Statistics Minus the Math 2.0

Nathan Favero

2026-01-16

Table of contents

Preface

Note: This is an unstable, early version of a substantial rewrite (which may become version 2.0).

For young professionals entering the workforce in people-oriented fields, a basic familiarity with fundamental statistics is increasingly expected. This introduction to quantitative research methods for the social sciences teaches foundational skills in data description, quantitative reasoning, and statistical inference. The power of simple statistical techniques is demonstrated using examples from throughout the social sciences, with ample attention to practical uses of data like those students are likely to encounter on the job.

Several features make the approach here a bit unique:

- This text does not assume any prior training in statistics or a strong mathematical background. Explanations prioritize conceptual understanding over extended mathematical treatments. The text also emphasizes the importance of drawing on subject-matter knowledge to critically evaluate assumptions underlying measurement, visualization, and inference.
- Contemporary approaches to quantitative social science are prioritized, resulting in greater attention to causality, regression modeling, and confidence intervals.
- The ordering of topics is designed to equip students to start working with data and reading quantitative social science literature as quickly as possible. The first three chapters not only teach data description and visualization but also provide basic guidelines for reading a regression results table. This also allows students to quickly begin work on term papers that may utilize regression. Later chapters add coverage of more abstract topics like probabilistic modeling, statistical inference, and regression assumptions.
- Chapters are relatively short (usually less than 5000 words), allowing for quick reading and making clear the most essential material for students to master on each topic. Chapter appendices sometimes provide additional content that may be important for more technically-oriented versions of an intro course.

This comprehensive introduction to quantitative analysis will equip students to independently evaluate and produce simple statistical analyses. Students will realize through the straightforward applications of statistical concepts to relevant and clear examples how the reasoning skills they've utilized their entire lives can help them effectively describe, contextualize, and interpret patterns in data.

Substantial portions of this book are adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (https://onlinestatbook.com Project Leader: David M. Lane, Rice University). A huge thanks to David Lane and his colleagues at Rice University for their creation of this wonderful resource. I use footnotes throughout to indicate precisely where the various sections of each chapter came from. Currently, Chapter 2, Chapter 3, Chapter 7, Chapter 8, and Chapter 9 are mostly derived from this public domain resource.

Acknowledgements

I would like to thank Natasha Kallish for help transforming an earlier version of this text from Word to a Quarto document.

Part I

Describing Data

1 Getting Started with Data

Data is everywhere. Yet many of us find numbers intimidating. It can be tempting to assume that someone sophisticated enough to fluently cite numerical information must know what they're talking about. At the same time, there is a kind of backlash to using numbers to describe the world. The title of the popular book *How to Lie with Statistics* captures a cynicism many people feel: since statistics can easily trick you, you shouldn't believe them at all. Such extreme reactions reflect people's lack of self-confidence in evaluating quantitative information. Because we don't trust ourselves, we reflexively either dismiss or accept numerical claims—perhaps depending on whether the claim is one we already agree with or the source is one we trust.[1]

Fortunately, statistics is a topic that most anyone can learn. Even if you think you are not a "numbers person," I am confident you can learn fundamentals that will allow you to critically assess statistical claims. I say this as someone who has spent much of my academic life surrounded by people who lacked confidence in their numerical abilities and yet decided to study public affairs—sometimes not realizing the extent to which they were signing up to work with numbers along the way. Many peers during my time studying for a PhD thought their abilities in math were not strong, but by the time they completed their degrees, they were all exceptionally competent in statistics. It is true that some people learn numerical material quicker than others. For most people, lots of repetition is required before many statistical concepts are grasped. But I have yet to encounter a student who is unable to learn statistics. If you set your mind to it, you will learn it.

In my opinion, the most important skill for dealing with statistics is critical thinking. While statistics is certainly quantitative, doing statistical analysis does not require you to perform complex math because software will handle the number crunching. Thus, being comfortable with statistics is mostly about developing familiarity with statistical concepts, mastering a few technical details, and thinking carefully about how to best apply statistical tools to real-world data. "Subject-matter expertise," by which I mean familiarity with the policy or program area described by the data, is almost always required for good interpretation of statistical results. Sometimes, a bit of math will help with explaining how a statistic works or gaming out the implications of a statistical result. But even then, most of the math we need for an introductory course is relatively simple arithmetic.

[1]This chapter was written by Nathan Favero.

1.1 Three Questions to Always Ask about Data

To help focus your attention on thinking critically, I encourage you to ask the following three questions whenever you encounter statistical information:

1. What is being measured?
2. Who (or what) is in the dataset?
3. How big are the differences?

Example: Long-Term Trends in U.S. Education

Many have argued that despite huge investments in public schooling, the U.S. has little to show in terms of learning gains.[2] In Figure 1.1, the solid line indicates a clear upward trend in spending on education over the decades, although there is a slight dip at the very end of the series (following the 2008 financial crisis).[3] The trajectory for math scores (the dashed line) is much less clear: an initial drop is followed by steep gains for a little while. Looking to the beginning and end points, we see the average score in 2012 is a bit higher than in 1973 (306 vs. 304), but this gain is tiny compared to the fluctuations seen across years.[4]

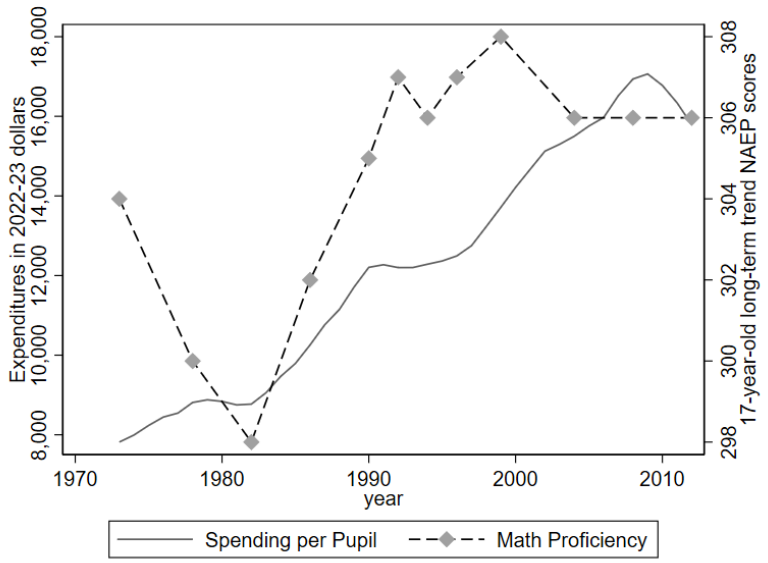

Figure 1.1: Trends in U.S. education spending and average math proficiency (17-year-olds)

Notice that there are two y-axes on this graph: one on the left edge showing the units for expenditures and one on the right edge displaying units that are described as NAEP

scores. That is because the trend in Math proficiency (measured using NAEP scores) is superimposed over the trend in spending levels, despite the fact that these two variables are measured in entirely different units. In my experience, the use of two y-axes often leads to confusion, and it is usually better to "keep it simple" by separating the data into two separate graphs.

