit, while the $y$-intercept $b$ tells us what would be the $y$ value where the line crosses the $y$-axis, so when the $x$ has the value 0. In each particular situation that we have bivariate quantitative data and compute an LSRL, we can then use these interpretations to make statements about the relationship between the independent and dependent variables.

EXAMPLE 3.2.3. Look one more time at the data on students' homework and total course points in a statistics class from Example 3.1.4, and the the LSRL computed there. We said that the slope of the LSRL was $m = .754$ and the $y$-intercept was $b = 26.976$. In context, what this means, is that *On average, each additional point of homework corresponded to an increase of* $.754$ *total course points.* We may hope that this is actually a causal relationship, that the extra work a student does to earn that additional point of homework score helps the student learn more statistics and therefore get $.75$ more total course points. But the mathematics here does not require that causation, it merely tells us the increase in $x$ is *associated* with that much increase in $y$.

Likewise, we can also conclude from the LSRL that *In general, a student who did no homework at all would earn about* $26.976$ *total course points.* Again, we cannot conclude that doing no homework *causes* that terrible final course point total, only that there is an association.

## 3.3. Cautions

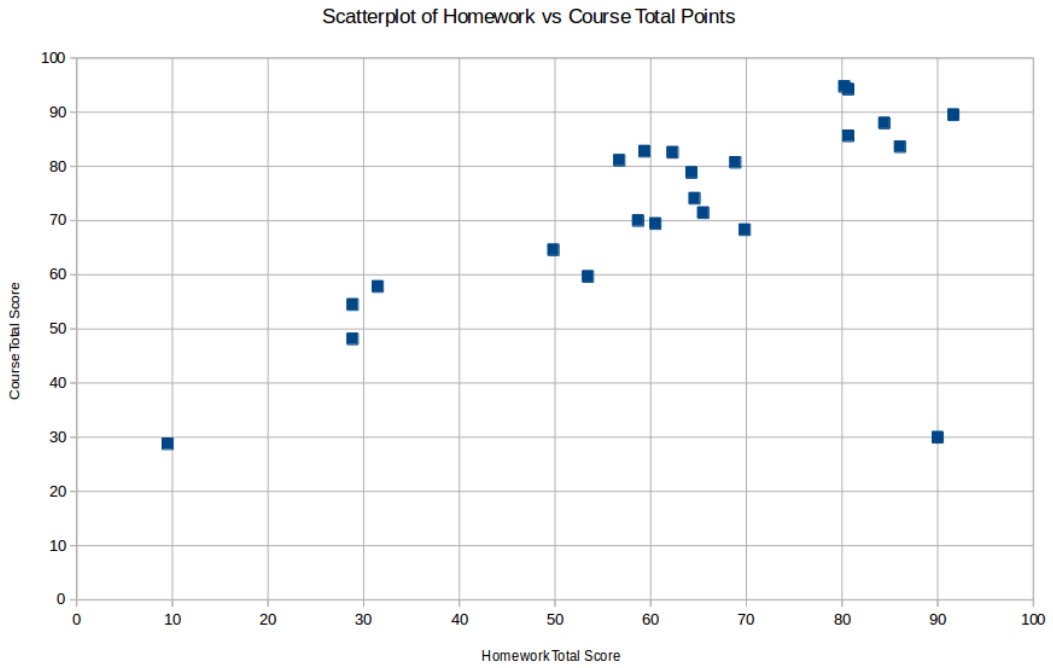
**3.3.1. Sensitivity to Outliers.** The correlation coefficient and the (coefficients of the) LSRL are built out of means and standard deviations and therefore the following fact is completely unsurprising

FACT 3.3.1. The correlation coefficient and the (coefficients of the) LSRL are very sensitive to outliers.

What perhaps is surprising here is that the outliers for bivariate data are a little different from those for 1-variable data.

DEFINITION 3.3.2. An **outlier** for a bivariate quantitative dataset is one which is far away from the curve which has been identified as underlying the shape of the scatterplot of that data. In particular, a point $(x, y)$ can be a bivariate outlier even if both $x$ is not an outlier for the independent variable data considered alone and $y$ is not an outlier for the dependent variable data alone.

EXAMPLE 3.3.3. Suppose we add one more point $(90, 30)$ to the dataset in Example 3.1.4. Neither the $x$- nor $y$-coordinates of this point are outliers with respect to their respective single-coordinate datasets, but it is nevertheless clearly a bivariate outlier, as can be seen in the new scatterplot




In fact recomputing the correlation coefficient and LSRL, we find quite a change from what we found before, in Example 3.1.4:

$$r = .704 \qquad \textit{[which used to be } .935\textit{]}$$

and

$$\widehat{y} = .529x + 38.458 \qquad \textit{[which used to be } .754x + 26.976\textit{]}$$

all because of one additional point!

**3.3.2. Causation.** The attentive reader will have noticed that we started our discussion of bivariate data by saying we hoped to study when one thing *causes* another. However, what we've actually done instead is find *correlation* between variables, which is quite a different thing.

Now philosophers have discussed what exactly causation *is* for millennia, so certainly it is a subtle issue that we will not resolve here. In fact, careful statisticians usually dodge the complexities by talking about *relationships*, *association*, and, of course, the *correlation coefficient*, being careful always not to commit to *causation* – at least based only on an analysis of the statistical data.

As just one example, where we spoke about the meaning of the square $r^2$ of the correlation coefficient (we called it Fact 2.3.3), we were careful to say that $r^2$ measures the variation of the dependent variable which is *associated* with the variation of the independent variable. A more reckless description would have been to say that one *caused* the other – but don't fall into that trap!

This would be a bad idea because (among other reasons) the correlation coefficient is symmetric in the choice of explanatory and response variables (meaning $r$ is the same no matter which is chosen for which role), while any reasonable notion of causation is asymmetric. *E.g.,* while the correlation is exactly the same very large value with either variable being $x$ and which $y$, most people would say that *smoking causes cancer* and not the other way[1]!

We do need to make one caution about this caution, however. If there is a causal relationship between two variables that are being studied carefully, then there will be correlation. So, to quote the great data scientist Edward Tufte [**Tuf06**],

> *Correlation is not causation but it sure is a hint.*

The first part of this quote (up to the "but") is much more famous and, as a very first step, is a good slogan to live by. Those with a bit more statistical sophistication might instead learn this version, though. A more sophisticated-sounding version, again due to Tufte [**Tuf06**], is

> *Empirically observed covariation is a necessary but not sufficient condition for causality.*

[1]Although in the 1950s a doctor (who later was found to be in the pay of the tobacco industry) did say that the clear statistical evidence of association between smoking and cancer might be a sign that cancer causes smoking (I know: crazy!). His theory was that people who have lung tissue which is more prone to developing cancer are more likely to start smoking because somehow the smoke makes that particular tissue feel better. Needless to say, this is not the accepted medical view, because lots of evidence goes against it.

**3.3.3. Extrapolation.** We have said that visual intuition often allows humans to sketch fairly good approximations of the LSRL on a scatterplot, so long as the correlation coefficient tells us there is a strong linear association. If the diligent reader did that with the first scatterplot in Example 3.1.4, probably the resulting line looked much like the line which **LibreOffice Calc** produced – except humans usually sketch their line all the way to the left and right edges of the graphics box. Automatic tools like **LibreOffice Calc** do not do that, for a reason.

DEFINITION 3.3.4. Given a bivariate quantitative dataset and associated LSRL with equation $\widehat{y} = mx + b$, the process of guessing that the value of the dependent variable in this relationship to have the value $mx_0 + b$, for $x_0$ any value for the independent variable which *does not satisfy* $x_{min} \leq x_0 \leq x_{max}$ [so, instead, either $x_0 < x_{min}$ or $x_0 > x_{max}$], is called **extrapolation**.

Extrapolation is considered a bad, or at least risky, practice. The idea is that we used the evidence in the dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ to build the LSRL, but, by definition, all of this data lies in the interval on the $x$-axis from $x_{min}$ to $x_{max}$. There is literally no evidence from this dataset about what the relationship between our chosen explanatory and response variables will be for $x$ outside of this interval. So in the absence of strong reasons to believe that the precise linear relationship described by the LSRL will continue for more $x$'s, we should not assume that it does, and therefore we should not use the LSRL equation to guess values by extrapolation.

The fact is, however, that often the best thing we can do with available information when we want to make predictions out into uncharted territory on the $x$-axis is extrapolation. So while it is perilous, it is reasonable to extrapolate, so long as you are clear about what exactly you are doing.

EXAMPLE 3.3.5. Using again the statistics students' homework and total course points data from Example 3.1.4, suppose the course instructor wanted to predict what would be the total course points for a student who had earned a perfect $100$ points on their homework. Plugging into the LSRL, this would have yielded a guess of $.754 \cdot 100 + 26.976 = 102.376$. Of course, this would have been impossible, since the maximum possible total course score was $100$. Moreover, making this guess is an example of extrapolation, since the $x$ value of $100$ is beyond the largest $x$ value of $x_{max} = 92$ in the dataset. Therefore we should not rely on this guess – as makes sense, since it is invalid by virtue of being larger than $100$.

**3.3.4. Simpson's Paradox.** Our last caution is not so much a way using the LSRL can go wrong, but instead a warning to be ready for something very counter-intuitive to happen – so counter-intuitive, in fact, that it is called a paradox.

It usually seems reasonable that if some object is cut into two pieces, both of which have a certain property, then probably the whole object also has that same property. But

if the object in question is *a population* and the property is *has positive correlation*, then maybe the unreasonable thing happens.

DEFINITION 3.3.6. Suppose we have a population for which we have a bivariate quantitative dataset. Suppose further that the population is broken into two (or more) subpopulations for all of which the correlation between the two variables is *positive*, but the correlation of the variables for the whole dataset is *negative*. Then this situation is called **Simpson's Paradox**. [It's also called Simpson's Paradox if the role of *positive* and *negative* is reversed in our assumptions.]

The bad news is that Simpson's paradox can happen.

EXAMPLE 3.3.7. Let $\mathcal{P} = \{(0,1),(1,0),(9,10),(10,9)\}$ be a bivariate dataset, which is broken into the two subpopulations $\mathcal{P}_1 = \{(0,1),(1,0)\}$ and $\mathcal{P}_2 = \{(9,10),(10,9)\}$. Then the correlation coefficients of both $\mathcal{P}_1$ and $\mathcal{P}_2$ are $r = -1$, but the correlation of all of $\mathcal{P}$ is $r = .9756$. This is Simpson's Paradox!

Or, in applications, we can have situations like

EXAMPLE 3.3.8. Suppose we collect data on two sections of a statistics course, in particular on how many hours per work the individual students study for the course and how they do in the course, measured by their total course points at the end of the semester. It is possible that there is a strong positive correlation between these variables for each section by itself, but there is a strong negative correlation when we put all the students into one dataset. In other words, it is possible that the rational advice, based on both individual sections, is *study more and you will do better in the course*, but that the rational advice based on all the student data put together is *study less and you will do better*.

## Exercises

EXERCISE 3.1. The age ($x$) and resting heart rate (RHR, $y$) were measured for nine men, yielding this dataset:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ : | 20 | 23 | 30 | 37 | 35 | 45 | 51 | 60 | 63 |
| $y$ : | 72 | 71 | 73 | 74 | 74 | 73 | 75 | 75 | 77 |

Make a scatterplot of these data.

Based on the scatterplot, what do you think the correlation coefficient $r$ will be?

Now compute $r$.

Compute the LSRL for these data, write down its equation, and sketch it on top of your scatterplot.

*[You may, of course, do as much of this with electronic tools as you like. However, you should explain what tool you are using, how you used it, and what it must have been doing behind the scenes to get the results which it displayed and you are turning in.]*

EXERCISE 3.2. Continuing with the data and computations of the previous problem:

What percentage of the variation in RHR is associated with variation in age?

Write the following sentences with blanks filled in: "If I measured the RHR of a 55 year-old man, I would expect it to be _____. Making an estimate like this is called _____."

Just looking at the equation of the LSRL, what does it suggest should be the RHR of a newborn baby? Explain.

Also explain what an estimate like yours for the RHR of a baby is called. This kind of estimate is considered a bad idea in many cases – explain why in general, and also use specifics from this particular case.

EXERCISE 3.3. Write down a bivariate quantitative dataset for a population of only two individuals whose LSRL is $\widehat{y} = 2x - 1$.

What is the correlation coefficient of your dataset?

Next, add one more point to the dataset in such a way that you don't change the LSRL or correlation coefficient.

Finally, can your find a dataset with the same LSRL but having a larger correlation coefficient than you just had?

*[Hint: fool around with modifications or additions to the datasets in you already found in this problem, using an electronic tool to do all the computational work. When you find a good one, write it down and explain what you thinking was as you searched for it.]*

# Part 2

# Good Data

It is something of an aphorism among statisticians that

*The plural of anecdote is not data.*[2]

The distinction being emphasized here is between the information we might get from a personal experience or a friend's funny story – an anecdote – and the cold, hard, objective information on which we want to base our scientific investigations of the world – data.

In this Part, our goal is to discuss aspects of getting good data. It may seem counter-intuitive, but the first step in that direction is to develop some of the foundations of *probability theory*, the mathematical study of systems which are non-deterministic – random – but in a consistent way. The reason for this is that the easiest and most reliable way to ensure objectivity in data, to suppress personal choices which may result in biased information from which we cannot draw universal, scientific conclusions, is to collect your data *randomly*. Randomness is a tool which the scientist introduces intentionally and carefully, as barrier against bias, in the collection of high quality data. But this strategy only works if we can understand how to extract precise information even in the presence of randomness – hence the importance of studying probability theory.

After a chapter on probability, we move on to a discussion of some fundamentals of *experimental design* – starting, not surprisingly, with *randomization*, but finishing with the gold standard for experiments (on humans, at least): *randomized, placebo-controlled, double-blind experiments [RCTs]*. Experiments whose subjects are not humans share some, but not all, of these design goals

It turns out that, historically, a number of experiments with human subjects have had very questionable moral foundations, so it is very important to stop, as we do in the last chapter of this Part, to build a outline of *experimental ethics*.

[2]It is hard to be certain of the true origins of this phrase. The political scientist Raymond Wolfinger is sometimes given credit [**PB**] – for a version *without the* "not," actually. Sometime later, then, it became widespread with the "not."

CHAPTER 4

# Probability Theory

We want to imagine doing an experiment in which there is no way to predict what the outcome will be. Of course, if we stop our imagination there, there would be nothing we could say and no point in trying to do any further analysis: the outcome would just be whatever it wanted to be, with no pattern.

So let us add the additional assumption that while we *cannot predict* what will happen any particular time we do the experiment, we *can predict* general trends, in the long run, if we repeat the experiment many times. To be more precise, we assume that, for any collection $E$ of possible outcomes of the experiment there is a number $p(E)$ such that, no matter who does the experiment, no matter when they do it, if they repeat the experiment many times, the fraction of times they would have seen any of the outcomes of $E$ would be close to that number $p(E)$.

This is called the *frequentist* approach to the idea of probability. While it is not universally accepted – the *Bayesian* alternative does in fact have many adherents – it has the virtue of being the most internally consistent way of building a foundation for probability. For that reason, we will follow the frequentist description of probability in this text.

Before we jump into the mathematical formalities, we should motivate two pieces of what we just said. First, why talk about *sets* of outcomes of the experiment instead of talking about individual outcomes? The answer is that we are often interested in sets of outcomes, as we shall see later in this book, so it is nice to set up the machinery from the very start to work with such sets. Or, to give a particular concrete example, suppose you were playing a game of cards and could see your hand but not the other players' hands. You might be very interested in how likely is it that your hand is a winning hand, *i.e.*, what is the likelihood of the set of all possible configurations of all the rest of the cards in the deck and in your opponents' hands for which what you have will be the winning hand? It is situations like this which motivate an approach based on *sets* of outcomes of the random experiment.

Another question we might ask is: where does our uncertainty about the experimental results come from? From the beginnings of the scientific method through the turn of the $20^{th}$ century, it was thought that this uncertainty came from our incomplete knowledge of the system on which we were experimenting. So if the experiment was, say, flipping a coin, the precise amount of force used to propel the coin up into the air, the precise angular motion imparted to the coin by its position just so on the thumbnail of the person doing

the flipping, the precise drag that the coin felt as it tumbled through the air caused in part by eddies in the air currents coming from the flap of a butterfly's wings in the Amazon rainforest – all of these things could significantly contribute to changing whether the coin would eventually come up *heads* or *tails*. Unless the coin-flipper was a robot operating in a vacuum, then, there would just be no way to know all of these physical details with enough accuracy to predict the toss.

After the turn of the $20^{th}$ century, matters got even worse (at least for physical determinists): a new theory of physics came along then, called *Quantum Mechanics*, according to which true randomness is built into the laws of the universe. For example, if you have a very dim light source, which produces the absolutely smallest possible "chunks" of light (called *photons*), and you shine it through first one polarizing filter and then see if it goes through a second filter at a $45^{\circ}$ angle to the first, then half the photons will get through the second filter, but there is *absolutely no way ever to predict whether any particular photon will get though or not*. Quantum mechanics is full of very weird, non-intuitive ideas, but it is one of the most well-tested theories in the history of science, and it has passed every test.

## 4.1. Definitions for Probability

**4.1.1. Sample Spaces, Set Operations, and Probability Models.** Let's get right to the definitions.

DEFINITION 4.1.1. Suppose we have a repeatable experiment we want to investigate probabilistically. The things that happen when we do the experiment, the results of running it, are called the [**experimental**] **outcomes**. The set of all outcomes is called the **sample space** of the experiment. We almost always use the symbol $S$ for this sample space.

EXAMPLE 4.1.2. Suppose the experiment we are doing is "flip a coin." Then the sample space would be $S = \{H, T\}$.

EXAMPLE 4.1.3. For the experiment "roll a [normal, six-sided] die," the sample space would be $S = \{1, 2, 3, 4, 5, 6\}$.

EXAMPLE 4.1.4. For the experiment "roll two dice," the sample space would be

$$\begin{aligned} S = \{&11, 12, 13, 14, 15, 16, \\ &21, 22, 23, 24, 25, 26 \\ &31, 23, 33, 34, 35, 36 \\ &41, 42, 43, 44, 45, 46 \\ &51, 52, 53, 54, 55, 56 \\ &61, 62, 63, 64, 65, 66 \end{aligned}$$

where the notation "$nm$" means "$1^{st}$ roll resulted in an $n$, $2^{nd}$ in an $m$."

EXAMPLE 4.1.5. Consider the experiment "flip a coin as many times as necessary to see the first *Head*." This would have the infinite sample space

$$S = \{H, TH, TTH, TTTH, TTTTH, \dots\} \quad .$$

EXAMPLE 4.1.6. Finally, suppose the experiment is "point a Geiger counter at a lump of radioactive material and see how long you have to wait until the next click." Then the sample space $S$ is the set of all positive real numbers, because potentially the waiting time could be any positive amount of time.

As mentioned in the chapter introduction, we are more interested in

DEFINITION 4.1.7. Given a repeatable experiment with sample space $S$, an **event** is any collection of [some, all, or none of the] outcomes in $S$; *i.e.,* an event is any **subset** $E$ of $S$, written $E \subset S$.

There is one special set which is a subset of any other set, and therefore is an event in any sample space.

DEFINITION 4.1.8. The set $\{\}$ with no elements is called the **empty set**, for which we use the notation $\emptyset$.

EXAMPLE 4.1.9. Looking at the sample space $S = \{H, T\}$ in Example 4.1.2, it's pretty clear that the following are all the subsets of $S$:

$$\emptyset$$
$$\{H\}$$
$$\{T\}$$
$$S\ [= \{H, T\}]$$

Two parts of that example are always true: $\emptyset$ and $S$ are always subsets of any set $S$.

Since we are going to be working a lot with events, which are subsets of a larger set, the sample space, it is nice to have a few basic terms from set theory:

DEFINITION 4.1.10. Given a subset $E \subset S$ of a larger set $S$, the **complement of** $E$, is the set $E^c = \{\text{all the elements of } S \text{ which } \textit{are not} \text{ in } E\}$.

If we describe an event $E$ in words as all outcomes satisfies some property $X$, the complementary event, consisting of all the outcomes not in $E$, can be described as all outcomes which *don't* satisfy $X$. In other words, we often describe the event $E^c$ as the event "**not** $E$."

DEFINITION 4.1.11. Given two sets $A$ and $B$, their **union** is the set

$$A \cup B = \{\text{all elements which are in } A \text{ or } B \text{ [or both]}\}\ .$$

Now if event $A$ is those outcomes having property $X$ and $B$ is those with property $Y$, the event $A \cup B$, with all outcomes in $A$ together with all outcomes in $B$ can be described as all outcomes satisfying $X$ or $Y$, thus we sometimes pronounce the event "$A \cup B$" as "$A$ **or** $B$."

DEFINITION 4.1.12. Given two sets $A$ and $B$, their **intersection** is the set

$$A \cap B = \{\text{all elements which are in both } A \text{ and } B\}\ .$$

If, as before, event $A$ consists of those outcomes having property $X$ and $B$ is those with property $Y$, the event $A \cap B$ will consist of those outcomes which satisfy both $X$ and $Y$. In other words, "$A \cap B$" can be described as "$A$ **and** $B$."

Putting together the idea of intersection with the idea of that special subset $\emptyset$ of any set, we get the

DEFINITION 4.1.13. Two sets $A$ and $B$ are called **disjoint** if $A \cap B = \emptyset$. In other words, sets are disjoint if they have nothing in common.

A exact synonym for disjoint that some authors prefer is **mutually exclusive**. We will use both terms interchangeably in this book.