In analyzing this example, I will mostly ignore the spending trend in favor of evaluating the math trend, for sake of brevity. Still, I will note that increased spending is largely driven by (1) increasing labor costs across the (high-skilled) economy and (2) greater investments in special education. The latter investments likely improve equity and fairness in society but may do little to improve test scores.

Regarding the math trend, let us consider our 3 Questions to Always Ask about Data:

1. **What is being measured?**

   If you were encountering this data for the first time on your own, I would recommend doing a quick online search for NAEP scores, perhaps focusing on descriptions from familiar sources like Wikipedia or government websites. You would find that NAEP scores come from standardized tests conducted by the Department of Education. If you wanted to go even deeper, you might do a Google Scholar search to see how researchers seem to be discussing/using NAEP scores. They are generally considered to be high-quality measures of student learning. You might also discover that there are actually two different kinds of NAEP scores: the long-term trend scores, shown here, as well as another set of scores based on a test that has only been administered since the 1990s.

2. **Who (or what) is in the dataset?**

   Typically, we ask this question because we wonder who (or what) might be *missing* from the data. Perhaps the first thing you notice is that the data stops in 2012. Why aren't students after 2012 included? It turns out that 2012 is the most recent year available for this trendline because the Department of Education has canceled data collection multiple times for the long-term trend NAEP among 17-year-olds, citing limited funding and COVID-19 pandemic disruptions. This reflects a frequent but unfortunate problem we face when analyzing data in the real world: data collectors often discontinue their work and leave us with an incomplete picture. Nonetheless, we can still use this data to say something meaningful about learning trends from prior decades.

   If you did some basic reading about the NAEP as suggested for the prior question, you would probably find that NAEP scores are available for various ages. Specifically, the long-term trend scores are available for kids aged 9, 13, and 17. Yet the graph only depicts the data for 17-year olds. On the one hand, it might seem that the oldest age is the most important since it indicates how students are doing near the end of their secondary education. On the other hand, it might be useful to

see what the other trends look like. A quick online search for "naep scores over time" will likely return a government website showing NAEP scores for younger students, such as those shown below.[5] Among 9 and 13-year-olds, there is clear improvement in math between the '70s and 2012. Then more recently, there is a large drop associated with the disruptions of COVID-19.

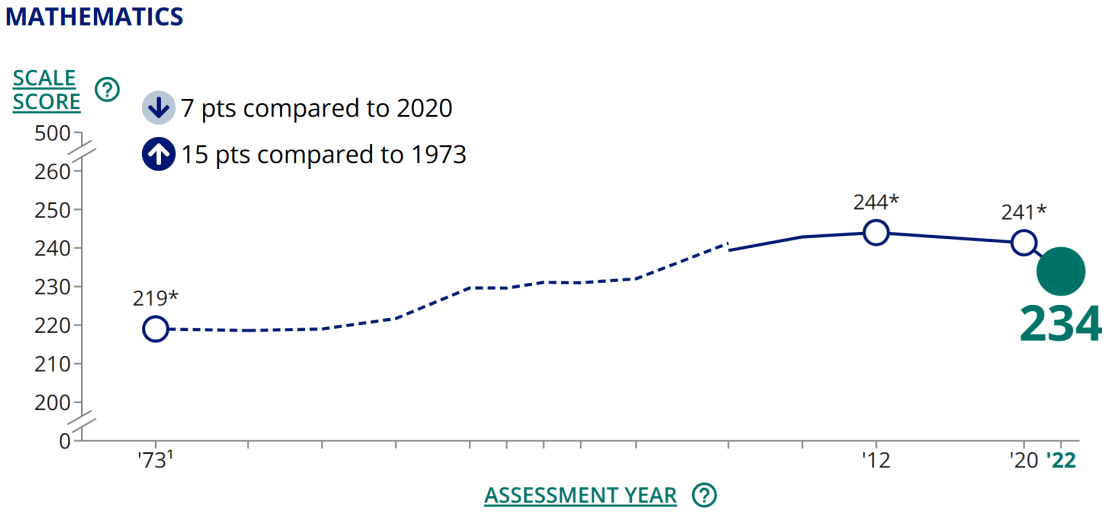

Figure 1.2: Trend in 9-year-old math proficiency

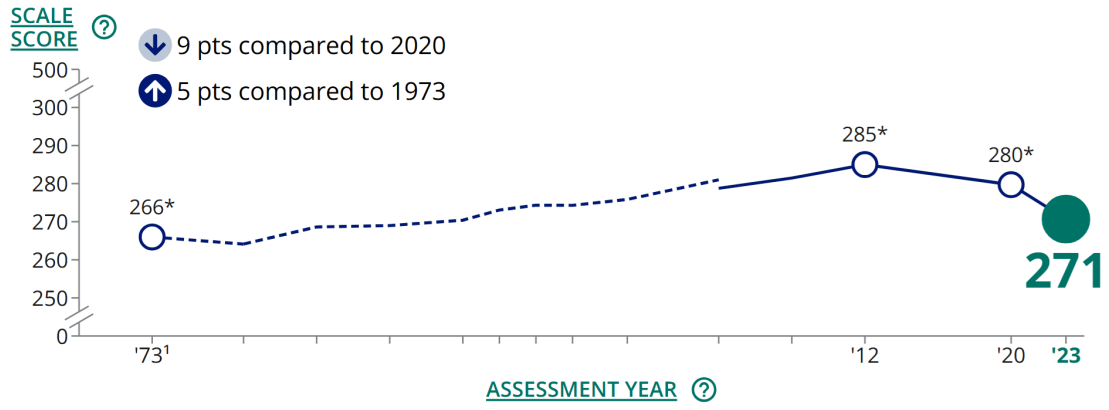

Figure 1.3: Trend in 13-year-old math proficiency

Given these results, you might conclude that there was some possible cherry picking in presenting the initial graph, since the trend among 17-year-olds shows much

less improvement from the 1970s to the early 2000s than the trends for younger students. That is an important conclusion to reach and helps put the original graph in context.

If you have some expertise in U.S. education, you can go even further with your evaluation of this data. One potentially important difference between 13 and 17-year-olds in the U.S. is that school attendance is typically compulsory at age 13 but not at 17. An online search (e.g., for "change in school dropout rate over time") will likely reveal government reports showing that the dropout rate has generally declined over previous decades. Thus, one potential explanation for the mostly-flat trendline in 17-year-old NAEP scores is that in prior decades the 17-year-old scores were artificially inflated due to lower-performing students dropping out of high school and thus not taking the exams. In more recent years, more of these lower-performing students remain in school (a positive policy outcome), and even if they are learning more than in the past, their continued presence in schools brings the NAEP score average down, all else equal. To be precise, the overall picture is muddy because of limited data, and it is not clear exactly how much the improving dropout rate has distorted the 17-year-old trendline. But there is probably at least some distortion, and it is thus difficult to make definitive statements about the trends in learning among 17-year-olds.[6]

One final note about what is (not) included in the data we've been discussing: in addition to math, there are reading scores available for NAEP. While I focus on a single subject here for simplicity, an important caveat to our story about 9 and 13-year-olds is that their reading scores improved less than their math scores between the 1970s and 2012.

3. **How big are the differences?**

   The 9-year-olds' 2012 scores are 25 points higher than in 1973 (244-219=25). Gains for 13-year-olds over the same period are a bit smaller: 19 points of improvement (285-266=19). How big is a 25-point increase? It's a bit hard to say because NAEP scores are measured in units that are not familiar to most of us. When working with unfamiliar units of measurement, sometimes the best we can do is to make relative comparisons. For example, we see that in 1973, 13-year-olds scored 47 points higher than 9-year-olds (266-219=47). Thus, a 25-point gain is over half the difference between 9 and 13-year-olds in terms of math proficiency on the 1973 exam. We could further contextualize these numbers by making the simplistic assumption that 11-year-olds—who don't take the NAEP—land exactly halfway between 9 and 13-year-olds. If so, 1973's 11-year-olds should have scored a 242.5. Thus, over the course of four decades, 9-year-olds seem to have improved to roughly the level of 11-year-olds in 1973. This is a substantial improvement!

   The term "back-of-the-envelope" calculations refers to the use of simple math to make approximations based on convenient assumptions, as we have just done. Such

calculations yield only rough approximations, but they can sometimes help us better interpret numerical results.

1.2 How Datasets are Structured

We often display data in the form of a spreadsheet. Table 1.1 shows part of a dataset describing countries. Each row represents one country, and we call the rows **observations**. Each column represents one characteristic of the countries. We call the columns **variables**. Variables allow us to describe whatever it is we care about—concepts like size of the public sector, the share of workers who are female in the public (or private) sector, and whether a country is a former British colony. We can easily imagine adding other variables like country wealth, average education level, or government spending on social programs. We call the columns variables because they display values of a characteristic that varies depending on which observation we are talking about: in Guatemala, the public sector size is 0.064, while in Hungary it is 0.261. What do these numbers mean? Here, public sector size is measured as the proportion of employment that is in the public sector. We can convert a proportion to a percentage by shifting the decimal point two places to the right. Thus, in Guatemala, 6.4% of jobs are found in the public sector, whereas in Hungary the figure is 26.1%. The public sector is much larger in Hungary than in any other country shown in this table.

Table 1.1: Preview of a dataset of countries and their characteristics[7]

country_name	publ_sector_size	females_publ	females_priv	former_british_colony
Guatemala	0.0642733	0.437011	0.2751051	0
Guinea-Bissau	0.0535029	0.2742133	0.3592181	0
Honduras	0.0602268	0.5613141	0.3063367	0
Hungary	0.2614872	0.694158	0.3943521	0
India	0.0851988	0.3021879	0.1555202	1
Indonesia	0.097229	0.4430795	0.3249187	0

[6]See https://www.hamiltonproject.org/wp-content/uploads/2023/01/092011_education_greenstone_looney_shevlin.pdf.

[6]Expenditure per pupil in fall enrollment from Table 236.55 of the *2023 Digest of Education Statistics,* National Center for Education Statistics. https://nces.ed.gov/programs/digest/d23/tables/dt23_236.55.asp

[6]National Center for Education Statistics (2013). *The Nation's Report Card: Trends in Academic Progress 2012* (NCES 2013–456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

[6]U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1971–2023 Long-Term Trend Reading and Mathematics Assessments. https://www.nationsreportcard.gov/ltt

[6]See https://www.chalkbeat.org/2022/3/31/23005371/high-school-test-scores-underestimate-naep-dropout-nces/ and https://www.edweek.org/teaching-learning/are-rising-grad-rates-pulling-down-naep-scores/2016/05.

Table 1.2 previews a dataset where each observation (row) is a different individual who responded to a survey. We use the term **unit of analysis** to describe what constitutes one observation in a dataset: in the prior table the unit of analysis was the country, whereas here it is the individual. Other common units of analysis include the organization, the work unit, and the subnational unit (e.g., city or region).

Table 1.2: Preview of data from a 2024 survey of U.S. government employees[8]

random_id	race	hispanic	sex	q1	q2
194868625278	White	No	Male	Neither Agree nor Disagree	Neither Agree nor Disagree
152966380283	White	No	Male	Disagree	Strongly Disagree
146904434378	Other	No	Female	Agree	Agree
161966059804	White	Yes		Strongly Agree	Strongly Agree
133516090099	Black or African American	No	Male	Agree	Agree

1.2.1 Types of variables

We also distinguish among different kinds of variables, which will often require different types of analysis.

Quantitative variables are expressed in numbers. The public sector size variable we examined above is quantitative.

For **qualitative variables**, it typically makes more sense to use text than numbers to describe the values. For example, race is a qualitative variable, as shown in the example survey data above. While you may sometimes encounter datasets that record the values of a qualitative variable using numbers (for ease of processing), the assignment of numbers to values will be arbitrary. This is because qualitative variables take on values that are unordered categories (cannot be arranged from least to most). Hispanic and sex are also qualitative variables in the example survey data.

A qualitative variable that takes on only two values is called a **binary variable**. The variables hispanic and sex are binary. For the sex variable, there is also one blank cell, indicating a missing value. Binary variables often appear in spreadsheets with values shown as 1s and 0s.

[7]Data sources are the Worldwide Bureaucracy Indicators dataset (CC-BY 4.0) and the Colonial Dates Dataset.
[8]2024 U.S. Federal Employee Viewpoint Survey

In some cases, a 1 indicates "yes" while a 0 indicates "no." For example, in the country data shown above, the variable "former_british_colony" is coded as a 1 if "yes, this country is a former British colony" and 0 if it is not.