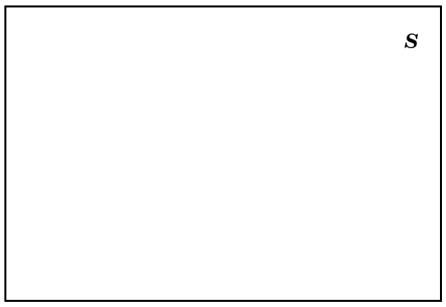
Now we are ready for the basic structure of probability.

DEFINITION 4.1.14. Given a sample space $S$, a **probability model** on $S$ is a choice of a real number $P(E)$ for every event $E \subset S$ which satisfies

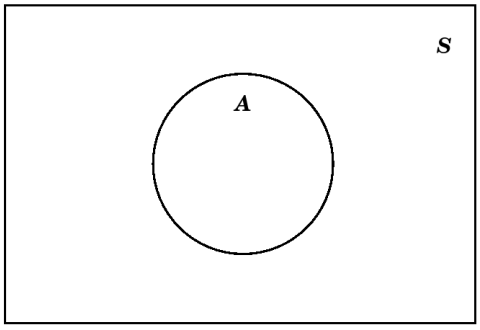
(1) For all events $E$, $0 \leq P(E) \leq 1$.
(2) $P(\emptyset) = 1$ and $P(S) = 1$.
(3) For all events $E$, $P(E^c) = 1 - P(E)$.
(4) If $A$ and $B$ are any two *disjoint* events, then $P(A \cup B) = P(A) + P(B)$. [This is called the **addition rule for disjoint events**.]

**4.1.2. Venn Diagrams.** Venn diagrams are a simple way to display subsets of a fixed set and to show the relationships between these subsets and even the results of various set operations (like *complement*, *union*, and *intersection*) on them. The primary use we will make of Venn diagrams is for events in a certain sample space, so we will use that terminology [even though the technique has much wider application].

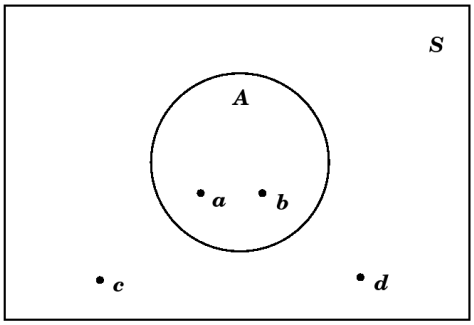
To make a Venn Diagram, *always start out by making a rectangle to represent the whole sample space*:



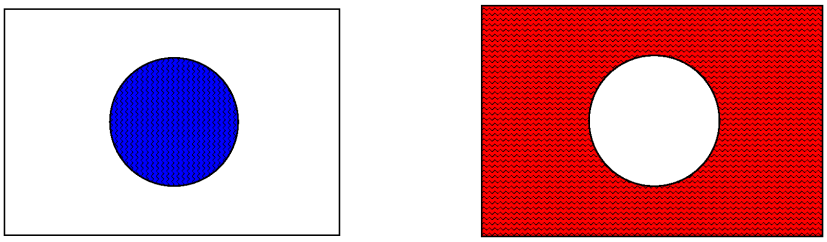

Within that rectangle, we make circles, ovals, or just blobs, to indicate that portion of the sample space which is some event $E$:




Sometimes, if the outcomes in the sample space $S$ and in the event $A$ might be indicated in the different parts of the Venn diagram. So, if $S = \{a, b, c, d\}$ and $A = \{a, b\} \subset S$, we might draw this as



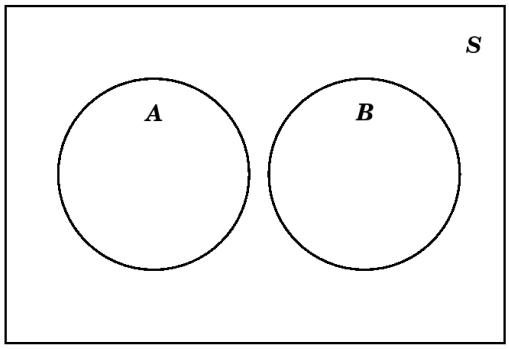

The *complement* $E^c$ of an event $E$ is easy to show on a Venn diagram, since it is simply everything which is not in $E$:



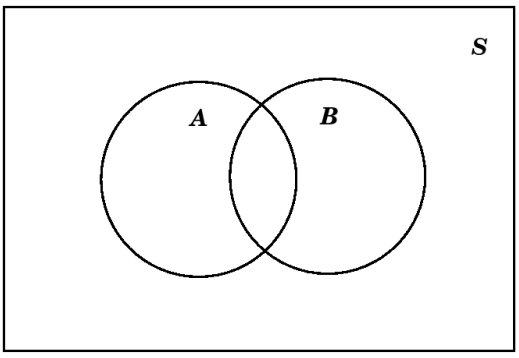
If the filled part here is $E$ ... then the filled part here is $E^c$

This can actually be helpful in figuring out what must be in $E^c$. In the example above with $S = \{a, b, c, d\}$ and $A = \{a, b\} \subset S$, by looking at what is in the shaded exterior part for our picture of $E^c$, we can see that for that $A$, we would get $A^c = \{c, d\}$.

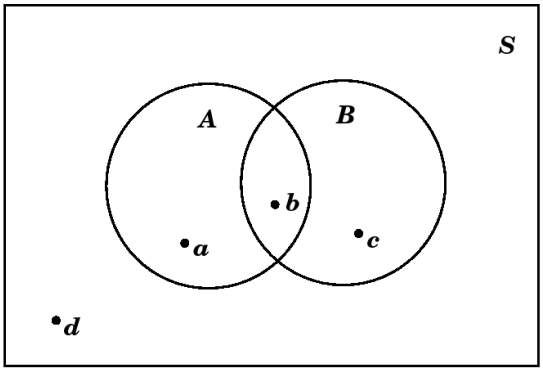
Moving now to set operations that work with two events, suppose we want to make a Venn diagram with events $A$ and $B$. If we know these events are disjoint, then we would make the diagram as follows:




while if they are known not to be disjoint, we would use instead this diagram:

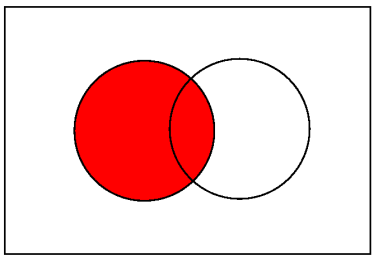


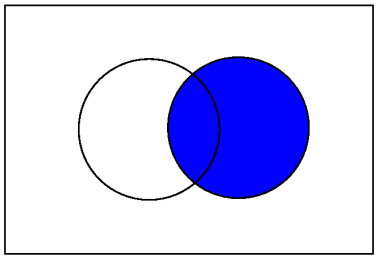

For example, it $S = \{a, b, c, d\}$, $A = \{a, b\}$, and $B = \{b, c\}$, we would have




When in doubt, it is probably best to use the version with overlap, which then could simply not have any points in it (or could have zero probability, when we get to that, below).

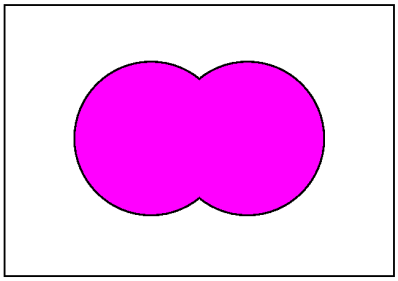
Venn diagrams are very good at showing unions, and intersection:

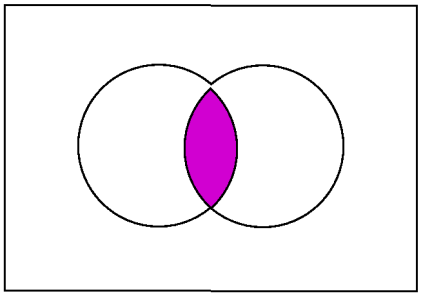




If the filled part here is $A$ and the filled part here is $B$

then

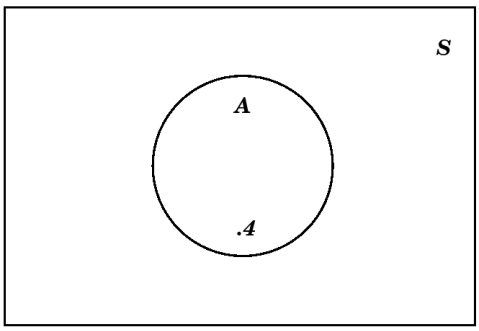




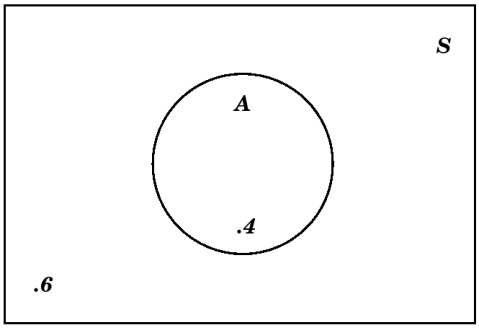
the filled part here is $A \cup B$ and the filled part here is $A \cap B$

Another nice thing to do with Venn diagrams is to use them as a visual aid for probability computations. The basic idea is to make a diagram showing the various events sitting inside the usual rectangle, which stands for the sample space, and to put numbers in various parts of the diagram showing the probabilities of those events, or of the results of operations (unions, intersection, and complement) on those events.

For example, if we are told that an event $A$ has probability $P(A) = .4$, then we can immediately fill in the $.4$ as follows:



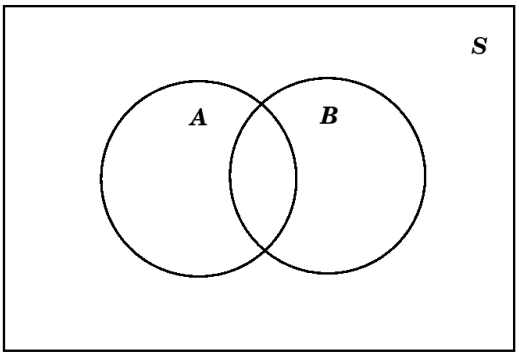
But we can also put a number in the exterior of that circle which represents $A$, taking advantage of the fact that that exterior is $A^c$ and the rule for probabilities of complements (point (3) in Definition 4.1.14) to conclude that the appropriate number is $1 - .4 = .6$:




We recommend that, in a Venn diagram showing probability values, *you always put a number in the region exterior to all of the events [but inside the rectangle indicating the sample space, of course]*.

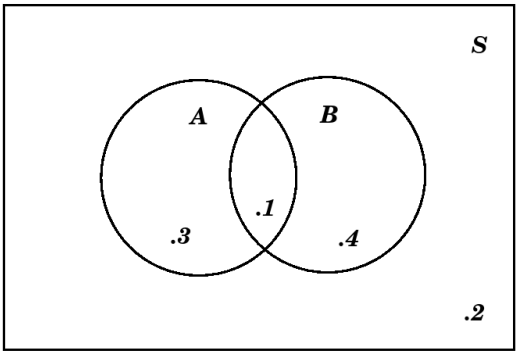
Complicating a little this process of putting probability numbers in the regions of a Venn diagram is the situation where we are giving for both an event and a subsetsubset, $\subset$ of that event. This most often happens when we are told probabilities both of some events and of their intersection(s). Here is an example:

EXAMPLE 4.1.15. Suppose we are told that we have two events $A$ and $B$ in the sample space $S$, which satisfy $P(A) = .4$, $P(B) = .5$, and $P(A \cap B) = .1$. First of all, we know that $A$ and $B$ are not disjoint, since if they were disjoint, that would mean (by definition) that $A \cap B = \emptyset$, and since $P(\emptyset) = 0$ but $P(A \cap B) \neq 0$, that is not possible. So we draw a Venn diagram that we've see before:




However, it would be unwise simply to write those given numbers .4, .5, and .1 into the three central regions of this diagram. The reason is that the number .1 is the probability of

$A \cap B$, which is a part of $A$ already, so if we simply write $.4$ in the rest of $A$, we would be counting that $.1$ for the $A \cap B$ twice. Therefore, before we write a number in the rest of $A$, outside of $A \cap B$, we have to subtract the $.1$ for $P(A \cap B)$. That means that the number which goes in the rest of $A$ should be $.4 - .1 = .3$. A similar reasoning tells us that the number in the part of $B$ outside of $A \cap B$, should be $.5 - .1 = .4$. That means the Venn diagram with all probabilities written in would be:




The approach in the above example is our second important recommendation for who to put numbers in a Venn diagram showing probability values: *always put a number in each region which corresponds to the probability of that smallest connected region containing the number, not any larger region*.

One last point we should make, using the same argument as in the above example. Suppose we have events $A$ and $B$ in a sample space $S$ (again). Suppose we are not sure if $A$ and $B$ are disjoint, so we cannot use the addition rule for disjoint events to compute $P(A \cup B)$. But notice that the events $A$ and $A^c$ are disjoint, so that $A \cap B$ and $A^c \cap B$ are also disjoint and

$$A = A \cap S = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c)$$

is a decomposition of the event $A$ into the two disjoint events $A \cap B$ and $A^c \cap B$. From the addition rule for disjoint events, this means that

$$P(A) = P(A \cap B) + P(A \cap B^c) \, .$$

Similar reasoning tells us both that

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

and that

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$$

is a decomposition of $A \cup B$ into disjoint pieces, so that

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \, .$$

Combining all of these equations, we conclude that

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

This is important enough to state as a

FACT 4.1.16. **The Addition Rule for General Events** If $A$ and $B$ are events in a sample space $S$ then we have the addition rule for their probabilities

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)\,.$$

This rule is true whether or not $A$ and $B$ are disjoint.

**4.1.3. Finite Probability Models.** Here is a nice situation in which we can easily calculate a lot of probabilities fairly easily: if the sample space $S$ of some experiment is *finite*.

So let's suppose the sample space consists of just the outcomes $S=\{o_1,o_2,\ldots,o_n\}$. For each of the outcomes, we can compute the probability:

$$\begin{aligned} p_1 &= P(\{o_1\}) \\ p_2 &= P(\{o_2\}) \\ &\vdots \\ p_n &= P(\{o_n\}) \end{aligned}$$

Let's think about what the rules for probability models tell us about these numbers $p_1,p_2,\ldots,p_n$. First of all, since they are each the probability of an event, we see that

$$\begin{aligned} 0 &\le p_1 \le 1 \\ 0 &\le p_2 \le 1 \\ &\vdots \\ 0 &\le p_n \le 1 \end{aligned}$$

Furthermore, since $S=\{o_1,o_2,\ldots,o_n\}=\{o_1\}\cup\{o_2\}\cup\cdots\cup\{o_n\}$ and all of the events $\{o_1\},\{o_2\},\ldots,\{o_n\}$ are disjoint, by the addition rule for disjoint events we have

$$\begin{aligned} 1=P(S) &= P(\{o_1,o_2,\ldots,o_n\}) \\ &= P(\{o_1\}\cup\{o_2\}\cup\cdots\cup\{o_n\}) \\ &= P(\{o_1\})+P(\{o_2\})+\cdots+P(\{o_n\}) \\ &= p_1+p_2+\cdots+p_n\,. \end{aligned}$$

The final thing to notice about this situation of a finite sample space is that if $E \subset S$ is any event, then $E$ will be just a collection of some of the outcomes from $\{o_1, o_2, \ldots, o_n\}$ (maybe none, maybe all, maybe an intermediate number). Since, again, the events like $\{o_1\}$ and $\{o_2\}$ and so on are disjoint, we can compute

$$\begin{aligned} P(E) &= P(\{\text{the outcomes } o_j \text{ which make up } E\}) \\ &= \sum \{\text{the } p_j\text{'s for the outcomes in } E\} \ . \end{aligned}$$

In other words

FACT 4.1.17. A probability model on a sample space $S$ with a finite number, $n$, of outcomes, is nothing other than a choice of real numbers $p_1, p_2, \ldots, p_n$, all in the range from 0 to 1 and satisfying $p_1 + p_2 + \cdots + p_n = 1$. For such a choice of numbers, we can compute the probability of any event $E \subset S$ as

$$P(E) = \sum \{\text{the } p_j\text{'s corresponding to the outcomes } o_j \text{ which make up } E\} \ .$$

EXAMPLE 4.1.18. For the coin flip of Example 4.1.2, there are only the two outcomes $H$ and $T$ for which we need to pick two probabilities, call them $p$ and $q$. In fact, since the total must be 1, we know that $p + q = 1$ or, in other words, $q = 1 - p$. The the probabilities for all events (which we listed in Example 4.1.9) are

$$\begin{aligned} P(\emptyset) &= 0 \\ P(\{H\}) &= p \\ P(\{T\}) &= q = 1 - p \\ P(\{H, T\}) &= p + q = 1 \end{aligned}$$

What we've described here is, potentially, a **biased coin**, since we are not assuming that $p = q$ – the probabilities of getting a head and a tail are not assumed to be the same. The alternative is to assume that we have a **fair coin**, meaning that $p = q$. Note that in such a case, since $p + q = 1$, we have $2p = 1$ and so $p = 1/2$. That is, the probability of a head (and, likewise, the probability of a tail) in a single throw of a fair coin is $1/2$.

EXAMPLE 4.1.19. As in the previous example, we can consider the die of Example 4.1.3 to a **fair die**, meaning that the individual face probabilities are all the same. Since they must also total to 1 (as we saw for all finite probability models), it follows that

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6.$$

We can then use this basic information and the formula (for $P(E)$) in Fact 4.1.17 to compute the probability of any event of interest, such as

$$P(\text{“roll was even”}) = P(\{2, 4, 6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \ .$$

We should immortalize these last two examples with a

DEFINITION 4.1.20. When we are talking about dice, coins, individuals for some task, or another small, practical, finite experiment, we use the term **fair** to indicate that the probabilities of all individual outcomes are equal (and therefore all equal to the the number $1/n$, where $n$ is the number of outcomes in the sample space). A more technical term for the same idea is **equiprobable**, while a more casual term which is often used for this in very informal settings is "**at random**" (such as "pick a card *at random* from this deck" or "pick a random patient from the study group to give the new treatment to...").

EXAMPLE 4.1.21. Suppose we look at the experiment of Example 4.1.4 and add the information that the two dice we are rolling are *fair*. This actually isn't quite enough to figure out the probabilities, since we also have to assure that the fair rolling of the first die doesn't in any way affect the rolling of the second die. This is technically the requirement that the two rolls be *independent*, but since we won't investigate that carefully until §4.2, below, let us instead here simply say that we assume the two rolls are fair and are in fact completely uninfluenced by anything around them in the world including each other.

What this means is that, in the long run, we would expect the first die to show a $1$ roughly $\frac{1}{6}^{th}$ of the time, and in the very long run, the second die would show a $1$ roughly $\frac{1}{6}^{th}$ of *those* times. This means that the outcome of the "roll two dice" experiment should be $11$ with probability $\frac{1}{36}$ – and the same reasoning would show that all of the outcomes have that probability. In other words, this is an equiprobable sample space with $36$ outcomes each having probability $\frac{1}{36}$. Which in turn enables us to compute any probability we might like, such as

$$\begin{aligned} P(\text{"sum of the two rolls is } 4\text{"}) &= P(\{13, 22, 31\}) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &= \frac{3}{36} \\ &= \frac{1}{12} \, . \end{aligned}$$

## 4.2. Conditional Probability

We have described the whole foundation of the theory of probability as coming from *imperfect knowledge*, in the sense that we don't know for sure if an event $A$ will happen any particular time we do the experiment but we do know, in the long run, in what fraction of times $A$ will happen. Or, at least, we claim that there is some number $P(A)$ such that after running the experiment $N$ times, out of which $n_A$ of these times are when $A$ happened, $P(A)$ is approximately $n_A/N$ (and this ratio gets closer and closer to $P(A)$ as $N$ gets bigger and bigger).

But what if we have *some* knowledge? In particular, what happens if we know for sure that the event $B$ has happened – will that influence our knowledge of whether $A$ happens or not? As before, when there is randomness involved, we cannot tell for sure if $A$ will happen, but we hope that, given the knowledge that $B$ happened, we can make a more accurate guess about the probability of $A$.

EXAMPLE 4.2.1. If you pick a person at random in a certain country on a particular date, you might be able to estimate the probability that the person had a certain height if you knew enough about the range of heights of the whole population of that country. [In fact, below we will make estimates of this kind.] That is, if we define the event

$$A = \text{“the random person is taller than } 1.829 \text{ meters } (6 \text{ feet})\text{”}$$

then we might estimate $P(A)$.

But consider the event

$$B = \text{“the random person's parents were both taller than } 1.829 \text{ meters”} .$$

Because there is a genetic component to height, if you know that $B$ happened, it would change your idea of how likely, given that knowledge, that $A$ happened. Because genetics are not the only thing which determines a person's height, you would not be certain that $A$ happened, even given the knowledge of $B$.

Let us use the frequentist approach to derive a formula for this kind of *probability of $A$ given that $B$ is known to have happened*. So think about doing the repeatable experiment many times, say $N$ times. Out of all those times, some times $B$ happens, say it happens $n_B$ times. Out of *those* times, the ones where $B$ happened, sometimes $A$ also happened. These are the cases where both $A$ and $B$ happened – or, converting this to a more mathematical descriptions, the times that $A \cap B$ happened – so we will write it $n_{A\cap B}$.

We know that the probability of $A$ happening in the cases where we know for sure that $B$ happened is approximately $n_{A\cap B}/n_B$. Let's do that favorite trick of multiplying and dividing by the same number, so finding that the probability in which we are interested is

approximately

$$\frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}\cdot N}{N\cdot n_B} = \frac{n_{A\cap B}}{N}\cdot\frac{N}{n_B} = \frac{n_{A\cap B}}{N}\Big/\frac{n_B}{N} \approx P(A\cap B)\Big/P(B)$$

Which is why we make the

DEFINITION 4.2.2. The **conditional probability of the event $A$ given the event $B$** is

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

Here $P(A|B)$ is pronounced *the probability of $A$ given $B$*.