There is also a third type of variable called an **ordinal variable**, where values can be arranged in order but cannot be easily quantified. This type exists in a somewhat grey zone between quantitative and qualitative. Many surveys utilize Likert scales, which allow respondents to choose from a range of options along a continuum (e.g., strongly disagree, disagree, agree, strongly agree). Variables q1 and q2 in the survey data example use this kind of scale. For q1, people are responding to the statement "I am given a real opportunity to improve my skills in my organization." Q2 asks them to indicate whether they agree that "I feel encouraged to come up with new and better ways of doing things." Likert scales result in ordinal variables: while we can arrange response options from most agreement to least agreement, it is not clear how to assign precise numbers because we don't know if the distances between response options are equal. If "strongly disagree" is a 1 and "disagree" is a 2, should "neither agree nor disagree" be a 3? Or a 4? Maybe 3.5? It is difficult to say what numbering scheme would most accurately represent respondents' attitudes because this is an ordinal variable.

1.2.2 Types of datasets

So far, the datasets we have seen are what we call **cross sections**. Each observation is a different unit. There are also **time series** datasets, where each observation is a different point in time. We saw an example of time series data on U.S. education being depicted graphically at the beginning of the chapter. Table 1.3 shows some of this same data displayed as a spreadsheet. With time series data, the unit of analysis could be the year (as in this example), the quarter, the month, the week, the day, etc.

Table 1.3: Time series data (U.S. education spending)

year	spending_per_pupil
1973	7817
1974	7994
1975	8235
1976	8445

You may also sometimes encounter more complicated data structures that combine the attributes of cross-sectional and time series data. A **panel** dataset tracks multiple units over multiple time periods. For example, a dataset might track several countries over several years (each row describing one country in one year). A **repeated cross section** is similar, except that different units are observed in each time period. A survey that is conducted annually but where the respondents are different each year is a repeated cross section.

1.2.3 Varying terminology

Unfortunately, there are many cases where statistical terminology is inconsistent from one source to the next. Since you will probably encounter research reports or articles that use different terminology than I use here, it is important to be familiar with these alternative terms:

- Qualitative variables can also be called **categorical variables** or **nominal variables**.
- Binary variables are often called **dummy variables**.
- Time series data is sometimes called **longitudinal data**.

Sources also disagree on whether ordinal variables should be considered a subcategory of quantitative or qualitative variables. For this text, I consider ordinal variables to be a distinct third category, but different authors classify them differently.

1.3 Visualization Basics

Creating simple graphs is often a great first step for getting familiar with a dataset. In this chapter, we will focus mainly on graphing just one variable at a time.

1.3.1 Qualitative and ordinal variables

For qualitative and ordinal variables, we often use bar charts to depict the frequencies of different values. (We can also sometimes use bar charts to depict quantitative variables if all values are whole numbers, as we will see next chapter in Figure 2.5.)

One can quickly see from Figure 1.4 that the most common value for the race variable is White, meaning that most U.S. government employees responding to the survey are White. We also see that the values of Asian and Other are fairly rare (each occurring in fewer than 10% of observations). Black or African American occurs a bit more frequently (around 15% of observations).

Another type of graph you can create with qualitative data is a pie chart, using the frequency of each value to create the size of the pie slice. Pie charts visually suggest a zero-sum approach to thinking about the sizes of the different categories: you can't make one slice bigger without making at least one other slice smaller. This makes pie charts particularly effective for depicting something like a budget allocation where a scarce resource is being distributed across categories. However, a zero-sum allocation is not always what we wish to emphasize when summarizing a qualitative variable. Pie charts can also be hard to read with certain configurations of data (i.e., when there are too many categories or the slices get too small). Thus, it should come as

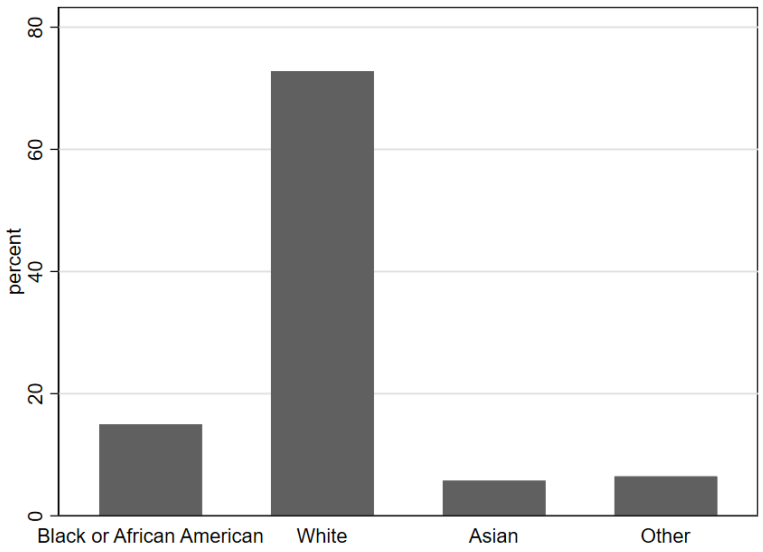

Figure 1.4: Bar chart of race variable from 2024 survey of U.S. government employees

no surprise that research scientists tend to rely on bar charts more than pie charts for depicting qualitative variables.

Note that bar charts and pie charts can be used to represent things other than the frequency with which different values of a variable occur, even though that is the main way we use them while studying statistics. When you encounter a graph, it is always important to carefully read the titles and labels to make sure you understand exactly what is depicted. A bar graph might, for example, indicate the size of the change (positive or negative) in the budget from 2024 to 2025 for various categories of spending. Or as noted above, you might see the levels for budget categories themselves depicted in a pie chart—highlighting how some categories make up a much larger share of the budget than others. In such cases, the charts would not be telling you how many observations (rows of data) record different values of a variable, as in the example graphs shown here.

1.3.2 Quantitative variables

Perhaps the most popular type of graph for visualizing a single quantitative variable is a histogram. An example is shown below, using the publ_sector_size variable we discussed above. In a histogram, each bar represents a range of values, known as a "bin." The x-axis (along the bottom of the graph) shows us the range of values for the publ_sector_size variable being described by each bar. The height of each bar represents how many data points fall within the corresponding bin. In this particular histogram, the y-axis (along the left side of the graph) shows how different heights indicate different numbers of observations. For example,