Let's do a simple

EXAMPLE 4.2.3. Building off of Example 4.1.19, note that the probability of rolling a 2 is $P(\{2\}) = 1/6$ (as is the probability of rolling any other face – it's a *fair die*). But suppose that you were told that the roll was even, which is the event $\{2,4,6\}$, and asked for the probability that the roll was a 2 given this prior knowledge. The answer would be

$$P(\{2\}\mid\{2,4,6\}) = \frac{P(\{2\}\cap\{2,4,6\})}{P(\{2,4,6\})} = \frac{P(\{2\})}{P(\{2,4,6\})} = \frac{1/6}{1/2} = 1/3\,.$$

In other words, the probability of rolling a 2 on a fair die with no other information is $1/6$, which the probability of rolling a 2 given that we rolled an even number is $1/3$. So the probability doubled with the given information.

Sometimes the probability changes even more than merely doubling: the probability that we rolled a 1 with no other knowledge is $1/6$, while the probability that we rolled a 1 given that we rolled an even number is

$$P(\{1\}\mid\{2,4,6\}) = \frac{P(\{1\}\cap\{2,4,6\})}{P(\{2,4,6\})} = \frac{P(\emptyset)}{P(\{2,4,6\})} = \frac{0}{1/2} = 0\,.$$

But, actually, sometimes the conditional probability for some event is the same as the unconditioned probability. In other words, sometimes knowing that $B$ happened doesn't change our estimate of the probability of $A$ at all, they are no really related events, at least from the point of view of probability. This motivates the

DEFINITION 4.2.4. Two events $A$ and $B$ are called **independent** if $P(A\mid B) = P(A)$.

Plugging the defining formula for $P(A\mid B)$ into the definition of *independent*, it is easy to see that

FACT 4.2.5. Events $A$ and $B$ are independent if and only if $P(A\cap B) = P(A)\cdot P(B)$.

EXAMPLE 4.2.6. Still using the situation of Example 4.1.19, we saw in Example 4.2.3 that the events $\{2\}$ and $\{2,3,4\}$ are not independent since

$$P(\{2\}) = 1/6 \neq 1/3 = P(\{2\}\mid\{2,4,6\})$$

nor are $\{1\}$ and $\{2,3,4\}$, since

$$P(\{1\}) = 1/6 \neq 0 = P(\{1\} \mid \{2,4,6\}) .$$

However, look at the events $\{1,2\}$ and $\{2,4,6\}$:

$$\begin{aligned} P(\{1,2\}) = P(\{1\}) + P(\{2\}) &= 1/6 + 1/6 \\ &= 1/3 \\ &= \frac{1/6}{1/2} \\ &= \frac{P(\{1\})}{P(\{2,4,6\})} \\ &= \frac{P(\{1,2\} \cap \{2,4,6\})}{P(\{2,4,6\})} \\ &= P(\{1,2\} \mid \{2,4,6\}) \end{aligned}$$

which means that they are independent!

EXAMPLE 4.2.7. We can now fully explain what was going on in Example 4.1.21. The two fair dice were supposed to be rolled in a way that the first roll had no effect on the second – this exactly means that the dice were rolled *independently*. As we saw, this then means that each individual outcome of sample space $S$ had probability $\frac{1}{36}$. But the first roll having any particular value is independent of the second roll having another, *e.g.*, if $A = \{11, 12, 13, 14, 15, 16\}$ is the event in that sample space of getting a $1$ on the first roll and $B = \{14, 24, 34, 44, 54, 64\}$ is the event of getting a $4$ on the second roll, then events $A$ and $B$ are independent, as we check by using Fact 4.2.5:

$$\begin{aligned} P(A \cap B) &= P(\{14\}) \\ &= \frac{1}{36} \\ &= \frac{1}{6} \cdot \frac{1}{6} \\ &= \frac{6}{36} \cdot \frac{6}{36} \\ &= P(A) \cdot P(B) . \end{aligned}$$

On the other hand, the event "the sum of the rolls is $4$," which is $C = \{13, 22, 31\}$ as a set, *is not independent* of the value of the first roll, since $P(A \cap C) = P(\{13\}) = \frac{1}{36}$ but $P(A) \cdot P(C) = \frac{6}{36} \cdot \frac{3}{36} = \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{72}$.

## 4.3. Random Variables

**4.3.1. Definition and First Examples.** Suppose we are doing a random experiment and there is some consequence of the result in which we are interested that can be measured by a number. The experiment might be playing a game of chance and the result could be how much you win or lose depending upon the outcome, or the experiment could be which part of the drives' manual you randomly choose to study and the result how many points we get on the driver's license test we make the next day, or the experiment might be giving a new drug to a random patient in medical study and the result would be some medical measurement you make after treatment (blood pressure, white blood cell count, whatever), *etc.* There is a name for this situation in mathematics

DEFINITION 4.3.1. A choice of a number for each outcome of a random experiment is called a **random variable** [**RV**]. If the values an RV takes can be counted, because they are either finite or countably infinite[1] in number, the RV is called **discrete**; if, instead, the RV takes on all the values in an interval of real numbers, the RV is called **continuous**.

We usually use capital letters to denote RVs and the corresponding lowercase letter to indicate a particular numerical value the RV might have, like $X$ and $x$.

EXAMPLE 4.3.2. Suppose we play a silly game where you pay me \$5 to play, then I flip a fair coin and I give you \$10 if the coin comes up heads and \$0 if it comes up tails. Then your net winnings, which would be +\$5 or -\$5 each time you play, are a random variable. Having only two possible values, this RV is certainly discrete.

EXAMPLE 4.3.3. Weather phenomena vary so much, due to such small effects – such as the famous butterfly flapping its wings in the Amazon rain forest causing a hurricane in North America – that they appear to be a random phenomenon. Therefore, observing the temperature at some weather station is a continuous random variable whose value can be any real number in some range like $-100$ to $100$ (we're doing *science*, so we use $^\circ C$).

EXAMPLE 4.3.4. Suppose we look at the "*roll two fair dice independently*" experiment from Example 4.2.7 and Example 4.1.21, which was based on the probability model in Example 4.1.21 and sample space in Example 4.1.4. Let us consider in this situation the random variable $X$ whose value for some pair of dice rolls is the sum of the two numbers showing on the dice. So, for example, $X(11) = 2$, $X(12) = 3$, *etc.*

[1]There many kinds of infinity in mathematics – in fact, an infinite number of them. The smallest is an infinity that can be counted, like the whole numbers. But then there are many larger infinities, describing sets that are too big even to be counted, like the set of all real numbers.

In fact, let's make a table of all the values of $X$:

$$\begin{aligned}
X(11) &= 2 \\
X(21) = X(12) &= 3 \\
X(31) = X(22) = X(13) &= 4 \\
X(41) = X(32) = X(23) = X(14) &= 5 \\
X(51) = X(42) = X(33) = X(24) = X(15) &= 6 \\
X(61) = X(52) = X(43) = X(34) = X(25) = X(16) &= 7 \\
X(62) = X(53) = X(44) = X(35) = X(26) &= 8 \\
X(63) = X(54) = X(45) = X(36) &= 9 \\
X(64) = X(55) = X(46) &= 10 \\
X(65) = X(56) &= 11 \\
X(66) &= 12
\end{aligned}$$

**4.3.2. Distributions for Discrete RVs.** The first thing we do with a random variable, usually, is talk about the probabilities associate with it.

DEFINITION 4.3.5. Given a discrete RV $X$, its **distribution** is a list of all of the values $X$ takes on, together with the probability of it taking that value.

[Note this is quite similar to Definition 1.3.5 – because it is essentially the same thing.]

EXAMPLE 4.3.6. Let's look at the RV, which we will call $X$, in the silly betting game of Example 4.3.2. As we noticed when we first defined that game, there are two possible values for this RV, \$5 and -\$5. We can actually think of "$X = 5$" as describing an event, consisting of the set of all outcomes of the coin-flipping experiment which give you a net gain of \$5. Likewise, "$X = -5$" describes the event consisting of the set of all outcomes which give you a net gain of -\$5. These events are as follows:

| $x$ | Set of outcomes $o$ such that $X(o) = x$ |
|---|---|
| 5 | $\{H\}$ |
| $-5$ | $\{T\}$ |

Since it is a fair coin so the probabilities of these events are known (and very simple), we conclude that the distribution of this RV is the table

| $x$ | $P(X = x)$ |
|---|---|
| 5 | $1/2$ |
| $-5$ | $1/2$ |

EXAMPLE 4.3.7. What about the $X =$ "*sum of the face values*" RV on the "*roll two fair dice, independently*" random experiment from Example 4.3.4? We have actually already done most of the work, finding out what values the RV can take and which outcomes cause each of those values. To summarize what we found:

| $x$ | Set of outcomes $o$ such that $X(o) = x$ |
|---|---|
| 2 | $\{11\}$ |
| 3 | $\{21, 12\}$ |
| 4 | $\{31, 22, 13\}$ |
| 5 | $\{41, 32, 23, 14\}$ |
| 6 | $\{51, 42, 33, 24, 15\}$ |
| 7 | $\{61, 52, 43, 34, 25, 16\}$ |
| 8 | $\{62, 53, 44, 35, 26\}$ |
| 9 | $\{63, 54, 45, 36\}$ |
| 10 | $\{64, 55, 46\}$ |
| 11 | $\{65, 56\}$ |
| 12 | $\{66\}$ |

But we have seen that this is an equiprobable situation, where the probability of any event $A$ contain $n$ outcomes is $P(A) = n \cdot 1/36$, so we can instantly fill in the distribution table for this RV as

| $x$ | $P(X = x)$ |
|---|---|
| 2 | $\frac{1}{36}$ |
| 3 | $\frac{2}{36} = \frac{1}{18}$ |
| 4 | $\frac{3}{36} = \frac{1}{12}$ |
| 5 | $\frac{4}{36} = \frac{1}{6}$ |
| 6 | $\frac{5}{36}$ |
| 7 | $\frac{6}{36} = \frac{1}{6}$ |
| 8 | $\frac{5}{36}$ |
| 9 | $\frac{4}{36} = \frac{1}{6}$ |
| 10 | $\frac{3}{36} = \frac{1}{12}$ |
| 11 | $\frac{2}{36} = \frac{1}{18}$ |
| 12 | $\frac{1}{36}$ |

One thing to notice about distributions is that if we make a preliminary table, as we just did, of the events consisting of all outcomes which give a particular value when plugged into the RV, then we will have a collection of disjoint events which exhausts all of the sample space. What this means is that the sum of the probability values in the distribution table of an RV is the probability of the whole sample space of that RV's experiment. Therefore

FACT 4.3.8. The sum of the probabilities in a distribution table for a random variable must always equal 1.

It is quite a good idea, whenever you write down a distribution, to check that this Fact is true in your distribution table, simply as a sanity check against simple arithmetic errors.

**4.3.3. Expectation for Discrete RVs.** Since we cannot predict what exactly will be the outcome each time we perform a random experiment, we cannot predict with precision what will be the value of an RV on that experiment, each time. But, as we did with the basic idea of probability, maybe we can at least learn something from the long-term trends. It turns out that it is relatively easy to figure out the mean value of an RV over a large number of runs of the experiment.

Say $X$ is a discrete RV, for which the distribution tells us that $X$ takes the values $x_1, \ldots, x_n$, each with corresponding probability $p_1, \ldots, p_n$. Then the frequentist view of probability says that the probability $p_i$ that $X = x_i$ is (approximately) $n_i/N$, where $n_i$ is the number of times $X = x_i$ out of a large number $N$ of runs of the experiment. But if

$$p_i = n_i/N$$

then, multiplying both sides by $N$,

$$n_i = p_i\, N \ .$$

That means that, out of the $N$ runs of the experiment, $X$ will have the value $x_1$ in $p_1\, N$ runs, the value $x_2$ in $p_2\, N$ runs, *etc.* So the sum of $X$ over those $N$ runs will be

$$(p_1\, N)x_1 + (p_2\, N)x_2 + \cdots + (p_n\, N)x_n \ .$$

Therefore the mean value of $X$ over these $N$ runs will be the total divided by $N$, which is $p_1\, x_1 + \cdots + p_n x_n$. This motivates the definition

DEFINITION 4.3.9. Given a discrete RV $X$ which takes on the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, the **expectation** [sometimes also called the **expected value**] of $X$ is the value

$$E(X) = \sum p_i\, x_i \ .$$

By what we saw just before this definition, we have the following

FACT 4.3.10. The expectation of a discrete RV is the mean of its values over many runs of the experiment.

*Note:* The attentive reader will have noticed that we dealt above only with the case of a finite RV, not the case of a countably infinite one. It turns out that all of the above works quite well in that more complex case as well, so long as one is comfortable with a bit of mathematical technology called "*summing an infinite series*." We do not assume such a

comfort level in our readers at this time, so we shall pass over the details of expectations of infinite, discrete RVs.

EXAMPLE 4.3.11. Let's compute the expectation of net profit RV $X$ in the silly betting game of Example 4.3.2, whose distribution we computed in Example 4.3.6. Plugging straight into the definition, we see

$$E(X) = \sum p_i\, x_i = \frac{1}{2} \cdot 5 + \frac{1}{2} \cdot (-5) = 2.5 - 2.5 = 0\ .$$

In other words, your average net gain playing this silly game many times will be **zero**. Note that does not mean anything like "*if you lose enough times in a row, the chances of starting to win again will go up*," as many gamblers seem to believe, it just means that, in the very long run, we can expect the average winnings to be approximately zero – but no one knows how long that run has to be before the balancing of wins and losses happens[2].

A more interesting example is

EXAMPLE 4.3.12. In Example 4.3.7 we computed the distribution of the random variable $X$ = "*sum of the face values*" on the "*roll two fair dice, independently*" random experiment from Example 4.3.4. It is therefore easy to plug the values of the probabilities and RV values from the distribution table into the formula for expectation, to get

$$\begin{aligned} E(X) &= \sum p_i\, x_i \\ &= \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \frac{4}{36} \cdot 5 + \frac{5}{36} \cdot 6 + \frac{6}{36} \cdot 7 + \frac{5}{36} \cdot 8 + \frac{4}{36} \cdot 9 + \frac{3}{36} \cdot 10 \\ &\qquad + \frac{2}{36} \cdot 11 + \frac{1}{36} \cdot 12 \\ &= \frac{2 \cdot 1 + 3 \cdot 2 + 4 \cdot 3 + 5 \cdot 4 + 6 \cdot 5 + 7 \cdot 6 + 8 \cdot 5 + 9 \cdot 4 + 10 \cdot 3 + 11 \cdot 2 + 12 \cdot 1}{36} \\ &= 7 \end{aligned}$$

So if you roll two fair dice independently and add the numbers which come up, then do this process many times and take the average, in the long run that average will be the value 7.

**4.3.4. Density Functions for Continuous RVs.** What about continuous random variables? Definition 4.3.5 of *distribution* explicitly excluded the case of continuous RVs, so does that mean we cannot do probability calculations in that case?

There is, when we think about it, something of a problem here. A distribution is supposed to be a list of possible values of the RV and the probability of each such value. But if some continuous RV has values which are an interval of real numbers, there is just no way to list all such numbers – it has been known since the late 1800s that there is no way to make a list like that (see [**Wik17a**], for a description of a very pretty proof of this fact). In

[2] In fact, in a very precise sense which we will not discuss in this book, the longer you play a game like this, the more you can expect there will be short-term, but very large, wins and losses.

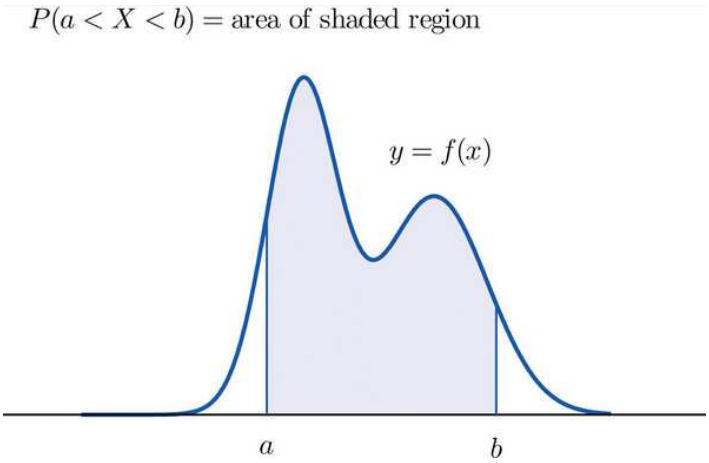
addition, the chance of some random process producing a real number that is *exactly* equal to some particular value really is zero: for two real numbers to be precisely equal requires infinite accuracy ... think of all of those decimal digits, marching off in orderly rows to infinity, which must match between the two numbers.

Rather than a distribution, we do the following:

DEFINITION 4.3.13. Let $X$ be a continuous random variable whose values are the real interval $[x_{min}, x_{max}]$, where either $x_{min}$ or $x_{max}$ or both may be $\infty$. A [**probability**] **density function** for $X$ is a function $f(x)$ defined for $x$ in $[x_{min}, x_{max}]$, meaning it is a curve with one $y$ value for each $x$ in that interval, with the property that

$$P(a < X < b) = \begin{cases} \text{the area in the } xy\text{-plane above the } x\text{-axis, below} \\ \text{the curve } y = f(x) \text{ and between } x = a \text{ and } x = b. \end{cases}.$$

Graphically, what is going on here is




Because of what we know about probabilities, the following is true (and fairly easy to prove):

FACT 4.3.14. Suppose $f(x)$ is a density function for the continuous RV $X$ defined on the real interval $[x_{min}, x_{max}]$. Then

- For all $x$ in $[x_{min}, x_{max}]$, $f(x) \geq 0$.
- The total area under the curve $y = f(x)$, above the $x$-axis, and between $x = x_{min}$ and $x = x_{max}$ is 1.

If we want the idea of *picking a real number on the interval* $[x_{min}, x_{max}]$ *at random*, where *at random* means that all numbers have the same chance of being picked (along the lines of *fair* in Definition 4.1.20, the height of the density function must be the same at all $x$. In other words, the density function $f(x)$ must be a constant $c$. In fact, because of the above Fact 4.3.14, that constant must have the value $\frac{1}{x_{max} - x_{min}}$. There is a name for this:

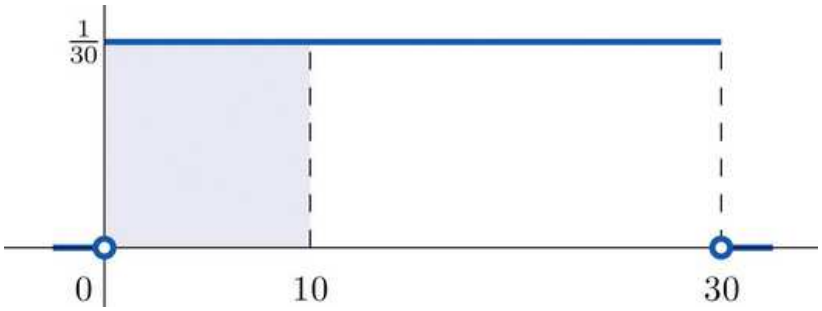
DEFINITION 4.3.15. The **uniform distribution on** $[x_{min}, x_{max}]$ is the distribution for the continuous RV whose values are the interval $[x_{min}, x_{max}]$ and whose density function is the constant function $f(x) = \frac{1}{x_{max} - x_{min}}$.

EXAMPLE 4.3.16. Suppose you take a bus to school every day and because of a chaotic home life (and, let's face it, you don't like mornings), you get to the bus stop at a pretty nearly perfectly random time. The bus also doesn't stick perfectly to its schedule – but it is guaranteed to come at least every $30$ minutes. What this adds up to is the idea that your waiting time at the bus stop is a uniformly distributed RV on the interval $[0, 30]$.

If you wonder one morning how likely it then is that you will wait for less than $10$ minutes, you can simply compute the area of the rectangle whose base is the interval $[0, 10]$ on the $x$-axis and whose height is $\frac{1}{30}$, which will be

$$P(0 < X < 10) = \textit{base} \cdot \textit{height} = 10 \cdot \frac{1}{30} = \frac{1}{3} .$$

A picture which should clarify this is




where the area of the shaded region represents the probability of having a waiting time from $0$ to $10$ minutes.

One technical thing that can be confusing about continuous RVs and their density functions is the question of whether we should write $P(a < X < b)$ or $P(a \leq X \leq b)$. But if you think about it, we really have three possible events here:

$$A = \{\textit{outcomes such that } X = a\},$$
$$M = \{\textit{outcomes such that } a < X < b\}, \text{ and}$$
$$B = \{\textit{outcomes such that } X = b\} .$$

Since $X$ always takes on exactly one value for any particular outcome, there is no overlap between these events: they are all disjoint. That means that

$$P(A \cup M \cup B) = P(A) + P(M) + P(B) = P(M)$$

where the last equality is because, as we said above, the probability of a continuous RV taking on exactly one particular value, as it would in events $A$ and $B$, is $0$. The same would be true if we added merely one endpoint of the interval $(a, b)$. To summarize:

FACT 4.3.17. If $X$ is a continuous RV with values forming the interval $[x_{min}, x_{max}]$ and $a$ and $b$ are in this interval, then

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) \ .$$

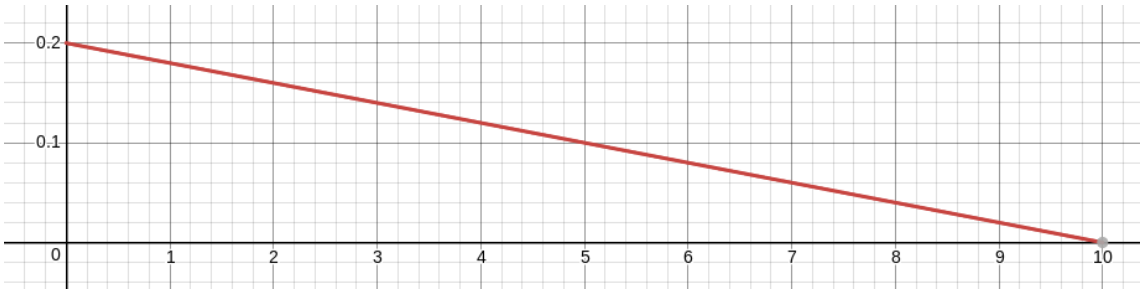
As a consequence of this fact, some authors write probability formulæ about continuous RVs with " $<$ " and some with " $\leq$ " and *it makes no difference*.

Let's do a slightly more interesting example than the uniform distribution:

EXAMPLE 4.3.18. Suppose you repeatedly throw darts at a dartboard. You're not a machine, so the darts hit in different places every time and you think of this as a repeatable random experiment whose outcomes are the locations of the dart on the board. You're interested in the probabilities of getting close to the center of the board, so you decide for each experimental outcome (location of a dart you threw) to measure its distance to the center – this will be your RV $X$.

Being good at this game, you hit near the center more than near the edge and you never completely miss the board, whose radius is $10cm$– so $X$ is more likely to be near $0$ than near $10$, and it is never greater than $10$. What this means is that the RV has values forming the interval $[0, 10]$ and the density function, defined on the same interval, should have its maximum value at $x = 0$ and should go down to the value $0$ when $x = 10$.

You decide to model this situation with the simplest density function you can think of that has the properties we just noticed: a straight line from the highest point of the density function when $x = 0$ down to the point $(10, 0)$. The figure that will result will be a triangle, and since the total area must be $1$ and the base is $10$ units long, the height must be $.2$ units. [To get that, we solved the equation $1 = \frac{1}{2}bh = \frac{1}{2}10h = 5h$ for $h$.] So the graph must be




and the equation of this linear density function would be $y = -\frac{1}{50}x + .2$ [why? – think about the slope and $y$-intercept!].

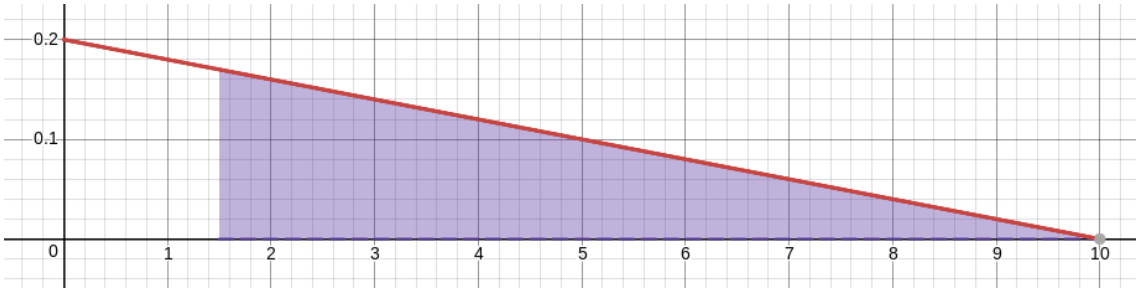
To the extent that you trust this model, you can now calculate the probabilities of events like, for example, "*hitting the board within that center bull's-eye of radius* $1.5cm$," which probability would be the area of the shaded region in this graph:


The upper-right corner of this shaded region is at $x$-coordinate $1.5$ and is on the line, so its $y$-coordinate is $-\frac{1}{50}1.5 + .2 = .17$ . Since the region is a trapezoid, its area is the distance between the two parallel sides times the average of the lengths of the other two sides, giving

$$P(0 < X < 1.5) = 1.5 \cdot \frac{.2 + .17}{2} = .2775\ .$$

In other words, the probability of hitting the bull's-eye, assuming this model of your dart-throwing prowess, is about $28\%$.

If you don't remember the formula for the area of a trapezoid, you can do this problem another way: compute the probability of the complementary event, and then take one minus that number. The reason to do this would be that the complementary event corresponds to the shaded region here




which is a triangle! Since we surely do remember the formula for the area of a triangle, we find that

$$P(1.5 < X < 10) = \frac{1}{2}bh = \frac{1}{2}.17 \cdot 8.5 = .7225$$

and therefore $P(0 < X < 1.5) = 1 - P(1.5 < X < 10) = 1 - .7225 = .2775$. [It's nice that we got the same number this way, too!]

**4.3.5. The Normal Distribution.** We've seen some examples of continuous RVs, but we have yet to meet the most important one of all.

DEFINITION 4.3.19. The **Normal distribution with mean $\mu_X$ and standard deviation** $\sigma_X$ is the continuous RV which takes on all real values and is governed by the probability density function

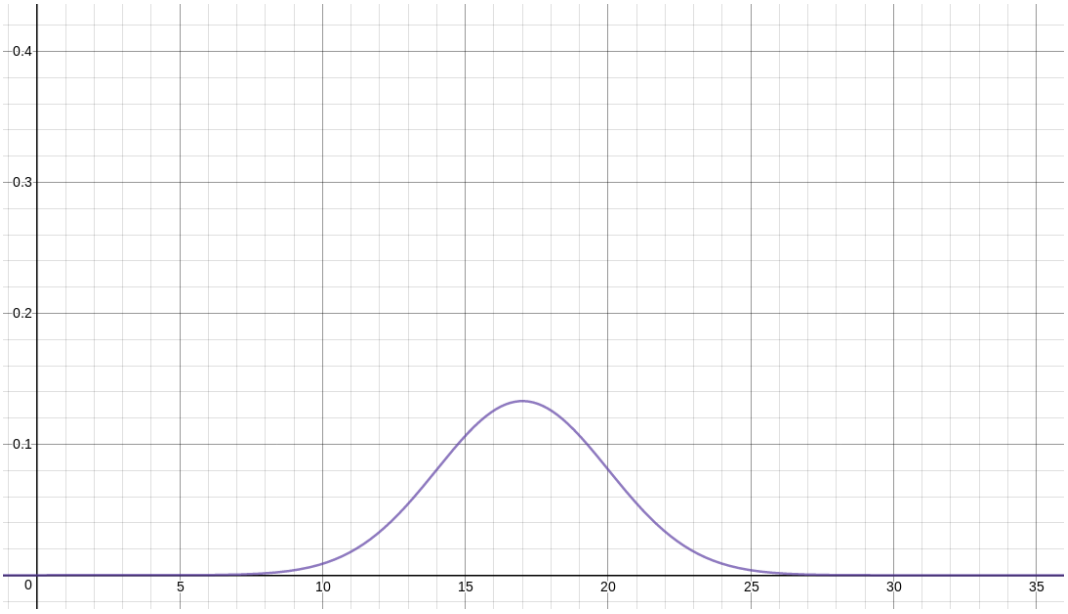
$$\rho(x) = \frac{1}{\sqrt{2{\sigma_X}^2\pi}} e^{-\frac{(x-\mu_X)^2}{2{\sigma_X}^2}}\ .$$

If $X$ is a random variable which follows this distribution, then we say that $X$ **is Normally distributed with mean** $\mu_X$ **and standard deviation** $\sigma_X$ or, in symbols, $X$ **is** $N(\mu_X, \sigma_X)$.

[More technical works also call this the *Gaussian distribution*, named after the great mathematician *Carl Friedrich Gauss*. But we will not use that term again in this book after this sentence ends.]

The good news about this complicated formula is that we don't really have to do anything with it. We will collect some properties of the Normal distribution which have been derived from this formula, but these properties are useful enough, and other tools such as modern calculators and computers which can find specific areas we need under the graph of $y = \rho(x)$, that we won't need to work directly with the above formula for $\rho(x)$ again. It is nice to know that $N(\mu_X, \sigma_X)$ does correspond to a specific, known density function, though, isn't it?

It helps to start with an image of what the Normal distribution looks like. Here is the density function for $\mu_X = 17$ and $\sigma_X = 3$:




Now let's collect some of these useful facts about the Normal distributions.

FACT 4.3.20. The density function $\rho$ for the Normal distribution $N(\mu_X, \sigma_X)$ is a positive function for all values of $x$ and the total area under the curve $y = \rho(x)$ is 1.

This simply means that $\rho$ is a good candidate for the probability density function for some continuous RV.

FACT 4.3.21. The density function $\rho$ for the Normal distribution $N(\mu_X, \sigma_X)$ is unimodal with maximum at $x$-coordinate $\mu_X$.

This means that $N(\mu_X, \sigma_X)$ is a possible model for an RV $X$ which tends to have one main, central value, and less often has other values farther away. That center is at the location given by the parameter $\mu_X$, so wherever we want to put the center of our model for $X$, we just use that for $\mu_X$.

FACT 4.3.22. The density function $\rho$ for the Normal distribution $N(\mu_X, \sigma_X)$ is is symmetric when reflected across the line $x = \mu_X$.

This means that the amount $X$ misses its center, $\mu_X$, tends to be about the same when it misses above $\mu_X$ and when it misses below $\mu_X$. This would correspond to situations were you hit as much to the right as to the left of the center of a dartboard. Or when randomly picked people are as likely to be taller than the average height as they are to be shorter. Or when the time it takes a student to finish a standardized test is as likely to be less than the average as it is to be more than the average. Or in many, many other useful situations.

FACT 4.3.23. The density function $\rho$ for the Normal distribution $N(\mu_X, \sigma_X)$ has has tails in both directions which are quite thin, in fact get extremely thin as $x \to \pm\infty$, but never go all the way to $0$.

This means that $N(\mu_X, \sigma_X)$ models situations where the amount $X$ deviates from its average has no particular cut-off in the positive or negative direction. So you are throwing darts at a dart board, for example, and there is no way to know how far your dart may hit to the right or left of the center, maybe even way off the board and down the hall – although that may be very unlikely. Or perhaps the time it takes to complete some task is usually a certain amount, but every once and a while it might take much more time, so much more that there is really no natural limit you might know ahead of time.

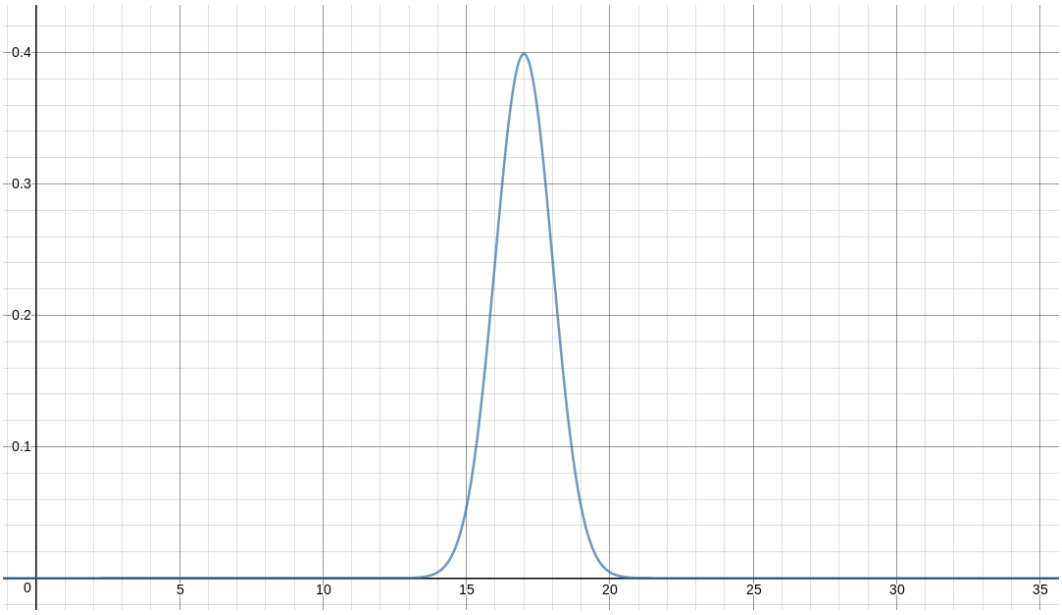
At the same time, those tails of the Normal distribution are so thin, for values far away from $\mu_X$, that it can be a good model even for a situation where there is a natural limit to the values of $X$ above or below $\mu_X$. For example, heights of adult males (in inches) in the United States are fairly well approximated by $N(69, 2.8)$, even though heights can never be less than $0$ and $N(69, 2.8)$ has an infinitely long tail to the left – because while that tail is non-zero all the way as $x \to -\infty$, it is very, very thin.

All of the above Facts are clearly true on the first graph we saw of a Normal distribution density function.

FACT 4.3.24. The graph of the density function $\rho$ for the Normal distribution $N(\mu_X, \sigma_X)$ has a taller and narrower peak if $\sigma_X$ is smaller, and a lower and wider peak if $\sigma_X$ is larger.

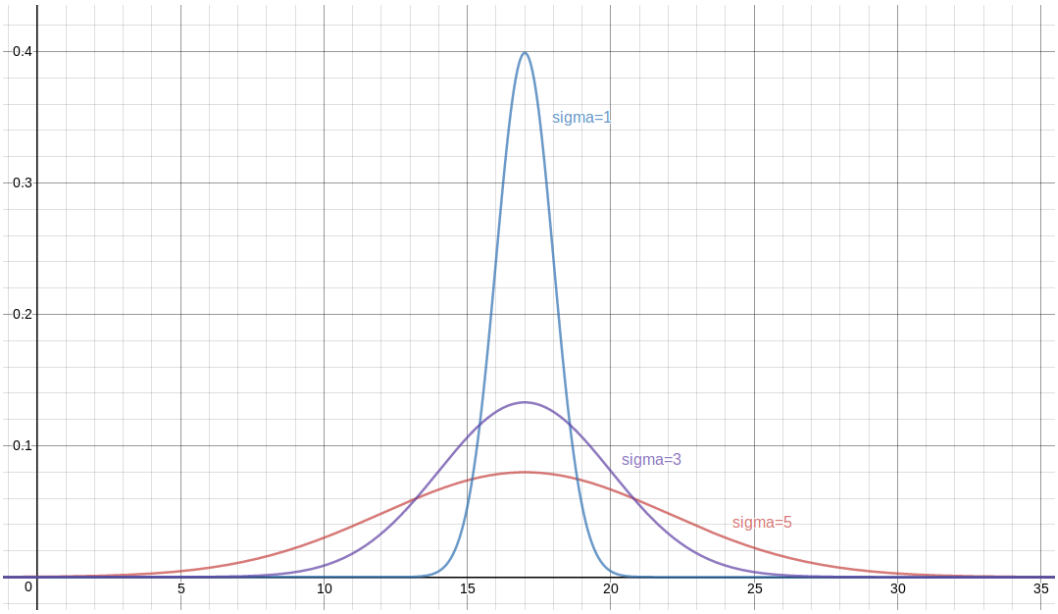
This allows the statistician to adjust how much variation there typically is in a normally distributed RV: By making $\sigma_X$ small, we are saying that an RV $X$ which is $N(\mu_X, \sigma_X)$ is very likely to have values quite close to its center, $\mu_X$. If we make $\sigma_X$ large, however, $X$ is more likely to have values all over the place – still, centered at $\mu_X$, but more likely to wander farther away.

Let's make a few versions of the graph we saw for $\rho$ when $\mu_X$ was 17 and $\sigma_X$ was 3, but now with different values of $\sigma_X$. First, if $\sigma_X = 1$, we get




If, instead, $\sigma_X = 5$, then we get

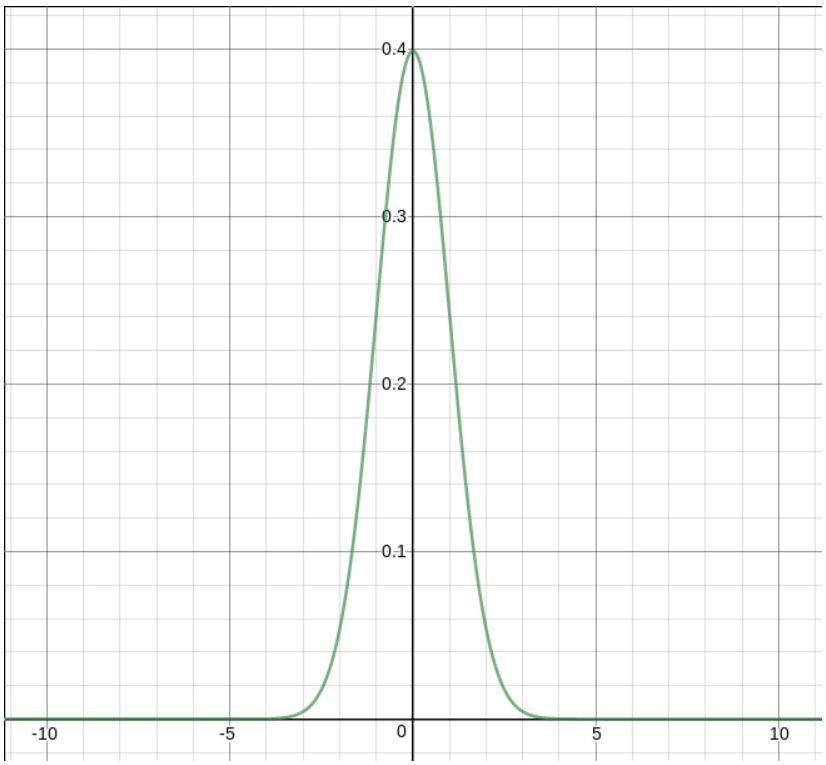

Finally, let's superimpose all of the above density functions on each other, for one, combined graph:




This variety of Normal distributions (one for each $\mu_X$ and $\sigma_X$) is a bit bewildering, so traditionally, we concentrate on one particularly nice one.

DEFINITION 4.3.25. The Normal distribution with mean $\mu_X = 0$ and standard deviation $\sigma_X = 1$ is called the **standard Normal distribution** and an RV [often written with the variable $Z$] that is $N(0, 1)$ is described as a **standard Normal RV**.

Here is what the standard Normal probability density function looks like:



One nice thing about the standard Normal is that all other Normal distributions can be related to the standard.

FACT 4.3.26. If $X$ is $N(\mu_X, \sigma_X)$, then $Z = (X - \mu_X)/\sigma_X$ is standard Normal.

This has a name.

DEFINITION 4.3.27. The process of replacing a random variable $X$ which is $N(\mu_X, \sigma_X)$ with the standard normal RV $Z = (X - \mu_X)/\sigma_X$ is called **standardizing a Normal RV**.

It used to be that standardization was an important step in solving problems with Normal RVs. A problem would be posed with information about some data that was modelled by a Normal RV with given mean $\mu_X$ and standardization $\sigma_X$. Then questions about probabilities for that data could be answered by standardizing the RV and looking up values in a single table of areas under the standard Normal curve.

Today, with electronic tools such as statistical calculators and computers, the standardization step is not really necessary.

EXAMPLE 4.3.28. As we noted above, the heights of adult men in the United States, when measured in inches, give a RV $X$ which is $N(69, 2.8)$. What percentage of the population, then, is taller than $6$ feet?

First of all, the frequentist point of view on probability tells us that what we are interested in is the probability that a randomly chosen adult American male will be taller than $6$ feet – that will be the same as the percentage of the population this tall. In other words, we must find the probability that $X > 72$, since in inches, $6$ feet becomes $72$. As $X$ is a continuous RV, we must find the area under its density curve, which is the $\rho$ for $N(69, 2.8)$, between $72$ and $\infty$.

That $\infty$ is a little intimidating, but since the tails of the Normal distribution are very thin, we can stop measuring area when $x$ is some large number and we will have missed only a very tiny amount of area, so we will have a very good approximation. Let's therefore find the area under $\rho$ from $x = 72$ up to $x = 1000$. This can be done in many ways:

- With a wide array of online tools – just search for "online normal probability calculator." One of these yields the value $.142$.
- With a **TI-8x** calculator, by typing

  **normalcdf(72, 1000, 69, 2.8)**

  which yields the value $.1419884174$. The general syntax here is

  **normalcdf(**$a$**,** $b$**,** $\mu_X$**,** $\sigma_X$**)**

  to find $P(a < X < b)$ when $X$ is $N(\mu_X, \sigma_X)$. Note you get **normalcdf** by typing

  $$\boxed{2ND} \rightarrow \boxed{VARS} \rightarrow 2$$

- Spreadsheets like **LibreOffice Calc** and **Microsoft Excel** will compute this by putting the following in a cell

  `=1-NORM.DIST(72, 69, 2.8, 1)`

  giving the value 0.1419883859. Here we are using the command

  `NORM.DIST(`$b$`, `$\mu_X$`, `$\sigma_X$`, 1)`

  which computes the area under the density function for $N(\mu_X, \sigma_X)$ from $-\infty$ to $b$. [The last input of "1" to `NORM.DIST` just tells it that we want to compute the area under the curve. If we used "0" instead, it would simple tell us the particular value of $\rho(b)$, which is of very direct little use in probability calculations.] Therefore, by doing $1 - NORM.DIST(72, 69, 2.8, 1)$, we are taking the total area of 1 and subtracting the area to the left of 72, yielding the area to the right, as we wanted.

  Therefore, if you want the area between $a$ and $b$ on an $N(\mu_X, \sigma_X)$ RV using a spreadsheet, you would put

  `=NORM.DIST(`$b$`, `$\mu_X$`, `$\sigma_X$`, 1) - NORM.DIST(`$a$`, `$\mu_X$`, `$\sigma_X$`, 1)`

  in a cell.