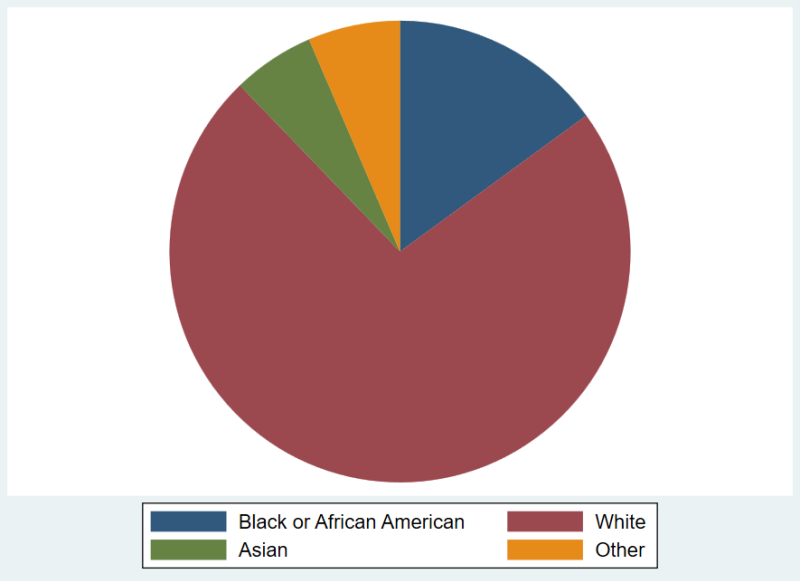

Figure 1.5: Pie chart of race variable from 2024 survey of U.S. government employees

the first bar on the left has a height indicating seven observations. The approximate range shown on the x-axis for this bar is .02 to .07. Therefore, seven countries have a public sector size between approximately .02 and .07, meaning that 2-7% of employment in those countries is within the public sector. The right-most bar indicates that 5 countries—those with the largest public sectors—have approximately 30-35% of all workers employed by the public sector. Most countries in this dataset lie between the two extremes depicted by the left-most and right-most bars. There are more observations in the middle three bins (indicated by the taller bars) than there are in the two left-most or the two right-most bins.

Now, let's examine a histogram for the variable females_publ, which indicates the proportion of public sector employees who are female. In this graph, we can notice a certain asymmetry: the bars on the right half of the graph tend to be taller than those on the left. The short bars indicate that relatively few countries have proportions in the approximate range of .1 to .45; in other words, for a small number of countries, we see that 10-45% of the public workforce is female. Most countries have a public workforce that is more like 45-75% female (the tall bars on the right side of the graph). Thus, it seems that in a majority of the 44 countries included in this dataset, the public sector employs more females than males.

There are several alternatives to histograms. You can see in Figure 1.8 several different types of graphs being used to depict the same data for the females_publ variable. Each one indicates the relative density of observations across the range of possible values. Much like a histogram, a kernel density (k-density) plot uses height to indicate the frequency with which values occur, but rather than dividing values up into discrete ranges (bins), an algorithm is used to create a smooth, continuous line that is drawn across the values. A violin plot (which can be oriented

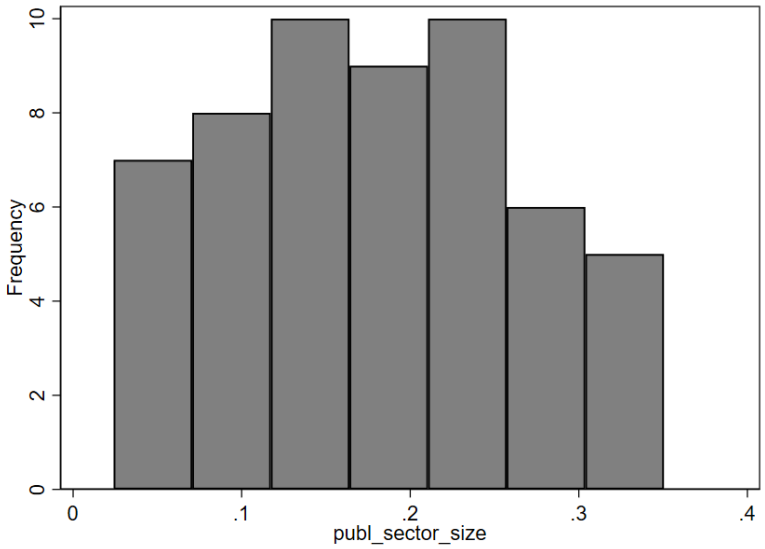

Figure 1.6: Histogram of public_sector_size variable from a dataset of countries and their characteristics

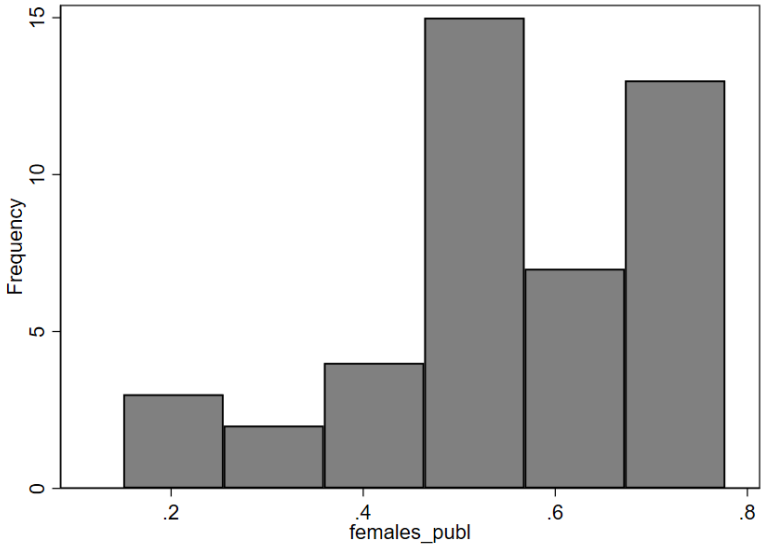

Figure 1.7: Histogram of females_publ variable from a dataset of countries and their characteristics

horizontally or vertically) is somewhat similar but uses thickness rather than height to indicate density of observations across the range of values. A strip plot uses one dot per observation but usually adds some random noise called "jitter" to the data. Without this jitter, dots often stack on top of one another such that the actual density of observations is obscured. Finally, a box plot uses a box to indicate the range of values containing the middle 50% of observations, with other lines indicating other important quantities like the full range of values. Given the importance and complexity of box plots, we will examine them in more detail in the next chapter.