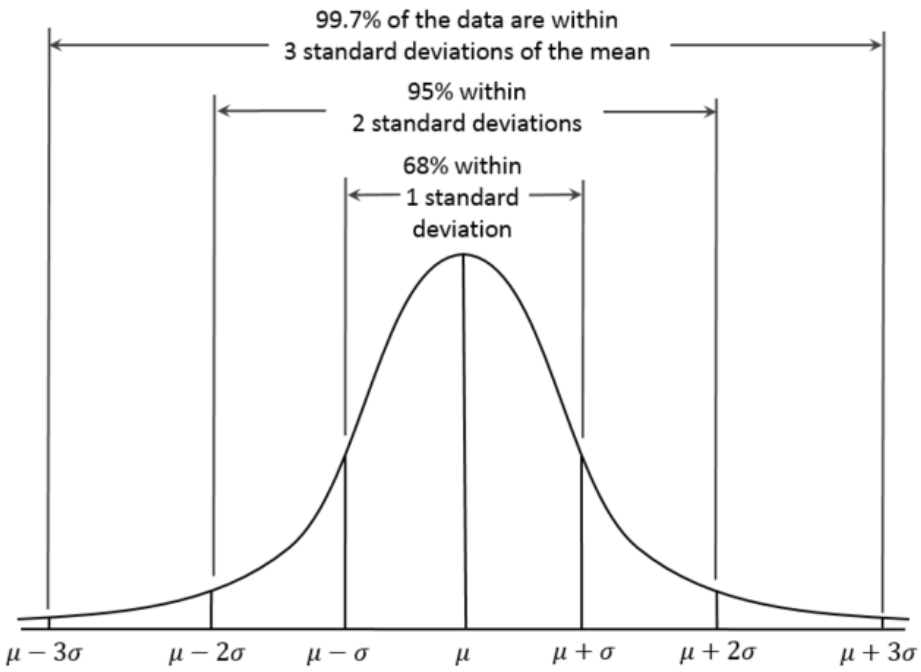
While standardizing a non-standard Normal RV and then looking up values in a table is an old-fashioned method that is tedious and no longer really needed, one old technique still comes in handy some times. It is based on the following:

FACT 4.3.29. **The 68-95-99.7 Rule**: Let $X$ be an $N(\mu_X, \sigma_X)$ RV. Then some special values of the area under the graph of the density curve $\rho$ for $X$ are nice to know:

- The area under the graph of $\rho$ from $x = \mu_X - \sigma_X$ to $x = \mu_X + \sigma_X$, also known as $P(\mu_X - \sigma_X < X < \mu_X + \sigma_X)$, is **.68** .
- The area under the graph of $\rho$ from $x = \mu_X - 2\sigma_X$ to $x = \mu_X + 2\sigma_X$, also known as $P(\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X)$, is **.95** .
- The area under the graph of $\rho$ from $x = \mu_X - 3\sigma_X$ to $x = \mu_X + 3\sigma_X$, also known as $P(\mu_X - 3\sigma_X < X < \mu_X + 3\sigma_X)$, is **.997** .

This is also called **The Empirical Rule** by some authors. Visually[3]:

[3]By Dan Kernler - Own work, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=36506025`.




In order to use the 68-95-99.7 Rule in understanding a particular situation, it is helpful to keep an eye out for the numbers that it talks about. Therefore, when looking at a problem, one should notice if the numbers $\mu_X + \sigma_X$, $\mu_X - \sigma_X$, $\mu_X + 2\sigma_X$, $\mu_X - 2\sigma_X$, $\mu_X + 3\sigma_X$, or $\mu_X - 3\sigma_X$ are ever mentioned. If so, perhaps this Rule can help.

EXAMPLE 4.3.30. In Example 4.3.28, we needed to compute $P(X > 72)$ where $X$ was known to be $N(69, 2.8)$. Is 72 one of the numbers for which we should be looking, to use the Rule? Well, it's greater than $\mu_X = 69$, so we could hope that it was $\mu_X + \sigma_X$, $\mu_X + 2\sigma_X$, or $\mu_X + 3\sigma_X$. But values are

$$\begin{aligned} \mu_X + \sigma_X &= 69 + 2.8 = 71.8, \\ \mu_X + 2\sigma_X &= 69 + 5.6 = 74.6, \text{and} \\ \mu_X + 3\sigma_X &= 69 + 8.4 = 77.4, \end{aligned}$$

none of which is what we need.

Well, it is true that $72 \approx 71.8$, so we could use that fact and accept that we are only getting an approximate answer – an odd choice, given the availability of tools which will give us extremely precise answers, but let's just go with it for a minute.

Let's see, the above Rule tells us that

$$P(66.2 < X < 71.8) = P(\mu_X - \sigma_X < X < \mu_X + \sigma_X) = .68\,.$$

Now since the total area under any density curve is 1,

$$P(X < 66.2 \text{ or } X > 71.8) = 1 - P(66.2 < X < 71.8) = 1 - .68 = .32\,.$$

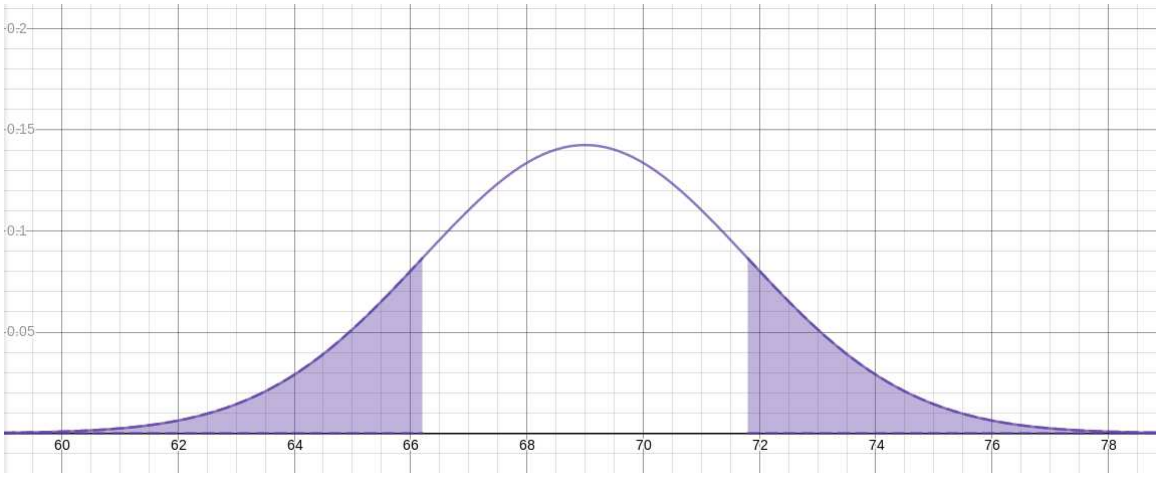
Since the event "$X < 66.2$" is disjoint from the event "$X > 71.8$" ($X$ only takes on one value at a time, so it cannot be simultaneously less than 66.2 and greater than 71.8), we can use the simple rule for addition of probabilities:

$$.32 = P(X < 66.2 \text{ or } X > 71.8) = P(X < 66.2) + P(X > 71.8) \, .$$

Now, since the density function of the Normal distribution is symmetric around the line $x = \mu_X$, the two terms on the right in the above equation are equal, which means that

$$P(X > 71.8) = \frac{1}{2}\left(P(X < 66.2) + P(X > 71.8)\right) = \frac{1}{2}.32 = .16 \, .$$

It might help to visualize the symmetry here as the equality of the two shaded areas in the following graph




Now, using the fact that $72 \approx 71.8$, we may say that

$$P(X > 72) \approx P(X > 71.8) = .16$$

which, since we know that in fact $P(X > 72) = .1419883859$, is not a completely terrible approximation.

EXAMPLE 4.3.31. Let's do one more computation in the context of the heights of adult American males, as in the immediately above Example 4.3.30, but now one in which the 68-95-99.7 Rule gives a more precise answer.

So say we are asked this time what proportion of adult American men are shorter than 63.4 inches. Why that height, in particular? Well, it's how tall archaeologists have determined King Tut was in life. [No, that's made up. It's just a good number for this problem.]

Again, looking through the values $\mu_X \pm \sigma_X$, $\mu_X \pm 2\sigma_X$, and $\mu_X \pm 3\sigma_X$, we notice that

$$63.4 = 69 - 5.6 = \mu_X - 2\sigma_X \, .$$

Therefore, to answer what fraction of adult American males are shorter than 63.4 inches amounts to asking what is the value of $P(X < \mu_X - 2\sigma_X)$.

What we know about $\mu_X \pm 2\sigma_X$ is that the probability of $X$ being between those two values is $P(\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X) = .95$. As in the previous Example, the complementary event to "$\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X$," which will have probability $.05$, consists of two pieces "$X < \mu_X - 2\sigma_X$" and "$X > \mu_X + 2\sigma_X$," which have the same area by symmetry. Therefore

$$\begin{aligned}
P(X < 63.4) &= P(X < \mu_X - 2\sigma_X) \\
&= \frac{1}{2}\left[P(X < \mu_X - 2\sigma_X) + P(X > \mu_X + 2\sigma_X)\right] \\
&= \frac{1}{2}P(X < \mu_X - 2\sigma_X \text{ or } X > \mu_X + 2\sigma_X) \quad \text{since they're disjoint} \\
&= \frac{1}{2}P((\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X)^c) \\
&= \frac{1}{2}\left[1 - P(\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X)\right] \quad \text{by prob. for complements} \\
&= \frac{1}{2}\,.05 \\
&= .025
\end{aligned}$$

Just the way finding the particular $X$ values $\mu_X \pm \sigma_X$, $\mu_X \pm 2\sigma_X$, and $\mu_X \pm 3\sigma_X$ in a particular situation would tell us the 68-95-99.7 Rule might be useful, so also would finding the probability values $.68$, $.95$, $99.7$, or their complements $.32$, $.05$, or $.003$, – or even half of one of those numbers, using the symmetry.

EXAMPLE 4.3.32. Continuing with the scenario of Example 4.3.30, let us now figure out what is the height above which there will only be .15% of the population.

Notice that .15%, or the proportion .0015, is not one of the numbers in the 68-95-99.7 Rule, nor is it one of their complements – but it is half of one of the complements, being half of .003 . Now, .003 is the complementary probability to .997, which was the probability in the range $\mu_X \pm 3\sigma_X$. As we have seen already (twice), the complementary area to that in the region between $\mu_X \pm 3\sigma_X$ consists of two thin tails which are of equal area, each of these areas being $\frac{1}{2}(1 - .997) = .0015$. This all means that the beginning of that upper tail, above which value lies .15% of the population, is the $X$ value $\mu_X + 3\sigma_X = 68 + 3 \cdot 2.8 = 77.4$.

Therefore .15% of adult American males are taller than 77.4 inches.

## Exercises

EXERCISE 4.1. A basketball player shoots four free throws, and you write down the sequence of hits and misses. Write down the sample space for thinking of this whole thing as a random experiment.

In another game, a basketball player shoots four free throws, and you write down the number of baskets she makes. Write down the sample space for this different random experiment.

EXERCISE 4.2. You take a normal, six-sided die, paint over all the sides, and then write the letter **A** on all six sides. You then roll the die. What is the sample space of this experiment? Also, list all the possible events for this experiment. *[Hint: it may help to look at Example 4.1.9.]*

Now you paint it over again, and write **A** on half the sides and **B** on the other half. Again, say what is the sample space and list all possible events.

One more time you paint over the sides, then write **A** on one third of the faces, **B** on one third of the other faces, and **C** on the remaining third. Again, give the sample space and all events.

Make a conjecture about how many events there will be if the sample space has $n$ outcomes in it.

EXERCISE 4.3. Describe a random experiment whose sample space will be the set of all points on the (standard, 2-dimensional, $xy$-) plane.

EXERCISE 4.4. The most common last [family] name in the world seems to be Wang [or the variant Wong]. Approximately 1.3% of the global population has this last name.

The most common first name in the world seems to be Mohammad [or one of several variants]. Some estimates suggest that perhaps as many as 2% of the global population has this first name.

Can you tell, from the above information, what percentage of the world population has the name "Mohammad Wang?" If so, why and what would it be? If not, why not, and can you make any guess about what that percentage would be, anyway?

*[Hint: think of all the above percentages as probabilities, where the experiment is picking a random person on Earth and asking their name. Carefully describe some events for this experiment, relevant to this problem, and say what their probabilities are. Tell how combining events will or will not compute the probability of the desired event, corresponding to the desired percentage.]*

*[Note: don't bet on the numbers given in this problem being too accurate – they might be, but there is a wide range of published values for them in public information from different sources, so probably they are only a very crude approximation.]*

EXERCISE 4.5. Suppose that when people have kids, the chance of having a boy or a girl is the same. Suppose also that the sexes of successive children in the same family are independent. [Neither of these is exactly true in real life, but let's pretend for this problem.]

The Wang family has two children. If we think of the sexes of these children as the result of a random experiment, what is the sample space? Note that we're interested in birth order as well, so that should be apparent from the sample space.

What are the probabilities of each of the outcomes in your sample space? Why?

Now suppose we know that at least one of the Wang children is a boy. Given this information, what is the probability that the Wangs have two boys?

Suppose instead that we know that the Wangs' older child is a boy. What is the probability, given this different information, that both Wang children are boys?

To solve this, clearly define events in words and with symbols, compute probabilities, and combine these to get the desired probability. Explain everything you do, of course.

EXERCISE 4.6. Imagine you live on a street with a stop light at both ends of the block. You watch cars driving down the street and notice which ones have to stop at the $1^{st}$ and/or $2^{nd}$ light (or none). After counting cars and stops for a year, you have seen what a very large number – call it $N$ – of cars did. Now imagine you decide to think about the experiment "*pick a car on this street from the last year at random and notice at which light or lights it has to stop*."

Let $A$ be the event "*the car had to stop at the* $1^{st}$ *light*" and $B$ be the event "*the car had to stop at the* $2^{nd}$ *light*." What else would you have to count, over your year of data collection, to estimate the probabilities of $A$ and of $B$? Pick some numbers for all of these variables and show what the probabilities would then be.

Make a Venn diagram of this situation. Label each of the four connected regions of this diagram (the countries, if this were a map) with a number from ① to ④, then provide a key which gives, for each of these numbered regions, **both** a formula in terms of $A$, $B$, unions, intersections, and/or complements, and then **also** a description entirely in words which do not mention $A$ or $B$ or set operations at all. Then put a decimal number in each of the regions indicating the probability of the corresponding event.

Wait – for one of the regions, you can't fill in the probability yet, with the information you've collected so far. What else would you have had to count over the data-collection year to estimate this probability? Make up a number and show what the corresponding probability would then be, and add that number to your Venn diagram.

Finally, using the probabilities you have chosen, are the events $A$ and $B$ independent? Why or why not? Explain in words what this means, in this context.

EXERCISE 4.7. Here is a table of the prizes for the **EnergyCube** Lottery:

| Prize | Odds of winning |
|---|---|
| \$1,000,000 | 1 in 12,000,000 |
| \$50,000 | 1 in 1,000,000 |
| \$100 | 1 in 10,000 |
| \$7 | 1 in 300 |
| \$4 | 1 in 25 |

We want to transform the above into the [probability] distribution of a random variable $X$.

First of all, let's make $X$ represent the **net gain** a Lottery player would have for the various outcomes of playing – note that the ticket to play costs \$2. How would you modify the above numbers to take into account the ticket costs?

Next, notice that the above table gives winning **odds**, not probabilities. How will you compute the probabilities from those odds? Recall that saying something has odds of "1 in $n$" means that it tends to happen about once out of $n$ runs of the experiment. You might use the word *frequentist* somewhere in your answer here.

Finally, something is missing from the above table of outcomes. What prize – actually the most common one! – is missing from the table, and how will you figure out its probability?

After giving all of the above explanations, now write down the full, formal, probability distribution for this "net gain in **EnergyCube** Lottery plays" random variable, $X$.

In this problem, some of the numbers are quite small and will disappear entirely if you round them. So use a calculator or computer to compute everything here and keep as much accuracy as your device shows for each step of the calculation.

EXERCISE 4.8. Continuing with the same scenario as in the previous Exercise 4.7, with the **EnergyCube** Lottery: What would be your expectation of the average gain per play of this Lottery? Explain fully, of course.

So if you were to play every weekday for a school year (so: five days a week for the 15 weeks of each semester, two semesters in the year), how much would you expect to win or lose in total?

Again, use as much accuracy as your computational device has, at every step of these calculations.

EXERCISE 4.9. Last problem in the situation of the above Exercise 4.7 about the **EnergyCube** Lottery: Suppose your friend plays the lottery and calls you to tell you that she won ... but her cell phone runs out of charge in the middle of the call, and you don't know how much she won. Given the information that she won, what is the probability that she won more than \$1,000?

Continue to use as much numerical accuracy as you can.

EXERCISE 4.10. Let's make a modified version of Example 4.3.18. You are again throwing darts at a dartboard, but you notice that you are very left-handed so your throws pull to the right much more than they pull to the left. What this means is that it is not a very good model of your dart throws just to notice how far they are from the center of the dartboard, it would be better to notice the $x$-coordinate of where the dart hits, measuring (in $cm$) with the center of the board at $x$ location $0$. This will be your new choice of RV, which you will still call $X$.

You throw repeatedly at the board, measure $X$, and find out that you *never* hit more than $10cm$ to the right of the center, while you are more accurate to the left and never hit more than $5cm$ in that direction. You do hit the middle ($X = 0$) the most often, and you guess that the probability decreases linearly to those edges where you never hit.

Explain why your $X$ is a *continuous* RV, and what its interval $[x_{min}, x_{max}]$ of values is.

Now sketch the graph of the probability density function for $X$. *[Hint: it will be a triangle, with one side along the interval of values $[x_{min}, x_{max}]$ on the $x$-axis, and its maximum at the center of the dartboard.]* Make sure that you put tick marks and numbers on the axes, enough so that the coordinates of the corners of the triangular graph can be seen easily. *[Another hint: it is a useful fact that the total area under the graph of any probability density function is $1$.]*

What is the probability that your next throw will be in the bull's-eye, whose radius, remember, is $1.5cm$ and which therefore stretches from $x$ coordinate $-1.5$ to $x$-coordinate $1.5$?

EXERCISE 4.11. Here's our last discussion of dartboards [maybe?]: One of the problems with the probability density function approaches from Example 4.3.18 and Exercise 4.10 is the assumption that the functions were *linear* (at least in pieces). It would be much more sensible to assume they were more *bell-shaped*, maybe like the Normal distribution.

Suppose your friend Mohammad Wang is an excellent dart-player. He throws at a board and you measure the $x$-coordinate of where the dart goes, as in Exercise 4.10 with the center corresponding to $x = 0$. You notice that his darts are rarely – only 5% of the time in total! – more than $5cm$ from the center of the board.

Fill in the blanks: "MW's dart hits' $x$-coordinates are an RV $X$ which is Normally distributed with mean $\mu_X =$_____ and standard deviation $\sigma_X =$_____." Explain, of course.

How often does MW completely miss the dartboard? Its radius is $10cm$.

How often does he hit the bull's-eye? Remember its radius is $1.5cm$, meaning that it stretches from $x$ coordinate $-1.5$ to $x$-coordinate $1.5$.

CHAPTER 5

# Bringing Home the Data

In this chapter, we start to get very practical on the matter of tracking down good data in the wild and bringing it home. This is actually a very large and important subject – there are entire courses and books on *Experimental Design*, *Survey Methodology*, and *Research Methods* specialized for a range of particular disciplines (medicine, psychology, sociology, criminology, manufacturing reliability, *etc.*) – so in this book we will only give a broad introduction to some of the basic issues and approaches.

The first component of this introduction will give several of the important definitions for experimental design in the most direct, simplest context: collecting sample data in an attempt to understand a single number about an entire population. As we have mentioned before, usually a population is too large or simply inaccessible and so to determine an important feature of a population of interest, a researcher must use the accessible, affordable data of a sample. If this approach is to work, the sample must be chosen carefully, so as to avoid the dreaded *bias*. The basic structure of such studies, the meaning of bias, and some of the methods to select bias-minimizing samples, are the subject of the first section of this chapter.

It is more complicated to collect data which will give evidence for *causality*, for a causal relationship between two variables under study. But we are often interested in such relationships – which drug is a more effective treatment for some illness, what advertisement will induce more people to buy a particular product, or what public policy leads to the strongest economy. In order to investigate causal relationships, it is necessary not merely to observe, but to do an actual experiment; for causal questions about human subjects, the gold standard is a *randomized, placebo-controlled, double-blind experiment*, sometimes called simply a *randomized, controlled trial [RCT]*, which we describe in the second section.

There is something in the randomized, controlled experiment which makes many people nervous: those in the control group are not getting what the experimenter likely thinks is the best treatment. So, even though society as a whole may benefit from the knowledge we get through RCTs, it almost seems as if some test subjects are being mistreated. While the scientific research community has come to terms with this apparent injustice, there are definitely experiments which could go too far and cross an important ethical lines. In fact, history has shown that a number of experiments have actually been done which we now consider to be clearly unethical. It is therefore important to state clearly some ethical

guidelines which future investigations can follow in order to be confident to avoid mistreatment of test subjects. One particular set of such guidelines for ethical experimentation on human subjects is the topic of the third and last section of this chapter.

## 5.1. Studies of a Population Parameter

Suppose we are studying some population, and in particular a variable defined on that population. We are typically interested in finding out the following kind of characteristic of our population:

DEFINITION 5.1.1. A [**population**] **parameter** is a number which is computed by knowing the values of a variable for every individual in the population.

EXAMPLE 5.1.2. If $X$ is a quantitative variable on some population, the population mean $\mu_X$ of $X$ is a population parameter – to compute this mean, you need to add together the values of $X$ for *all* of individuals in the population. Likewise, the population standard deviation $\sigma_X$ of $X$ is another parameter.

For example, we asserted in Example 4.3.28 that the heights of adult American men are $N(69, 2.8)$. Both the $69$ and $2.8$ are population parameters here.

EXAMPLE 5.1.3. If, instead, $X$ were a categorical variable on some population, then the relative frequency (also called the **population proportion**) of some value $A$ of $X$ – the fraction of the population that has that value – is another population parameter. After all, to compute this fraction, you have to look at every single individual in the population, all $N$ of them, say, and see how many of them, say $N_A$, make the $X$ take the value $A$, then compute the relative frequency $N_A/N$.

Sometimes one doesn't have to look at the specific individuals and compute that fraction $n_A/N$ to find a population proportion. For example, in Example 4.3.28, we found that $14.1988\%$ of adult American men are taller than $6$ feet, assuming, as stated above, that adult American men's heights are distributed like $N(69, 2.8)$ – using, notice, those parameters $\mu_X$ and $\sigma_X$ of the height distribution, for which the entire population must have been examined. What this means is that the relative frequency of the value *"yes"* for the categorical variable *"is this person taller than* $6$ *feet?"* is $.141988$. This relative frequency is also a parameter of the same population of adult American males.

Parameters must be thought of as fixed numbers, out there in the world, which have a single, specific value. However, they are very hard for researchers to get their hands on, since to compute a parameter, the variable values for the entire population must be measured. So while the parameter is a single, fixed value, usually that value is *unknown.*

What can (and does change) is a value coming from a sample.

DEFINITION 5.1.4. A [**sample**] **statistic** is a number which is computed by knowing the values of a variable for the individuals from only a sample.

EXAMPLE 5.1.5. Clearly, if we have a population and quantitative variable $X$, then any time we choose a sample out of that population, we get a sample mean and sample standard deviation $S_x$, both of which are statistics.

Similarly, if we instead have a categorical variable $Y$ on some population, we take a sample of size $n$ out of the population and count how many individuals in the sample – say $n_A$ – have some value $A$ for their value of $Y$, then the $n_A/n$ is a statistic (which is also called the **sample proportion** and frequently denoted $\widehat{p}$).

Two different researchers will choose different samples and so will almost certainly have different values for the statistics they compute, even if they are using the same formula for their statistic and are looking at the same population. Likewise, one researcher taking repeated samples from the same population will probably get different values each time for the statistics they compute. So we should think of a statistic as an easy, accessible number, changing with each sample we take, that is merely an estimate of the thing we want, the parameter, which is one, fixed number out in the world, but hidden from out knowledge.

So while getting sample statistics is practical, we need to be careful that they are good estimates of the corresponding parameters. Here are some ways to get better estimates of this kind:

(1) *Pick a larger sample.* This seems quite obvious, because the larger is the sample, the closer it is to being the whole population and so the better its approximating statistics will estimate the parameters of interest. But in fact, things are not really quite so simple. In many very practical situations, it would be completely infeasible to collect sample data on a sample which was anything more than a miniscule part of the population of interest. For example, a national news organization might want to survey the American population, but it would be entirely prohibitive to get more than a few thousand sample data values, out of a population of hundreds of millions – so, on the order of tenths of a percent.

Fortunately, there is a general theorem which tells us that, in the long run, one particular statistic is a good estimator of one particular parameter:

FACT 5.1.6. **The Law of Large Numbers**: Let $X$ be a quantitative variable on some population. Then as the sizes of samples (each made up of individuals chosen randomly and *independently* from the population) get bigger and bigger, the corresponding sample means $\overline{x}$ get closer and closer to the population mean $\mu_X$.

(2) *Pick a better statistic.* It makes sense to use the sample mean as a statistic to estimate the population mean and the sample proportion to estimate the population proportion. But it is less clear where the somewhat odd formula for the sample standard deviation came from – remember, it differs from the population standard deviation by having an $n-1$ in the denominator instead of an $n$. The reason, whose proof is too technical to be included here, is that the formula we gave for $S_X$ is a better estimator for $\sigma_X$ than would have be the version which simply had the same $n$ in the denominator.

In a larger sense, "picking a better statistic" is about getting higher quality estimates from your sample. Certainly using a statistic with a clever formula is one way to do that. Another is to make sure that your data is of the highest quality possible. For example, if you are surveying people for their opinions, the way you ask a question can have enormous consequences in how your subjects answer: *"Do you support a woman's right to control her own body and her reproduction?"* and *"Do you want to protect the lives of unborn children?"* are two heavy-handed approaches to asking a question about abortion. Collectively, the impacts of how a question is asked are called **wording effects**, and are an important topic social scientists must understand well.

(3) *Pick a **better** sample.* Sample quality is, in many ways, the most important and hardest issue in this kind of statistical study. What we want, of course, is a sample for which the statistic(s) we can compute give good approximations for the parameters in which we are interested. There is a name for this kind of sample, and one technique which is best able to create these good samples: *randomness*.

DEFINITION 5.1.7. A sample is said to be **representative** of its population if the values of its sample means and sample proportions for all variables relevant to the subject of the research project are good approximations of the corresponding population means and proportions.

It follows almost by definition that a representative sample is a good one to use in the process of, as we have described above, using a sample statistic as an estimate of a population parameter in which you are interested. The question is, of course, *how to get a representative sample*.

The answer is that it is extremely hard to build a procedure for choosing samples which guarantees representative samples, but there is a method – using randomness – which at least can reduce as much as possible one specific kind of problem samples might have.

DEFINITION 5.1.8. Any process in a statistical study which tends to produce results which are *systematically different* from the true values of the population parameters under investigation is called **biased**. Such a systematic deviation from correct values is called **bias**.

The key word in this definition is *systematically:* a process which has a lot of variation might be annoying to use, it might require the researcher to collect a huge amount of data to average together, for example, in order for the estimate to settle down on something near the true value – but it might nevertheless not be *biased*. A biased process might have less variation, might seem to get close to some particular value very quickly, with little data, but would never give the correct answer, because of the systematic deviation it contained.

The hard part of finding bias is to figure out what might be causing that systematic deviation in the results. When presented with a sampling method for which we wish to think about sources of possible bias, we have to get creative.

EXAMPLE 5.1.9. In a democracy, the opinion of citizens about how good a job their elected officials are doing seems like an interesting measure of the health of that democracy. At the time of this writing, approximately two months after the inauguration of the $45^{th}$ president of the United States, the widely respected Gallup polling organization reports [**Gal17**] that $56\%$ of the population approve of the job the president is doing and $40\%$ disapprove. [Presumably, $4\%$ were neutral or had no opinion.]

According to the site from which these numbers are taken,

> *"Gallup tracks daily the percentage of Americans who approve or disapprove of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults...."*

Presumably, Gallup used the sample proportion as an estimator computed with the responses from their sample of $1500$ adults. So it was a good statistic for the job, and the sample size is quite respectable, even if not a very large fraction of the entire adult American population, which is presumably the target population of this study. Gallup has the reputation for being a quite neutral and careful organization, so we can also hope that the way they worded their questions did not introduce any bias.

A source of bias that does perhaps cause some concern here is that phrase "telephone interviews." It is impossible to do telephone interviews with people who don't have telephones, so there is one part of the population they will miss completely. Presumably, also, Gallup knew that if they called during normal working days and hours, they would not get working people at home or even on cell phones. So perhaps they called also, or only, in the evenings and on weekends – but this approach would tend systematically to miss people who had to work very long and/or late hours.

So we might worry that a strategy of telephone interviews only would be biased against those who work the longest hours, and those people might tend to have similar political views. In the end, that would result in a systematic error in this sampling method.

Another potential source of bias is that even when a person is able to answer their phone, it is their choice to do so: there is little reward in taking the time to answer an opinion survey, and it is easy simply not to answer or to hang up. It is likely, then, that only those who have quite strong feelings, either positive or negative, or some other strong personal or emotional reason to take the time, will have provided complete responses to this telephone survey. This is potentially distorting, even if we cannot be sure that the effects are systematically in one direction or the other.

[Of course, Gallup pollsters have an enormous amount of experience and have presumably thought the above issues through completely and figure out how to work around it – but we have no particular reason to be completely confident in their results other than our faith in their reputation, without more details about what work-arounds they used. In science, doubt is always appropriate.]

One of the issues we just mentioned about the Gallup polling of presidential approval ratings has its own name:

DEFINITION 5.1.10. A sample selection method that involves any substantial choice of whether to participate or not suffers from what is called **voluntary sample bias**.

Voluntary sample bias is incredibly common, and yet is such a strong source of bias that it should be taken as a reason to disregard completely the supposed results of any study that it affects. Volunteers tend to have strong feelings that drive them to participate, which can have entirely unpredictable but systematic distorting influence on the data they provide. Web-based opinion surveys, numbers of *thumbs-up* or *-down* or of positive or negative comments on a social media post, percentages of people who call in to vote for or against some public statement, *etc.*, *etc.* – such widely used polling methods produce nonsensical results which will be instantly rejected by anyone with even a modest statistical knowledge. Don't fall for them!

We did promise above one technique which can robustly combat bias: randomness. Since bias is based on a *systematic* distortion of data, any method which completely breaks all systematic processes in, for example, sample selection, will avoid bias. The strongest such sampling method is as follows.

DEFINITION 5.1.11. A **simple random sample** [**SRS**] is a sample of size $n$, say, chosen from a population by a method which produces all samples of size $n$ from that population with equal probability.

It is oddly difficult to tell if a particular sample is an SRS. Given just a sample, in fact, there is no way to tell – one must ask to see the procedure that had been followed to make that sample and then check to see if that procedure would produce any subset of the population, of the same size as the sample, with equal probability. Often, it is easier to see that a sampling method *does not* make SRSs, by finding some subsets of the population which have the correct size but which the sampling method *would never choose*, meaning that they have probability zero of being chosen. That would mean some subsets of the correct size would have zero probability and others would have a positive probability, meaning that not all subsets of that size would have the same probability of being chosen.

Note also that in an SRS it is not that every *individual* has the same probability of being chosen, it must be that every *group of individuals of the size of the desired sample* has the same probability of being chosen. These are not the same thing!

EXAMPLE 5.1.12. Suppose that on Noah's Ark, the animals decide they will form an advisory council consisting of an SRS of 100 animals, to help Noah and his family run a tight ship. So a chimpanzee (because it has good hands) puts many small pieces of paper in a basket, one for each type of animal on the Ark, with the animal's name written on the paper. Then the chimpanzee shakes the basket well and picks fifty names from the basket. Both members of the breeding pair of that named type of animal are then put on the advisory council. Is this an SRS from the entire population of animals on the Ark?

First of all, each animal name has a chance of $50/N$, where $N$ is the total number of types of animals on the Ark, of being chosen. Then both the male and female of that type of animal are put on the council. In other words, every individual animal has the same probability – $50/N$ – of being on the council. And yet there are certainly collections of $100$ animals from the Ark which do not consist of $50$ breeding pairs: for example, take 50 female birds and 50 female mammals; that collection of $100$ animals has no breeding pairs at all.

Therefore this is a selection method which picks each individual for the sample with equal probability, but *not* each collection of $100$ animals with the same probability. So it is not an SRS.

With a computer, it is fairly quick and easy to generate an SRS:

FACT 5.1.13. Suppose we have a population of size $N$ out of which we want to pick an SRS of size $n$, where $n < N$. Here is one way to do so: assign every individual in the population a unique ID number, with say $d$ digits (maybe student IDs, Social Security numbers, new numbers from $1$ to $N$ chosen in any way you like – randomness not needed here, there is plenty of randomness in the next step). Have a computer generate completely random $d$-digit number, one after the other. Each time, pick the individual from the population with that ID number as a new member of the sample. If the next random number generated by the computer is a repeat of one seen before, or if it is a $d$-digit number that doesn't happen to be any individual's ID number, then simply skip to the next random number from the computer. Keep going until you have $n$ individuals in your sample.

The sample created in this way will be an SRS.

## 5.2. Studies of Causality

If we want to draw conclusions about *causality*, observations are insufficient. This is because simply seeing $B$ always follow $A$ out in the world does not tell us that $A$ causes $B$. For example, maybe they are both caused by $Z$, which we didn't notice had always happened before those $A$ and $B$, and $A$ is simply a bit faster than $B$, so it seems always to proceed, even to cause, $B$. If, on the other hand, we go out in the world and do $A$ and then always see $B$, we would have more convincing evidence that $A$ causes $B$.

Therefore, we distinguish two types of statistical studies

DEFINITION 5.2.1. An **observational study** is any statistical study in which the researchers merely look at (measure, talk to, *etc.*) the individuals in which they are interested. If, instead, the researchers also change something in the environment of their test subjects before (and possibly after and during) taking their measurements, then the study is an **experiment**.

EXAMPLE 5.2.2. A simple survey of, for example, opinions of voters about political candidates, is an observational study. If, as is sometimes done, the subject is told something like "let me read you a statement about these candidates and then ask you your opinion again" [this is an example of something called **push-polling**], then the study has become an experiment.

Note that to be considered an experiment, it is not necessary that the study use principles of good experimental design, such as those described in this chapter, merely that the researchers *do something* to their subjects.

EXAMPLE 5.2.3. If I slap my brother, notice him yelp with pain, and triumphantly turn to you and say "See, slapping hurts!" then I've done an experiment, simply because I *did something*, even if it is a stupid experiment [tiny non-random sample, no comparison, *etc., etc.*].

If I watch you slap someone, who cries out with pain, and then I make the same triumphant announcement, then I've only done an observational study, since the action taken was not by me, the "researcher."

When we do an experiment, we typically impose our intentional change on a number of test subjects. In this case, no matter the subject of inquiry, we steal a word from the medical community:

DEFINITION 5.2.4. The thing we do to the test subjects in an experiment is called the **treatment**.

**5.2.1. Control Groups.** If we are doing an experiment to try to understand something in the world, we should not simply do the interesting new treatment to all of our subjects

and see what happens. In a certain sense, if we did that, we would simply be changing the whole world (at least the world of all of our test subjects) and then doing an observational study, which, as we have said, can provide only weak evidence of causality. To really do an experiment, we must *compare* two treatments.

Therefore any real experiment involves at least two groups.

DEFINITION 5.2.5. In an experiment, the collection of test subjects which gets the new, interesting treatment is called the **experimental group**, while the remaining subjects, who get some other treatment such as simply the past common practice, are collectively called the **control group**.

When we have to put test subjects into one of these two groups, it is very important to use a selection method which has no bias. The only way to be sure of this is [as discussed before] to use a random assignment of subjects to the experimental or control group.

**5.2.2. Human-Subject Experiments: The *Placebo Effect*.** Humans are particularly hard to study, because their awareness of their environments can have surprising effects on what they do and even what happens, physically, to their bodies. This is not because people fake the results: there can be real changes in patients' bodies even when you give them a medicine which is not physiologically effective, and real changes in their performance on tests or in athletic events when you merely convince them that they will do better, *etc*.

DEFINITION 5.2.6. A beneficial consequence of some treatment which should not directly [*e.g.*, physiologically] cause an improvement is called the **Placebo Effect**. Such a "fake" treatment, which looks real but has no actual physiological effect, is called a **placebo**.

Note that even though the Placebo Effect is based on giving subjects a "fake" treatment, the effect itself *is not fake*. It is due to a complex mind-body connection which really does change the concrete, objectively measurable situation of the test subjects.

In the early days of research into the Placebo Effect, the pill that doctors would give as a placebo would look like other pills, but would be made just of sugar (glucose), which (in those quite small quantities) has essentially no physiological consequences and so is a sort of neutral dummy pill. We still often call medical placebos **sugar pills** even though now they are often made of some even more neutral material, like the starch binder which is used as a matrix containing the active ingredient in regular pills – but without any active ingredient.

Since the Placebo Effect is a real phenomenon with actual, measurable consequences, when making an experimental design and choosing the new treatment and the treatment for the control group, it is important to give the control group *something*. If they get nothing, they do not have the beneficial consequences of the Placebo Effect, so they will not have as good measurements as the experimental group, even if the experimental treatment had

no actual useful effect. So we have to equalize for both groups the benefit provided by the Placebo Effect, and give them both an treatment which looks about the same (compare pills to pills, injections to injections, operations to operations, three-hour study sessions in one format to three-hour sessions in another format, *etc.*) to the subjects.

DEFINITION 5.2.7. An experiment in which there is a treatment group and a control group, which control group is given a convincing placebo, is said to be **placebo-controlled**.

**5.2.3. Blinding.** We need one last fundamental tool in experimental design, that of keeping subjects and experimenters ignorant of which subject is getting which treatment, experimental or control. If the test subjects are aware of into which group they have been put, that mind-body connection which causes the Placebo Effect may cause a systematic difference in their outcomes: this would be the very definition of bias. So we don't tell the patients, and make sure that their control treatment looks just like the real experimental one.

It also could be a problem if the experimenter knew who was getting which treatment. Perhaps if the experimenter knew a subject was only getting the placebo, they would be more compassionate or, alternatively, more dismissive. In either case, the systematically different atmosphere for that group of subjects would again be a possible cause of bias.

Of course, when we say that the experimenter doesn't know which treatment a particular patient is getting, we mean that they do not know that at the time of the treatment. Records must be kept somewhere, and at the end of the experiment, the data is divided between control and experimental groups to see which was effective.

DEFINITION 5.2.8. When one party is kept ignorant of the treatment being administered in an experiment, we say that the information has been **blinded**. If neither subjects nor experimenters know who gets which treatment until the end of the experiment (when both must be told, one out of fairness, and one to learn something from the data that was collected), we say that the experiment was **double-blind**.

**5.2.4. Combining it all: RCTs.** This, then is the gold standard for experimental design: to get reliable, unbiased experimental data which can provide evidence of causality, the design must be as follows:

DEFINITION 5.2.9. An experiment which is

- *randomized*
- *placebo-controlled*.
- *double-blind*

is called, for short, a **randomized, controlled trial** [**RCT**] (where the "placebo-" and "double-blind" are assumed even if not stated).

**5.2.5. Confounded Lurking Variables.** A couple of last terms in this subject are quite poetic but also very important.

DEFINITION 5.2.10. A **lurking variable** is a variable which the experimenter did not put into their investigation.

So a lurking variable is exactly the thing experimenters most fear: something they didn't think of, which might or might not affect the study they are doing.

Next is a situation which also could cause problems for learning from experiments.

DEFINITION 5.2.11. Two variables are **confounded** when we cannot statistically distinguish their effects on the results of our experiments.

When we are studying something by collecting data and doing statistics, confounded variables are a big problem, because we do not know which of them is the real cause of the phenomenon we are investigating: they are statistically indistinguishable.

The combination of the two above terms is the worst thing for a research project: what if there is a lurking variable (one you didn't think to investigate) which is confounded with the variable you did study? This would be bad, because then your conclusions would apply equally well (since the variables are statistically identical in their consequences) to that thing you didn't think of ... so your results could well be completely misunderstanding cause and effect.

The problem of confounding with lurking variables is particularly bad with observational studies. In an experiment, you can intentionally choose your subjects very randomly, which means that any lurking variables should be randomly distributed with respect to any lurking variables – but controlled with respect to the variables you are studying – so if the study finds a causal relationship in your study variables, in cannot be confounded with a lurking variable.

EXAMPLE 5.2.12. Suppose you want to investigate whether fancy new athletic shoes make runners faster. If you just do an observational study, you might find that those athletes with the new shoes do run faster. But a lurking variable here could be how rich the athletes are, and perhaps if you looked at rich and poor athletes they would have the same relationship to slow and fast times as the new- *vs* old-shoe wearing athletes. Essentially, the variable *what kind of shoe is the athlete wearing* (categorical with the two values *new* and *old*) is being confounded with the lurking variable *how wealthy is the athlete*. So the conclusion about causality *fancy new shoes make them run faster* might be false, and instead the real truth might be *wealthy athletes, who have lots of support, good coaches, good nutrition, and time to devote to their sport, run faster.*

If, instead, we did an experiment, we would not have this problem. We would select athletes at random – so some would be wealthy and some not – and give half of them (the experimental group) the fancy new shoes and the other half (the control group) the old type.

If the type of shoe was the real cause of fast running, we would see that in our experimental outcome. If really it is the lurking variable of the athlete's wealth which matters, then we would see neither group would do better than the other, since they both have a mixture of wealthy and poor athletes. If the type of shoe really is the cause of fast running, then we would see a difference between the two groups, even though there were rich and poor athletes in both groups, since only one group had the fancy new shoes.

In short, experiments are better at giving evidence for causality than observational studies in large part because an experiment which finds a causal relationship between two variables cannot be confounding the causal variable under study with a lurking variable.

## 5.3. Experimental Ethics

Experiments with human subjects are technically hard to do, as we have just seen, because of things like the Placebo Effect. Even beyond these difficulties, they are hard because human subjects just don't do what we tell them, and seem to want to express their free will and autonomy.

In fact, history has many (far too many) examples of experiments done on human subjects which did not respect their humanity and autonomy – see, for example, the Wikipedia page on **unethical human experimentation** [**Wik17b**].

The ethical principles for human subject research which we give below are largely based on the idea of respecting the humanity and autonomy of the test subjects, since the lack of that respect seems to be the crucial failure of many of the generally acknowledged unethical experiments in history. Therefore the below principles should always be taken as from the point of view of the test subjects, or as if they were designed to create systems which protect those subjects. In particular, a utilitarian calculus of *the greatest good for the greatest number* might be appealing to some, but modern philosophers of experimental ethics generally do not allow the researchers to make that decision themselves. If, for example, some subjects were willing and chose to experience some negative consequences from being in a study, that might be alright, but it is never to be left up to the researcher.

**5.3.1. "Do No Harm".** The Hippocratic Oath, a version of which is thought in popular culture to be sworn by all modern doctors, is actually not used much at all today in its original form. This is actually not that strange, since it sounds quite odd and archaic[1] to modern ears – it begins

> *I swear by Apollo the physician, and Asclepius, and Hygieia and Panacea and all the gods and goddesses as my witnesses that...*

It also has the odd requirements that physicians not use a knife, and will remain celibate, *etc.*

One feature, often thought to be part of the Oath, does not exactly appear in the traditional text but is probably considered the most important promise: **First, do no harm** [sometimes seen in the Latin version, **primum nil nocere**]. This principle is often thought of as constraining doctors and other care-givers, which is why, for example, the *American Medical Association* forbids doctors from participation in executions, even when they are legal in certain jurisdictions in the United States.