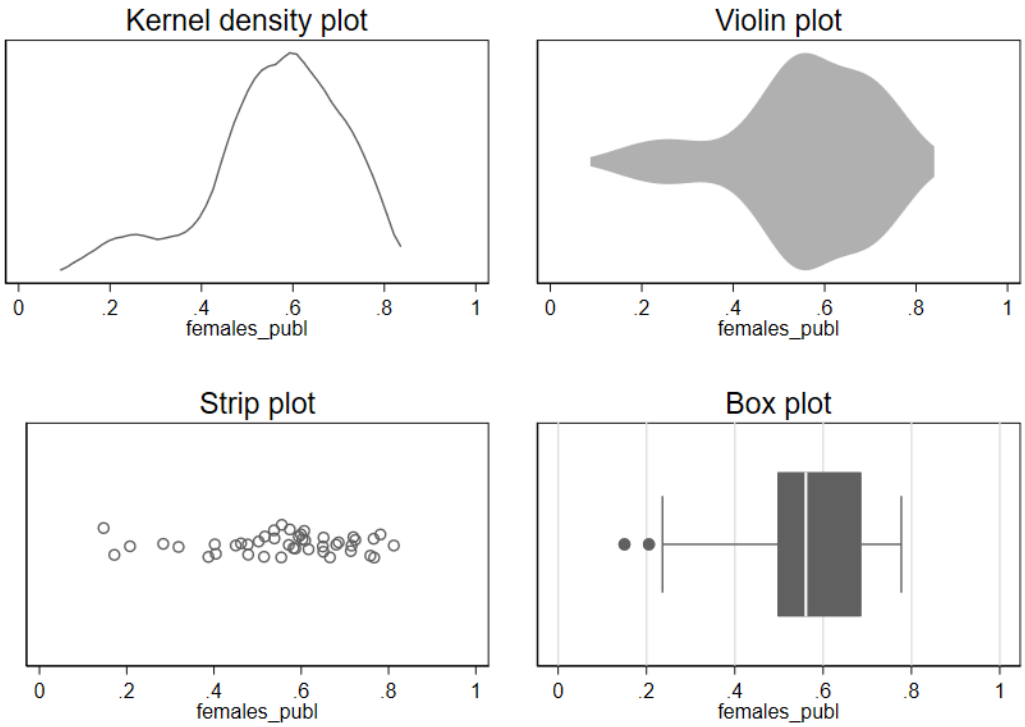

Figure 1.8: Various plots of public_sector_size variable from a dataset of countries and their characteristics

Regardless of which type of plot we use, we see the same asymmetry apparent in the original histogram for females_publ: most of the data lies in the range of .4-.8, with a few observations also occurring in the range of .1-.4.

1.3.3 Best practices for simple graphs

As useful as graphs are, it is also easy to go wrong with data visualization. Here are three simple guidelines to keep in mind when getting started with graphs.

First, the simple graphs we examined for quantative variables are typically a good starting point for learning about a variable, but there may be important details that are not apparent from the first graph we look at. For example, while histograms usually provide a nice overview of a variable, the overall shape depicted by the bars in a histogram can sometimes change in surprising ways if we alter the boundaries of the bins.

Example: Estimates of Small Donations

A survey of nonprofits asked each respondent to estimate the percentage of individual donations to their organization in 2019 that were smaller than \$250.[9] Look at how different the right-most bars of the histogram look, depending on whether we use 10 bins or 11 bins (a setting we can change in the software creating the histogram):

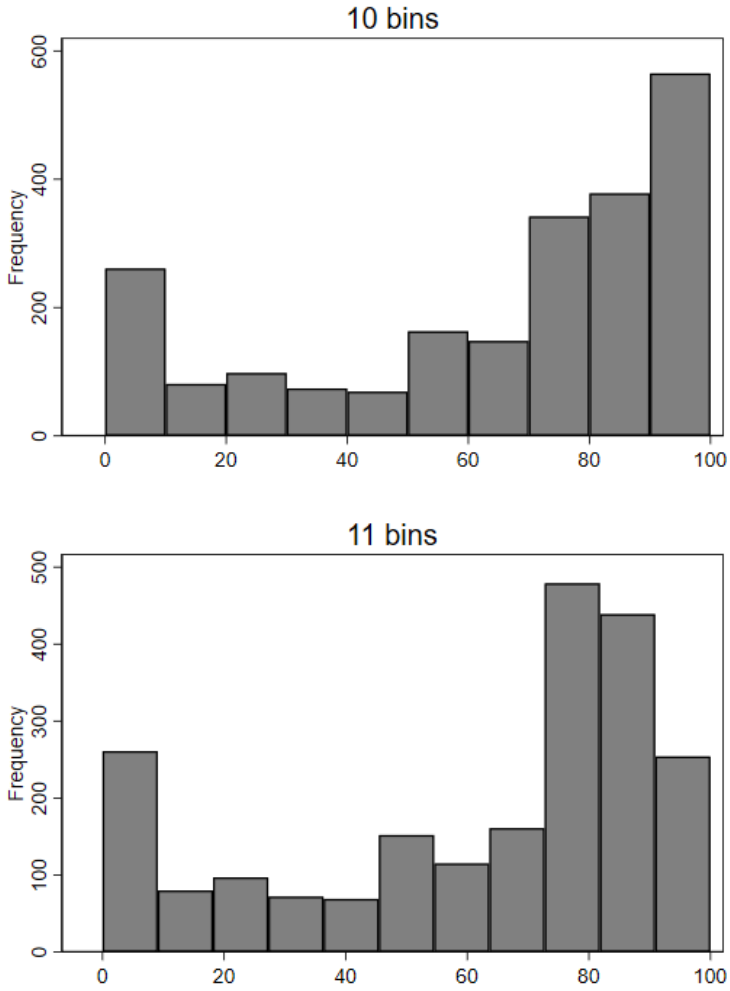

Figure 1.9: Histograms with different bin settings (depicting donations estimate variable)

Why the dramatic change? If we plot bins with a width of 1, we get a better sense of what is unusual about the underlying data:

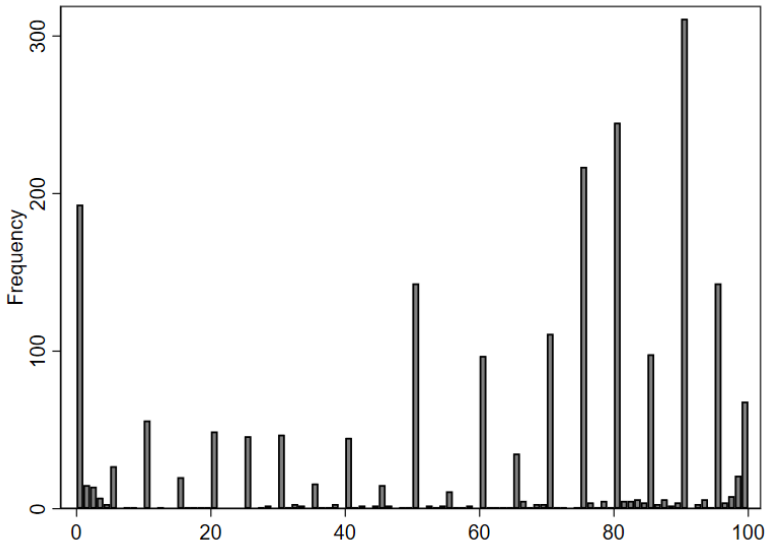

Figure 1.10: Histogram with a bin width of 1 (depicting donations estimate variable)

From this more fine-grained depiction of the data, we see that respondents appear to typically answer with numbers divisible by 5. Remember, survey respondents were asked to provide an estimate—not look up the exact number from their records. People are more likely to offer estimates that are "round" numbers like 90 or 95, as opposed to something like 92.
In the initial histogram with 10 bins, the right-most bin included the value 90—the most popular answer. When we switched to using 11 bins, the right-most bin was redrawn to exclude 90, so the number of observations in the bin (represented by its height) dropped dramatically. Because of the spikes in frequency at round numbers (divisible by 10 or 5), changes in bin settings can cause relatively large changes to the visual pattern of the histogram.