It does seem like good general idea, in any case, that those who have power and authority over others should, at the very least, not harm them. In the case of human subject experimentation, this is thought of as meaning that researchers must never knowingly harm their patients, and must in fact let the patients decide what they consider harm to be.

[1] It dates from the $5^{th}$ century BCE, and is attributed to Hippocrates of Kos [**US 12**].

**5.3.2. Informed Consent.** Continuing with the idea of letting subjects decide what harms they are willing to experience or risk, one of the most important ethical principles for human subject research is that test subjects must be asked for **informed consent**. What this means is that they must be informed of all of the possible consequences, positive and (most importantly) negative, of participation in the study, and then given the right to decide if they want to participate. The information part does not have to tell every detail of the experimental design, but it must give every possible consequence that the researchers can imagine.

It is important when thinking about *informed consent* to make sure that the subjects really have the ability to exercise fully free will in their decision to give consent. If, for example, participation in the experiment is the only way to get some good (health care, monetary compensation in a poor neighborhood, a good grade in a class, advancement in their job, *etc.*) which they really need or want, the situation itself may deprive them of their ability freely to say *no* – and therefore *yes*, freely.

**5.3.3. Confidentiality.** The Hippocratic Oath does also require healers to protect the privacy of their patients. Continuing with the theme of protecting the autonomy of test subjects, then, it is considered to be entirely the choice of subject when and how much information about their participation in the experiment will be made public.

The kinds of information protected here run from, of course, the subjects' performance in the experimental activities, all the way to the simple fact of participation itself. Therefore, ethical experimenters must make it possible for subject to sign up for and then do all parts of the experiment without anyone outside the research team knowing this fact, should the subject want this kind of privacy.

As a practical matter, something must be revealed about the experimental outcomes in order for the scientific community to be able to learn something from that experiment. Typically this public information will consist of measures like sample means and other data which are *aggregated* from many test subjects' results. Therefore, even if it were know what the mean was and that a person participated in the study, the public would not be able to figure out what that person's particular result was.

If the researchers want to give more precise information about one particular test subject's experiences, or about the experiences of a small enough number of subjects that individual results could be *disaggregated* from what was published, then the subjects' identities must be hidden, or **anonymized**. This is done by removing from scientific reports all *personally identifiable information [PII]* such as name, social security or other ID number, address, phone number, email address, *etc.*

**5.3.4. External Oversight [IRB].** One last way to protect test subjects and their autonomy which is required in ethical human subject experimentation is to give some other, disinterested, external group as much power and information as the researchers themselves.

In the US, this is done by requiring all human subject experimentation to get approval from a group of trained and independent observers, called the **Institutional Review Board** [**IRB**] *before the start of the experiment*. The IRB is given a complete description of all details of the experimental design and then chooses whether or not to give its approval. In cases when the experiment continues for a long period of time (such as more than one year), progress reports must be given to the IRB and its re-approval sought.

Note that the way this IRB requirement is enforced in the US is by requiring approval by a recognized IRB for experimentation by any organization which wants ever to receive US Federal Government monies, in the form of research grants, government contracts, or even student support in schools. IRBs tend to be very strict about following rules, and if the ever see a violation at some such organization, that organization will quickly get excluded from federal funds for a very long time. As a consequence, all universities, NGOs, and research institutes in the US, and even many private organizations or companies, are very careful about proper use of IRBs.

## Exercises

EXERCISE 5.1. In practice, *wording effects* are often an extremely strong influence on the answers people give when surveyed. So... Suppose you were doing a survey of American voters opinions of the president. Think of a way of asking a question which would tend to *maximize* the number of people who said they approved of the job he is doing. Then think of another way of asking a question which would tend to *minimize* that number [who say they approve of his job performance].

EXERCISE 5.2. Think of a survey question you could ask in a survey of the general population of Americans in response to which many [most?] people would *lie*. State what would be the issue you would be investigating with this survey question, as a clearly defined, formal *variable* and *parameter* on the population of all Americans. Also tell exactly what would be the wording of the question you think would get lying responses.

Now think of a way to do an observational study which would get more accurate values for this variable and for the parameter of interest. Explain in detail.

EXERCISE 5.3. Many parents believe that their small children get a bit hyperactive when they eat or drink sweets (candies, sugary sodas, *etc.*), and so do not let their kids have such things before nap time, for example. A pediatrician at Euphoria State University Teaching Hospital [ESUTH] thinks instead that it is the parents' expectations about the effects of sugar which cause their children to become hyperactive, and not the sugar at all.

Describe a randomized, placebo-controlled, double-blind experiment which would collect data about this ESUTH pediatrician's hypothesis. Make sure you are clear about both which part of your experimental procedure addresses each of those important components of good experimental design.

EXERCISE 5.4. Is the experiment you described in the previous exercise an ethical one? What must the ESUTH pediatrician do before, during, and after the experiment to make sure it is ethical? Make sure you discuss (at least) the checklist of ethical guidelines from this chapter and how each point applies to this particular experiment.

# Part 3

# Inferential Statistics

We are now ready to make (some) inferences about the real world based on data – this subject is called **inferential statistics**. We have seen how to display and interpret 1- and 2-variable data. We have seen how to design experiments, particularly experiments whose results might tell us something about cause and effect in the real world. We even have some principles to help us do such experimentation ethically, should our subjects be human beings. Our experimental design principles use randomness (to avoid bias), and we have even studied the basics of probability theory, which will allow us to draw the best possible conclusions in the presence of randomness.

What remains to do in this part is to start putting the pieces together. In particular, we shall be interested in drawing the best possible conclusions about some population parameter of interest, based on data from a sample. Since we know always to seek simple random samples (again, to avoid bias), our inferences will be never be completely sure, instead they will be built on (a little bit of) probability theory.

The basic tools we describe for this inferential statistics are the *confidence interval* and the *hypothesis test* (also called *test of significance*). In the first chapter of this Part, we start with the easiest cases of these tools, when they are applied to inferences about the population mean of a quantitative RV. Before we do that, we have to discuss the *Central Limit Theorem [CLT]*, which is both crucial to those tools and one of the most powerful and subtle theorems of statistics.

CHAPTER 6

# Basic Inferences

The purpose of this chapter is to introduce two basic but powerful tools of inferential statistics, the *confidence interval* and the *hypothesis test* (also called *test of significance*), in the simplest case of looking for the population mean of a quantitative RV.

This simple case of these tool is based, for both of them, on a beautiful and amazing theorem called the *Central Limit Theorem*, which is therefore the subject of the first section of the chapter. The following sections then build the ideas and formulæ first for confidence intervals and then for hypothesis tests.

Throughout this chapter, we assume that we are working with some (large) population on which there is defined a quantitative RV $X$. The population mean $\sigma_X$ is, of course, a fixed number, out in the world, unchanging but also probably unknown, simply because to compute it we would have to have access to the values of $X$ for the entire population.

Strangely, we assume in this chapter that while we do not know $\mu_X$, we do know the population standard deviation $\sigma_X$, of $X$. This is actually quite a silly assumption – how could we know $\sigma_X$ if we didn't already know $\mu_X$? But we make this assumption because if makes this first version of *confidence intervals* and *hypothesis tests* particularly simple. (Later chapters in this Part will remove this silly assumption.)

Finally, we always assume in this chapter that the samples we use are simple random samples, since by now we know that those are the best kind.

## 6.1. The Central Limit Theorem

Taking the average [mean] of a sample of quantitative data is actually a very nice process: the arithmetic is simple, and the average often has the nice property of being closer to the center of the data than the values themselves being combined or averaged. This is because while a random sample may have randomly picked a few particularly large (or particularly small) values from the data, it probably also picked some other small (or large) values, so that the mean will be in the middle. It turns out that these general observations of how nice a sample mean can be explained and formalized in a very important Theorem:

FACT 6.1.1. **The Central Limit Theorem** [**CLT**] Suppose we have a large population on which is defined a quantitative random variable $X$ whose population mean is $\mu_X$ and whose population standard deviation is $\sigma_X$. Fix a whole number $n \geq 30$. As we take repeated, independent SRSs of size $n$, the distribution of the sample means $\overline{x}$ of these SRSs is approximately $N(\mu_X, \sigma_X/\sqrt{n})$. That is, the distribution of $\overline{x}$ is approximately Normal with mean $\mu_X$ and standard deviation $\sigma_X/\sqrt{n}$.

Furthermore, as $n$ gets bigger, the Normal approximation gets better.

Note that the CLT has several nice pieces. First, it tells us that the middle of the histogram of sample means, as we get repeated independent samples, is the same as the mean of the original population – *the mean of the sample means is the population mean*. We might write this as $\mu_{\overline{x}} = \mu_X$.

Second, the CLT tells us precisely how much less variation there is in the sample means because of the process noted above whereby averages are closer to the middle of some data than are the data values themselves. The formula is $\sigma_{\overline{x}} = \sigma_X/\sqrt{n}$.

Finally and most amazingly, the CLT actually tells us exactly what is the shape of the distribution for $\overline{x}$ – and it turns out to be that complicated formula we gave Definition 4.3.19. This is completely unexpected, but somehow the universe knows that formula for the Normal distribution density function and makes it appear when we construct the histogram of sample means.

Here is an example of how we use the CLT:

EXAMPLE 6.1.2. We have said elsewhere that adult American males' heights in inches are distributed like $N(69, 2.8)$. Supposing this is true, let us figure out what is the probability that 52 randomly chosen adult American men, lying down in a row with each one's feet touching the next one's head, stretch the length of a football field. [Why 52? Well, an American football team may have up to 53 people on its active roster, and one of them has to remain standing to supervise everyone else's formation lying on the field....]

First of all, notice that a football field is 100 yards long, which is 300 feet or 3600 inches. If every single one of our randomly chosen men was exactly the average height for

adult men, that would a total of $52 * 69 = 3588$ inches, so they would not stretch the whole length. But there is variation of the heights, so maybe it will happen sometimes....

So imagine we have chosen 52 random adult American men. Measure each of their heights, and call those numbers $x_1, x_2, \ldots, x_{52}$. What we are trying to figure out is whether $\sum x_i \geq 3600$. More precisely, we want to know

$$P\left(\sum x_i \geq 3600\right) .$$

Nothing in that looks familiar, but remember that the 52 adult men were chosen randomly. The best way to choose some number, call it $n = 52$, of individuals from a population is to choose an SRS of size $n$.

Let's also assume that we did that here. Now, having an SRS, we know from the CLT that the sample mean $\overline{x}$ is $N(69, 2.8/\sqrt{52})$ or, doing the arithmetic, $N(69, .38829)$.

But the question we are considering here doesn't mention $\overline{x}$, you cry! Well, it almost does: $\overline{x}$ is the sample mean given by

$$\overline{x} = \frac{\sum x_i}{n} = \frac{\sum x_i}{52} .$$

What that means is that the inequality

$$\sum x_i \geq 3600$$

amounts to exactly the same thing, by dividing both sides by 52, as the inequality

$$\frac{\sum x_i}{52} \geq \frac{3600}{52}$$

or, in other words,

$$\overline{x} \geq 69.23077 .$$

Since these inequalities all amount to the same thing, they have the same probabilities, so

$$P\left(\sum x_i \geq 3600\right) = P\left(\overline{x} \geq 69.23077\right) .$$

But remember $\overline{x}$ was $N(69, .38829)$, so we can calculate this probability with **LibreOffice Calc** or **Microsoft Excel** as

$$\begin{aligned} P\left(\overline{x} \geq 69.23077\right) &= 1 - P\left(\overline{x} < 69.23077\right) \\ &= \texttt{NORM.DIST}(69.23077, 69, .38829, 1) \\ &= .72385 \end{aligned}$$

where here we first use the probability rule for complements to turn around the inequality into the direction that `NORM.DIST` calculates.

Thus the chance that 52 randomly chosen adult men, lying in one long column, are as long as a football field, is 72.385%.

## 6.2. Basic Confidence Intervals

As elsewhere in this chapter, we assume that we are working with some (large) population on which there is defined a quantitative RV $X$. The population mean $\mu_X$ is unknown, and we want to estimate it. world, unchanging but also probably unknown, simply because to compute it we would have to have access to the values of $X$ for the entire population.

We continue also with our strange assumption that while we do not know $\mu_X$, we do know the population standard deviation $\sigma_X$, of $X$.

Our strategy to estimate $\mu_X$ is to take an SRS of size $n$, compute the sample mean $\overline{x}$ of $X$, and then to guess that $\mu_X \approx \overline{x}$. But this leaves us wondering how good an approximation $\overline{x}$ is of $\mu_X$.

The strategy we take for this is to figure how close $\mu_X$ must be to $\overline{x}$ – or $\overline{x}$ to $\mu_X$, it's the same thing, and in fact to be precise enough to say what is the probability that $\mu_X$ is a certain distance from $\overline{x}$. That is, if we choose a target probability, call it $L$, we want to make an interval of real numbers centered on $\overline{x}$ with the probability of $\mu_X$ being in that interval being $L$.

Actually, that is not really a sensible thing to ask for: probability, remember, is the fraction of times something happens in repeated experiments. But we are not repeatedly choosing $\mu_X$ and seeing if it is in that interval. Just the opposite, in fact: $\mu_X$ is fixed (although unknown to us), and every time we pick a new SRS – that's the repeated experiment, choosing new SRSs! – we can compute a new interval and hope that that new interval might contain $\mu_X$. The probability $L$ will correspond to what fraction of those newly computed intervals which contain the (fixed, but unknown) $\mu_X$.

How to do even this?

Well, the Central Limit Theorem tells us that the distribution of $\overline{x}$ as we take repeated SRSs – exactly the repeatable experiment we are imagining doing – is approximately Normal with mean $\mu_X$ and standard deviation $\sigma_X/\sqrt{n}$. By the 68-95-99.7 Rule:

(1) 68% of the time we take samples, the resulting $\overline{x}$ will be within $\sigma_X/\sqrt{n}$ units on the number line of $\mu_X$. Equivalently (since the distance from A to B is the same as the distance from B to A!), 68% of the time we take samples, $\mu_X$ will be within $\sigma_X/\sqrt{n}$ of $\overline{x}$. In other words, 68% of the time we take samples, $\mu_X$ will happen to lie in the interval from $\overline{x} - \sigma_X/\sqrt{n}$ to $\overline{x} + \sigma_X/\sqrt{n}$.
(2) Likewise, 95% of the time we take samples, the resulting $\overline{x}$ will be within $2\sigma_X/\sqrt{n}$ units on the number line of $\mu_X$. Equivalently (since the distance from A to B is still the same as the distance from B to A!), 95% of the time we take samples, $\mu_X$ will be within $2\sigma_X/\sqrt{n}$ of $\overline{x}$. In other words, 95% of the time we take samples, $\mu_X$ will happen to lie in the interval from $\overline{x} - 2\sigma_X/\sqrt{n}$ to $\overline{x} + 2\sigma_X/\sqrt{n}$.
(3) Likewise, 99.7% of the time we take samples, the resulting $\overline{x}$ will be within $3\sigma_X/\sqrt{n}$ units on the number line of $\mu_X$. Equivalently (since the distance from A

to B is still the same as the distance from B to A!), 99.7% of the time we take samples, $\mu_X$ will be within $3\sigma_X/\sqrt{n}$ of $\overline{x}$. In other words, 99.7% of the time we take samples, $\mu_X$ will happen to lie in the interval from $\overline{x} - 3\sigma_X/\sqrt{n}$ to $\overline{x} + 3\sigma_X/\sqrt{n}$.

Notice the general shape here is that the interval goes from $\overline{x} - z_L^*\sigma_X/\sqrt{n}$ to $\overline{x} + z_L^*\sigma_X/\sqrt{n}$, where this number $z_L^*$ has a name:

DEFINITION 6.2.1. The **critical value $z_L^*$ with probability** $L$ for the Normal distribution is the number such that the Normal distribution $N(\mu_X, \sigma_X)$ has probability $L$ between $\mu_X - z_L^*\sigma_X$ and $\mu_X + z_L^*\sigma_X$.

Note the probability $L$ in this definition is usually called the **confidence level**.

If you think about it, the 68-95-99.7 Rule is exactly telling us that $z_L^* = 1$ if $L = .68$, $z_L^* = 2$ if $L = .95$, and $z_L^* = 3$ if $L = .997$. It's actually convenient to make a table of similar values, which can be calculated on a computer from the formula for the Normal distribution.

FACT 6.2.2. Here is a useful table of critical values for a range of possible confidence levels:

| $L$ | .5 | .8 | .9 | .95 | .99 | .999 |
|---|---|---|---|---|---|---|
| $z_L^*$ | .674 | 1.282 | 1.645 | 1.960 | 2.576 | 3.291 |

Note that, oddly, the $z_L^*$ here for $L = .95$ is not 2, but rather $1.96$! This is actually more accurate value to use, which you may choose to use, or you may continue to use $z_L^* = 2$ when $L = .95$, if you like, out of fidelity to the 68-95-99.7 Rule.

Now, using these accurate critical values we can define an interval which tells us where we expect the value of $\mu_X$ to lie.

DEFINITION 6.2.3. For a probability value $L$, called the **confidence level**, the interval of real numbers from $\overline{x} - z_L^*\sigma_X/\sqrt{n}$ to $\overline{x} + z_L^*\sigma_X/\sqrt{n}$ is called the **confidence interval for $\mu_X$ with confidence level** $L$.

The meaning of *confidence* here is quite precise (and a little strange):

FACT 6.2.4. Any particular confidence interval with confidence level $L$ might or might not actually contain the sought-after parameter $\mu_X$. Rather, what it means to have confidence level $L$ is that if we take repeated, independent SRSs and compute the confidence interval again for each new $\overline{x}$ from the new SRSs, then a fraction of size $L$ of these new intervals will contain $\mu_X$.

Note that any particular confidence interval might or might not contain $\mu_X$ not because $\mu_X$ is moving around, but rather the SRSs are different each time, so the $\overline{x}$ is (potentially) different, and hence the interval is moving around. The $\mu_X$ is fixed (but unknown), while the confidence intervals move.

Sometimes the piece we add and subtract from the $\overline{x}$ to make a confidence interval is given a name of its own:

DEFINITION 6.2.5. When we write a confidence interval for the population mean $\mu_X$ of some quantitative variable $X$ in the form $\overline{x} - E$ to $\overline{x} + E$, where $E = z_L^* \sigma_X / \sqrt{n}$, we call $E$ the **margin of error** [or, sometimes, the **sampling error**] of the confidence interval.

Note that if a confidence interval is given without a stated confidence level, particularly in the popular press, we should assume that the implied level was .95 .

**6.2.1. Cautions.** The thing that most often goes wrong when using confidence intervals is that the sample data used to compute the sample mean $\overline{x}$ and then the endpoints $\overline{x} \pm E$ of the interval is not from a good SRS. It is hard to get SRSs, so this is not unexpected. But we nevertheless frequently assume that some sample is an SRS, so that we can use it to make a confidence interval, even of that assumption is not really justified.

Another thing that can happen to make confidence intervals less accurate is to choose too small a sample size $n$. We have seen that our approach to confidence intervals relies upon the CLT, therefore it typically needs samples of size at least 30.

EXAMPLE 6.2.6. A survey of 463 first-year students at Euphoria State University [ESU] found that the [sample] average of how long they reported studying per week was 15.3 hours. Suppose somehow we know that the population standard deviation of hours of study per week at ESU is 8.5 . Then we can find a confidence interval at the 99% confidence level for the mean study per week of all first-year students by calculating the margin of error to be

$$E == z_L^* \sigma_X / \sqrt{n} = 2.576 \cdot 8.5 / \sqrt{463} = 1.01759$$

and then noting that the confidence interval goes from

$$\overline{x} - E = 15.3 - 1.01759 = 14.28241$$

to

$$\overline{x} + E = 15.3 + 1.01759 = 16.31759 \, .$$

Note that for this calculation to be doing what we want it to do, we must assume that the 463 first-year students were an SRS out of the entire population of first-year students at ESU.

Note also that what it means that we have 99% confidence in this interval from 14.28241 to 16.31759 (or $[14.28241, 16.31759]$ in interval notation) is not, in fact, that we any confidence at all in those particular numbers. Rather, we have confidence in the *method*, in the sense that if we imagine independently taking many future SRSs of size 463 and recalculating new confidence intervals, then 99% of these future intervals will contain the one, fixed, unknown $\mu_X$.

## 6.3. Basic Hypothesis Testing

Let's start with a motivating example, described somewhat more casually than the rest of the work we usually do, but whose logic is exactly that of the scientific standard for hypothesis testing.

EXAMPLE 6.3.1. Suppose someone has a coin which they claim is a fair coin (including, in the informal notion of a fair coin, that successive flips are independent of each other). You care about this fairness perhaps because you will use the coin in a betting game.

How can you know if the coin really is fair?

Obviously, your best approach is to start flipping the coin and see what comes up. If the first flip shows *heads* [*H*], you wouldn't draw any particular conclusion. If the second was also an *H*, again, so what? If the third was still *H*, you're starting to think there's a run going. If you got all the way to ten *H*s in a row, you would be very suspicious, and if the run went to 100 *H*s, you would demand that some other coin (or person doing the flipping) be used.

Somewhere between two and 100 *H*s in a row, you would go from bland acceptance of fairness to nearly complete conviction that this coin is not fair – why? After all, the person flipping the coin and asserting its fairness could say, correctly, that it is possible for a fair coin to come up *H* any number of times in a row. Sure, you would reply, but it is very unlikely: that is, given that the coin is fair, the conditional probability that those long runs without *T*s would occur is very small.