Second, be careful about using line graphs. When working with time series data, we often depict quantitative variables using a line graph. As we already saw in our example on trends in U.S. education, line graphs allow one to visually observe trends over time. While line graphs are appropriate for depicting time series data, their use in other contexts is often misleading, since the lines suggest connections between points that may not be connected at all in reality.

Finally, it is no accident that the graphs we have just reviewed are visually quite simple—maybe even boring. As I alluded to near the beginning of the chapter, "keep it simple" is a great mantra to remember for data visualization (and much of data analysis). Some software programs

[9]Year 1 of the Nonprofit Trends Longitudinal Survey Public Use Data Files. https://doi.org/10.7910/DVN/T4OT1J

will point you toward features like 3-dimensional graphs or using images instead of bars to depict information. While such graphs may look impressive on the surface, the flourishes usually distract from the main point of the graph (and can sometimes even actively mislead the reader).[10] Focus on simplicity and clarity in data visualization, not trying to stand out.

1.4 Critically Evaluating Graphs

Graphs are often poorly constructed in ways that can mislead, so it's always important to think critically and exercise caution when interpreting data visualizations.

Let's look at some examples and see if any of the Three Questions to Always Ask about Data can help us identify misleading graphs.

Example: Differences in employee attitudes by sex

Q2 from the survey of federal employees asks whether there is a workplace environment supportive of innovation. As already noted, this variable is ordinal, meaning it is not obvious what numbers would be most appropriate to attach to the variable values. Nonetheless, we adopt here the common practice of starting from 1 and counting up by whole numbers for the different response option:

- Strong Disagree = 1
- Disagree = 2
- Neither Agree nor Disagree = 3
- Agree = 4
- Strongly Agree = 5

We're not sure if these numbers are really the ideal ones to assign, but they are a reasonable approximation of how attitudes might be mapped to numbers. And by assigning precise numbers to the response options, we enable treating this variable as quantitative. This is beneficial because quantitative variables are often easier to work with than qualitative variables. For example, we can calculate the average of a quantitative variable. The overall average of the q2 variable is 3.79, which is a bit less than the value we assigned to Agree. We can also compute averages for different subsets of survey respondents. For example, we can compute separate averages by sex (separating out males and females) and display these averages in a simple graph. Here is one possibility for what that graph could look like:

[10] See, for example, violations of the "principle of proportional ink": https://callingbullshit.org/tools/tools_proportional_ink.html

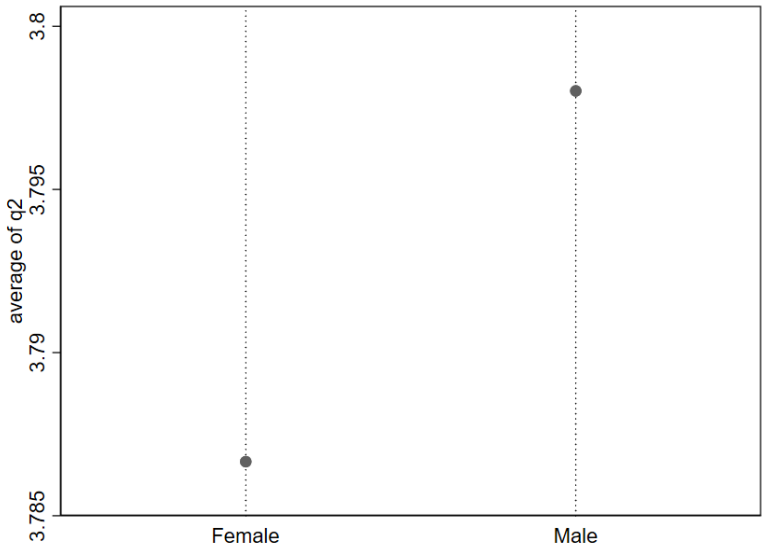

Figure 1.11: A tiny difference that looks large: Comparing average male and female responses to survey item "I feel encouraged to come up with new and better ways of doing things."

A quick glance at this graph suggests that males feel much more supported in pursuing innovation than than females. But our third Question to Always Ask about Data indicates we should consider "How big is the difference?" Look closely at the y-axis here. The average response for females appears to be around 3.787, while the average response for males is approximately 3.798. That is a tiny difference—just .011 on a 1-5 scale. In reality, male and female respondents report very similar average levels of support for pursing innovation. It is just that the figure is depicting a very narrow portion of the range for this variable, which makes a mole hill look like a mountain.
Let's look at what happens when we redraw the y-axis:

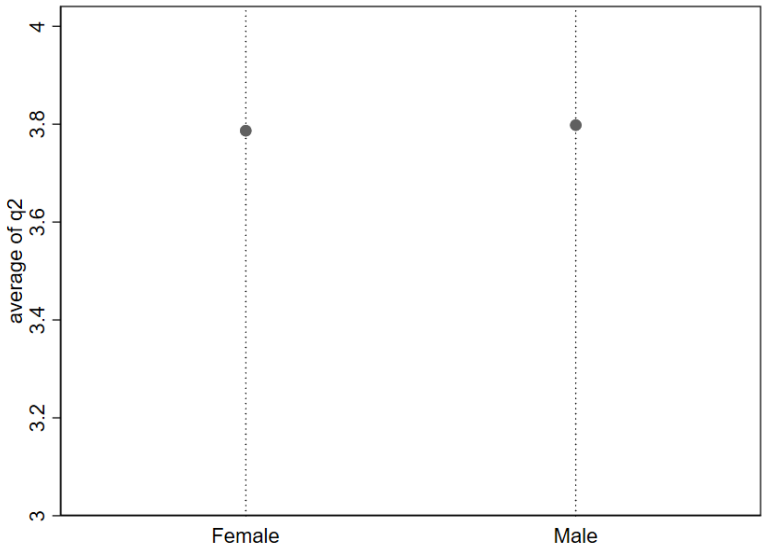

Figure 1.12: A tiny difference that looks tiny: Comparing average male and female responses to survey item "I feel encouraged to come up with new and better ways of doing things."