Which in turn also explains how you would draw the line, between two and 100 *H*s in a row, for when you thought the the improbability of that particular run of straight *H*s was past the level you would be willing to accept. Other observers might draw the line elsewhere, in fact, so there would not be an absolutely sure conclusion to the question of whether the coin was fair or not.

It might seem that in the above example we only get a probabilistic answer to a yes/no question (is the coin fair or not?) simply because the thing we are asking about is, by nature, a random process: we cannot predict how any particular flip of the coin will come out, but the long-term behavior is what we are asking about; no surprise, then, that the answer will involve likelihood. But perhaps other scientific hypotheses will have more decisive answers, which do not invoke probability.

Unfortunately, this will not be the case, because we have seen above that it is wise to introduce probability into an experimental situation, even if it was not there originally, in order to avoid bias. Modern theories of science (such as quantum mechanics, and also, although in a different way, epidemiology, thermodynamics, genetics, and many other sciences) also have some amount of randomness built into their very foundations, so we should expect probability to arise in just about every kind of data.

Let's get a little more formal and careful about what we need to do with hypothesis testing.

### 6.3.1. The Formal Steps of Hypothesis Testing.

(1) State what is the population under study.

(2) State what is the variable of interest for this population. *For us in this section, that will always be a quantitative variable $X$.*

(3) State which is the resulting population parameter of interest. *For us in this section, that will always be the population mean $\mu_X$ of $X$.*

(4) State two hypotheses about the value of this parameter. One, called the **null hypothesis** and written $H_0$, will be a statement that the parameter of interest has a particular value, so

$$H_0 : \mu_X = \mu_0$$

where $\mu_0$ is some specific number. The other is the interesting alternative we are considering for the value of that parameter, and is thus called the **alternative hypothesis**, written $H_a$. The alternative hypothesis can have one of three forms:

$$H_a : \mu_X < \mu_0 \ ,$$
$$H_a : \mu_X > \mu_0 \ , \text{ or}$$
$$H_a : \mu_X \neq \mu_0 \ ,$$

where $\mu_0$ is the same specific number as in $H_0$.

(5) Gather data from an SRS and compute the sample statistic which is best related to the parameter of interest. *For us in this section, that will always be the sample mean $\overline{X}$*

(6) Compute the following conditional probability

$$p = P\left(\begin{array}{c}\text{getting values of the statistic which are as extreme,} \\ \text{or more extreme, as the ones you did get}\end{array} \middle| \ H_0\right) \ .$$

This is called the **$p$-value of the test**.

(7) If the $p$-value is sufficiently small – typically, $p < .05$ or even $p < .01$ – announce

*"We reject $H_0$, with $p = \langle$number here$\rangle$."*

Otherwise, announce

*"We fail to reject $H_0$, with $p = \langle$number here$\rangle$."*

(8) Translate the result just announced into the language of the original question. As you do this, you can say *"There is strong statistical evidence that ..."* if the $p$-value is very small, while you should merely say something like *"There is evidence that..."* if the $p$-value is small but not particularly so.

Note that the hypotheses $H_0$ and $H_a$ are *statements*, not numbers. So ***don't*** *write something like* $H_0 = \mu_X = 17$; you might use

$$H_0 = \text{“}\mu_X = 17\text{”}$$

or

$$H_o : \mu_X = 17$$

(we always use the latter in this book).

**6.3.2. How Small is Small Enough, for $p$-values?** Remember how the $p$-value is defined:

$$p = P\left(\begin{array}{c}\text{getting values of the statistic which are as extreme,} \\ \text{or more extreme, as the ones you did get}\end{array} \middle| \; H_0\right) .$$

In other words, if the null hypothesis is true, maybe the behavior we saw with the sample data would sometimes happen, but if the probability is very small, it starts to seem that, under the assumption $H_0$ is true, the sample behavior was a crazy fluke. If the fluke is crazy enough, we might want simply to say that since the sample behavior actually happened, it makes us doubt that $H_0$ is true at all.

For example, if $p = .5$, that means that under the assumption $H_0$ is true, we would see behavior like that of the sample about every other time we take an SRS and compute the sample statistic. Not much of a surprise.

If the $p = .25$, that would still be behavior we would expect to see in about one out of every four SRSs, when the $H_0$ is true.

When $p$ gets down to $.1$, that is still behavior we expect to see about one time in ten, when $H_0$ is true. That's rare, but we wouldn't want to bet anything important on it.

Across science, in legal matters, and definitely for medical studies, we start to reject $H_0$ when $p < .05$. After all, if $p < .05$ and $H_0$ is true, then we would expect to see results as extreme as the ones we saw in fewer than one SRS out of 20.

There is some terminology for these various cut-offs.

DEFINITION 6.3.2. When we are doing a hypothesis test and get a $p$-value which satisfies $p < \alpha$, for some real number $\alpha$, we say the data are **statistically significant at level** $\alpha$. Here the value $\alpha$ is called the **significance level** of the test, as in the phrase *"We reject* $H_0$ *at significance level* $\alpha$*,"* which we would say if $p < \alpha$.

EXAMPLE 6.3.3. If we did a hypothesis test and got a $p$-value of $p = .06$, we would say about it that the result was statistically significant at the $\alpha = .1$ level, but not statistically significant at the $\alpha = .05$ level. In other words, we would say *"We reject the null hypothesis at the* $\alpha = .1$ *level,"* but also *"We fail to reject the null hypothesis at the* $\alpha = .05$ *level,"*.

FACT 6.3.4. The courts in the United States, as well as the majority of standard scientific and medical tests which do a formal hypothesis test, use the significance level of $\alpha = .05$.

In this chapter, when not otherwise specified, we will use that value of $\alpha = .05$ as a default significance level.

EXAMPLE 6.3.5. We have said repeatedly in this book that the heights of American males are distributed like $N(69, 2.8)$. Last semester, a statistics student named Mohammad Wong said he thought that had to be wrong, and decide to do a study of the question. MW is a bit shorter than 69 inches, so his conjecture was that the mean height must be less, also. He measured the heights of all of the men in his statistics class, and was surprised to find that the average of those 16 men's heights was 68 inches (he's only 67 inches tall, and he thought he was typical, at least for his class[1]). Does this support his conjecture or not?

Let's do the formal hypothesis test.

The population that makes sense for this study would be all adult American men today – MW isn't sure if the claim of American men's heights having a population mean of 69 inches was *always* wrong, he is just convinced that it is wrong *today*.

The quantitative variable of interest on that population is their height, which we'll call $X$.

The parameter of interest is the population mean $\mu_X$.

The two hypotheses then are

$$H_0 : \mu_X = 69 \quad \text{and}$$
$$H_a : \mu_X < 69 \, ,$$

where the basic idea in the null hypothesis is that the claim in this book of men's heights having mean 69 is true, while the new idea which MW hopes to find evidence for, encoded in alternative hypothesis, is that the true mean of today's men's heights is less than 69 inches (like him).

MW now has to make two bad assumptions: the first is that the 16 students in his class are an SRS drawn from the population of interest; the second, that the population standard deviation of the heights of individuals in his population of interest is the same as the population standard deviation of the group of all adult American males asserted elsewhere in this book, 2.8 . These are definitely **bad assumptions** – particularly that MW's male classmates are an SRS of the population of today's adult American males – but he has to make them nevertheless in order to get somewhere.

The sample mean height $\overline{X}$ for MW's SRS of size $n = 16$ is $\overline{X} = 68$.

[1]When an experimenter tends to look for information which supports their prior ideas, it's called **confirmation bias** – MW may have been experiencing a bit of this bias when he mistakenly thought he was average in height for his class.

MW can now calculate the $p$-value of this test, using the Central Limit Theorem. According to the CLT, the distribution of $\overline{X}$ is $N(69, 2.8/\sqrt{16})$. Therefore the $p$-value is

$$p = P\left(\begin{array}{c}\text{MW would get values of } \overline{X} \text{ which are as} \\ \text{extreme, or more extreme, as the ones he did get}\end{array}\middle| \; H_0\right) = P(\overline{X} < 69)\,.$$

Which, by what we just observed the CLT tells us, is computable by

$$\textbf{normalcdf}(-9999, 68, 69, 2.8/\sqrt{16})$$

on a calculator, or

```
NORM.DIST(68, 69, 2.8/SQRT(16), 1)
```

in a spreadsheet, either of which gives a value around .07656 .

This means that if MW uses the 5% significance level, as we often do, the result is not statistically significant. Only at the much cruder 10% significance level would MW say that he rejects the null hypothesis.

In other words, he might conclude his project by saying

> *"My research collected data about my conjecture which was statistically insignificant at the 5% significance level but the data, significant at the weaker 10% level, did indicate that the average height of American men is less than the 69 inches we were told it is ($p = .07656$)."*

People who talk to MW about his study should have additional concerns about his assumptions of having an SRS and of the value of the population standard deviation

**6.3.3. Calculations for Hypothesis Testing of Population Means.** We put together the ideas in §6.3.1 above and the conclusions of the Central Limit Theorem to summarize what computations are necessary to perform:

FACT 6.3.6. Suppose we are doing a formal hypothesis test with variable $X$ and parameter of interest the population mean $\mu_X$. Suppose that somehow we know the population standard deviation $\sigma_X$ of $X$. Suppose the null hypothesis is

$$H_0 : \mu_X = \mu_0$$

where $\mu_0$ is a specific number. Suppose also that we have an SRS of size $n$ which yielded the sample mean $\overline{X}$. Then exactly one of the following three situations will apply:

(1) If the alternative hypothesis is $H_a : \mu_X < \mu_0$ then the $p$-value of the test can be calculated in any of the following ways
   (a) the area to the left of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
   (b) **normalcdf**$(-9999, \overline{X}, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
   (c) `NORM.DIST(`$\overline{X}$`, `$\mu_0$`, `$\sigma_X$`/SQRT(n), 1)` on a spreadsheet.

(2) If the alternative hypothesis is $H_a : \mu_X > \mu_0$ then the $p$-value of the test can be calculated in any of the following ways

(a) the area to the right of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
(b) **normalcdf**$(\overline{X}, 9999, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
(c) `1-NORM.DIST(`$\overline{X}$`, `$\mu_0$`, `$\sigma_X$`/SQRT(`$n$`), 1)` on a spreadsheet.

(3) If the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ then the $p$-value of the test can be found by using the approach in exactly one of the following three situations:
(a) If $\overline{X} < \mu_0$ then $p$ is calculated by any of the following three ways:
(i) two times the area to the left of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
(ii) 2 * **normalcdf**$(-9999, \overline{X}, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
(iii) `2*NORM.DIST(`$\overline{X}$`, `$\mu_0$`, `$\sigma_X$`/SQRT(`$n$`), 1)` on a spreadsheet.
(b) If $\overline{X} > \mu_0$ then $p$ is calculated by any of the following three ways:
(i) two times the area to the right of $\overline{X}$ under the graph of a $N(\mu_0, \sigma_X/\sqrt{n})$ distribution,
(ii) 2 * **normalcdf**$(\overline{X}, 9999, \mu_0, \sigma_X/\sqrt{n})$ on a calculator, or
(iii) `2*(1-NORM.DIST(`$\overline{X}$`, `$\mu_0$`, `$\sigma_X$`/SQRT(`$n$`), 1))` on a spreadsheet.
(c) If $\overline{X} = \mu_0$ then $p = 1$.

Note the reason that there is that multiplication by two if the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ is that there are two directions – the distribution has two tails – in which the values can be more extreme than $\overline{X}$. For this reason we have the following terminology:

DEFINITION 6.3.7. If we are doing a hypothesis test and the alternative hypothesis is $H_a : \mu_X > \mu_0$ or $H_a : \mu_X < \mu_0$ then this is called a **one-tailed test**. If, instead, the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ then this is called a **two-tailed test**.

EXAMPLE 6.3.8. Let's do one very straightforward example of a hypothesis test:

A cosmetics company fills its best-selling 8-ounce jars of facial cream by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ounces per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either underfill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value .22 ounce. A quality control engineer randomly selects 30 jars from the assembly line each day to check the amounts filled. One day, the sample mean is $\overline{X} = 8.2$ ounces. Let us see if there is sufficient evidence in this sample to indicate, at the 1% level of significance, that the machine should be recalibrated.

The population under study is all of the jars of facial cream on the day of the 8.2 ounce sample.

The variable of interest is the weight $X$ of the jar in ounces.

The population parameter of interest is the population mean $\mu_X$ of $X$.

The two hypotheses then are

$$H_0 : \mu_X = 8.1 \quad \text{and}$$
$$H_a : \mu_X \neq 8.1 \, .$$

The sample mean is $\overline{X} = 8.2$, and the sample – which we must assume to be an SRS – is of size $n = 30$.

Using the case in Fact 6.3.6 where the alternative hypothesis is $H_a : \mu_X \neq \mu_0$ and the sub-case where $\overline{X} > \mu_0$, we compute the $p$-value by

$$\texttt{2 * (1-NORM.DIST(8.2, 8.1, .22/SQRT(30), 1))}$$

on a spreadsheet, which yields $p = .01278$.

Since $p$ is not less than .01, we fail to reject $H_0$ at the $\alpha = .01$ level of significance.

The quality control engineer should therefore say to company management

> *"Today's sample, though off weight, was not statistically significant at the stringent level of significance of $\alpha = .01$ that we have chosen to use in these tests, that the jar-filling machine is in need of recalibration today ($p = .01278$)."*

**6.3.4. Cautions.** As we have seen before, the requirement that the sample we are using in our hypothesis test is a valid SRS is quite important. But it is also quite hard to get such a good sample, so this is often something that can be a real problem in practice, and something which we must assume is true with often very little real reason.

It should be apparent from the above Facts and Examples that most of the work in doing a hypothesis test, after careful initial set-up, comes in computing the $p$-value.

Be careful of the phrase *statistically significant*. It does not mean that the effect is large! There can be a very small effect, the $\overline{X}$ might be very close to $\mu_0$ and yet we might reject the null hypothesis if the population standard deviation $\sigma_X$ were sufficiently small, or even if the sample size $n$ were large enough that $\sigma_X/\sqrt{n}$ became very small. Thus, oddly enough, a statistically significant result, one where the conclusion of the hypothesis test was statistically quite certain, might not be *significant* in the sense of mattering very much. With enough precision, we can be very sure of small effects.

Note that the meaning of the $p$-value is explained above in its definition as a conditional probability. So $p$ ***does not*** compute the probability that the null hypothesis $H_0$ is true, or any such simple thing. In contrast, the Bayesian approach to probability, which we chose not to use in the book, in favor of the frequentist approach, does have a kind of hypothesis test which includes something like the direct probability that $H_0$ is true. But we did not follow the Bayesian approach here because in many other ways it is more confusing.

In particular, one consequence of the real meaning of the $p$-value as we use it in this book is that sometimes we will reject a true null hypothesis $H_0$ just out of bad luck. In

fact, if $p$ is just slightly less than $.05$, we would reject $H_0$ at the $\alpha = .05$ significance level even though, in slightly less than one case in 20 (meaning 1 SRS out of 20 chosen independently), we would do this rejection even though $H_0$ was true.

We have a name for this situation.

DEFINITION 6.3.9. When we reject a true null hypothesis $H_0$ this is called a **type I error**. Such an error is usually (but not always: it depends upon how the population, variable, parameter, and hypotheses were set up) a **false positive**, meaning that something exciting and new (or scary and dangerous) was found even though it is not really present in the population.

EXAMPLE 6.3.10. Let us look back at the cosmetic company with a jar-filling machine from Example 6.3.8. We don't know what the median of the SRS data was, but it wouldn't be surprising if the data were symmetric and therefore the median would be the same as the sample mean $\overline{X} = 8.2$. That means that there were at least 15 jars with 8.2 ounces of cream in them, even though the jars are all labelled "8oz." The company is giving way at least $.2 \times 15 = 3$ ounces of the very valuable cream – in fact, probably much more, since that was simply the overfilling in that one sample.

So our intrepid quality assurance engineer might well propose to management to increase the significance level $\alpha$ of the testing regime in the factory. It is true that with a larger $\alpha$, it will be easier for simple randomness to result in type I errors, but unless the recalibration process takes a very long time (and so results in fewer jars being filled that day), the cost-benefit analysis probably leans towards fixing the machine slightly too often, rather than waiting until the evidence is extremely strong it must be done.

## Exercises

EXERCISE 6.1. You buy seeds of one particular species to plant in your garden, and the information on the seed packet tells you that, based on years of experience with that species, the mean number of days to germination is 22, with standard deviation 2.3 days.

What is the population and variable in that information? What parameter(s) and/or statistic(s) are they asserting have particular values? Do you think they can really know the actual parameter(s) and/or statistic's(s') value(s)? Explain.

You plant those seeds on a particular day. What is the probability that the first plant closest to your house will germinate within half a day of the reported mean number of days to germination – that is, it will germinate between 21.5 and 22.5 after planting?

You are interested in the whole garden, where you planted 160 seeds, as well. What is the probability that the average days to germination of all the plants in your garden is between 21.5 and 22.5 days? How do you know you can use the Central Limit Theorem to answer this problem – what must you assume about those 160 seeds from the seed packet in order for the CLT to apply?

EXERCISE 6.2. You decide to expand your garden and buy a packet of different seeds. But the printing on the seed packet is smudged so you can see that the standard deviation for the germination time of that species of plant is 3.4 days, but you cannot see what the mean germination time is.

So you plant 100 of these new seeds and note how long each of them takes to germinate: the average for those 100 plants is 17 days.

What is a 90% confidence interval for the population mean of the germination times of plants of this species? Show and explain all of your work. What assumption must you make about those 100 seeds from the packet in order for your work to be valid?

What does it mean that the interval you gave had *90% confidence*? Answer by talking about what would happen if you bought many packets of those kinds of seeds and planted 100 seeds in each of a bunch of gardens around your community.

EXERCISE 6.3. An SRS of size 120 is taken from the student population at the very large Euphoria State University [ESU], and their GPAs are computed. The sample mean GPA is 2.71 . Somehow, we also know that the population standard deviation of GPAs at ESU is .51 . Give a confidence interval at the 90% confidence level for the mean GPA of all students at ESU.

You show the confidence interval you just computed to a fellow student who is not taking statistics. They ask, "Does that mean that 90% of students at ESU have a GPA which is between $a$ and $b$?" where $a$ and $b$ are the lower and upper ends of the interval you computed. Answer this question, explaining why if the answer is *yes* and both why not and what is a better way of explaining this 90% confidence interval, if the answer is *no*.

EXERCISE 6.4. The recommended daily calorie intake for teenage girls is 2200 calories per day. A nutritionist at Euphoria State University believes the average daily caloric intake of girls in her state to be lower because of the advertising which uses underweight models targeted at teenagers. Our nutritionist finds that the average of daily calorie intake for a random sample of size $n = 36$ of teenage girls is 2150.

Carefully set up and perform the hypothesis test for this situation and these data. You may need to know that our nutritionist has been doing studies for years and has found that the standard deviation of calorie intake per day in teenage girls is about 200 calories.

Do you have confidence the nutritionist's conclusions? What does she need to be careful of, or to assume, in order to get the best possible results?

EXERCISE 6.5. The medication most commonly used today for post-operative pain relieve after minor surgery takes an average of 3.5 minutes to ease patients' pain, with a standard deviation of 2.1 minutes. A new drug is being tested which will hopefully bring relief to patients more quickly. For the test, 50 patients were randomly chosen in one hospital after minor surgeries. They were given the new medication and how long until their pain was relieved was timed: the average in this group was 3.1 minutes. Does this data provide statistically significant evidence, at the 5% significance level, that the new drug acts more quickly than the old?

Clearly show and explain all your set-up and work, of course!

EXERCISE 6.6. The average household size in a certain region several years ago was 3.14 persons, while the standard deviation was .82 persons. A sociologist wishes to test, at the 5% level of significance, whether the mean household size is different now. Perform the test using new information collected by the sociologist: in a random sample of 75 households this past year, the average size was 2.98 persons.

EXERCISE 6.7. A medical laboratory claims that the mean turn-around time for performance of a battery of tests on blood samples is 1.88 business days. The manager of a large medical practice believes that the actual mean is larger. A random sample of 45 blood samples had a mean of 2.09 days. Somehow, we know that the population standard deviation of turn-around times is 0.13 day. Carefully set up and perform the relevant test at the 10% level of significance. Explain everything, of course.

# Bibliography


[Gal17] Gallup, *Presidential Job Approval Center*, 2017, https://www.gallup.com/interactives/185273/presidential-job-approval-center.aspx, Accessed: 2 April 2017.

[Huf93] Darrell Huff, *How to Lie with Statistics*, W.W. Norton & Company, 1993.

[PB] Emily Pedersen and Alexander Barreira, *Former UC Berkeley political science professor Raymond Wolfinger dies at 83*, The Daily Californian, 11 February 2015, https://dailycal.org/2015/02/11/former-uc-berkeley-political-science-professor-raymond-wolfinger-dies-83/, Accessed: 21 February 2017.

[Sta02] Richard Stallman, *Free Software, Free Society: Selected Essays of Richard M. Stallman*, Free Software Foundation, 2002.

[TNA10] Mark Twain, Charles Neider, and Michael Anthony, *The Autobiography of Mark Twain*, Wiley Online Library, 2010.

[Tuf06] Edward Tufte, *The Cognitive Style of PowerPoint: Pitching Out Corrupts Within, ($2^{nd}$ ed.)*, Graphics Press, Cheshire, CT, USA, 2006.

[US 12] US National Library of Medicine, *Greek medicine*, 2012, https://www.nlm.nih.gov/hmd/greek/, Accessed 11-April-2017.

[Wik17a] Wikipedia, *Cantor's diagonal argument*, 2017, https://en.wikipedia.org/wiki/Cantor%27s_diagonal_argument, Accessed 5-March-2017.

[Wik17b] ———, *Unethical human experimentation*, 2017, https://en.wikipedia.org/wiki/Unethical_human_experimentation, Accessed 7-April-2017.

# Index