Now, the difference in averages between females and males is difficult to visually detect because the dot for Male is barely higher up than the dot for Female. Notice the values on the y-axis: the range begins at 3 (Neither Agree nor Disagree) and ends at 4 (Agree). We have "zoomed out" on the difference we saw in Figure 1.11.

Generally speaking, any difference can be made to visually look very large or very small depending on how the axes are drawn. It is all about how "zoomed in" or "zoomed out" you are relative to the range of possible values for the variable.

It is sometimes tempting to create simplistic rules to try to avoid misleading graphs like the one from the prior example. Such rules are generally a bad substitute for carefully thinking through how to best describe the data in front of us. As an example, it is sometimes advised to always including 0 in the y axis in order to avoid "zooming in" on the y axis so much that we make small differences appear larger than they really are. Let's consider this rule in the context of some data on trends in childhood vaccination.

Example: Childhood Vaccination Rates

Public health officials advise that at least 95% of the community should be vaccinated against measles in order to maintain herd immunity.[11] Figure 1.13 shows estimates for the rate of measles-mumps-rubella (MMR) vaccination in the U.S. among children starting

primary school (solid line).[12] The y-axis is drawn to start at 0, and the target rate of 95% is depicted as a dashed line. The solid line looks almost flat and is always close to the dashed line. It is hard to say much more based on this graph, except that the actual vaccination rate appears to fall a bit shy of the 95% target in recent years.

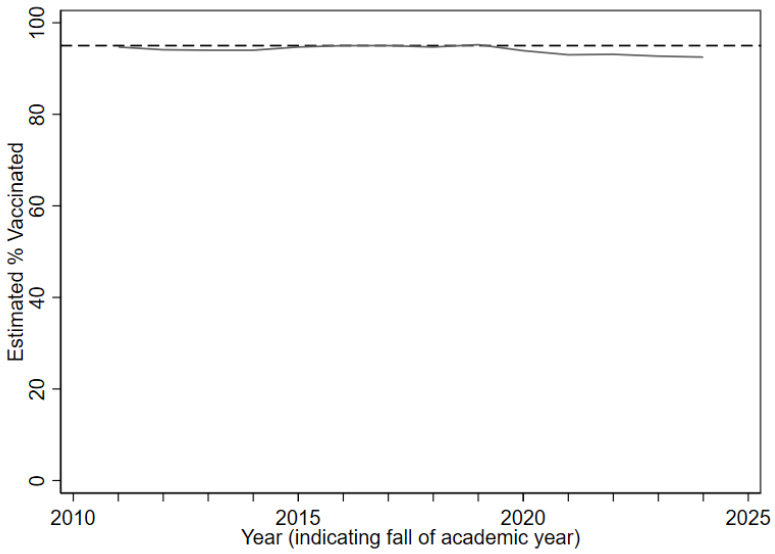

Figure 1.13: A depiction of the trend in MMR vaccination among children starting primary school (kindergarten), compared to 95% target

Now look at the following alternative. While the exact same data is being depicted in Figure 1.14, the redrawn y axis makes the solid line look like it is moving a lot (suggesting big year-to-year differences). Prior to 2020, rates were pretty consistently within the 94-95% range. But more recent years show rates two or three percentage points below the 95% target. Given the potential public health implications of falling just one or two percentage points below the target of 95% vaccine coverage, it does not seem like a smart choice to include 0 on the y axis for this data (as in Figure 1.13), since doing so makes it quite difficult to precisely discern changes of one or two percentage points.

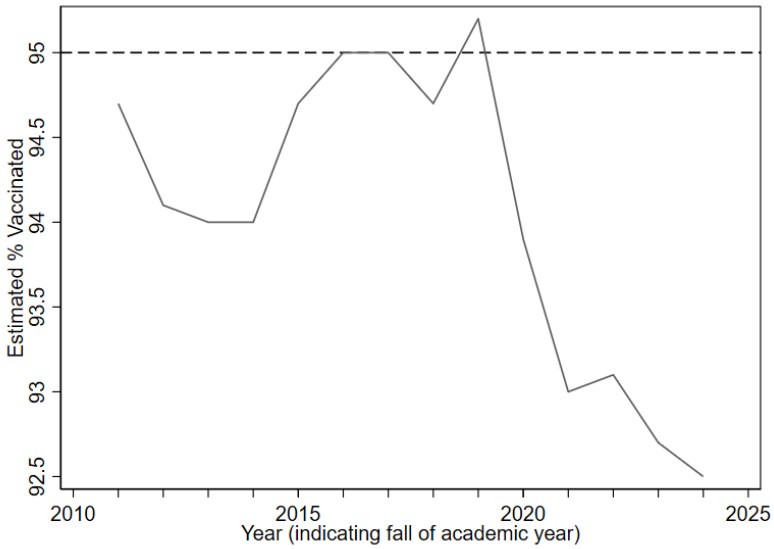

Figure 1.14: Another depiction of the trend in MMR vaccination among children starting primary school (kindergarten), compared to 95% target

For good visualization, we want differences that are meaningful in reality to be visually noticeable in our graphs. How do we define a "meaningful" difference? There is no universal rule. Typically, subject-matter expertise and careful judgment will be our best guides.

Whenever you see a difference in a graph that "looks big" or "looks small," take a moment to pause. Look carefully at the numbers on the axes, and think about how to evaluate the third Question to Always Ask about Data: "Based on what I know about this topic and how the variables are measured, *how big is this difference*? Does my understanding of the numbers match what my eyes see?"

[12]Pandey, A., & Galvani, A. P. (2023). Exacerbation of measles mortality by vaccine hesitancy worldwide. *The Lancet Global Health*, *11*(4), e478-e479. https://doi.org/10.1016/S2214-109X(23)00063-3

[12]Data from the U.S. Centers for Disease Control and Prevention.

2 Describing One Variable at a Time

2.1 Distributions[1]

2.1.1 Distributions of Qualitative or Discrete Variables

I recently purchased a bag of Plain M&M's. The M&M's were in six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table 2.1.

Table 2.1: Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a **frequency distribution**. Often a frequency distribution is shown graphically, as we saw in the prior chapter.

2.1.2 Continuous Variables

The variable "color of M&M" is a qualitative variable, and its distribution is called discrete because there are a finite number of values the variable can take on. Let us now extend the concept of a distribution to quantitative variables measured to many decimal places.

The data shown in Table 2.2 are the times it took David Lane (the author of much of the material appearing in this book) to move the cursor over a small target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous

[1]This section is adapted from David M. Lane and Heidi Ziemer. "Distributions." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/introduction/distributions.html

variable. With time measured accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in psychology. As you can see in Table 2.2, measuring David Lane's responses this way produced times no two of which were the same. As a result, a frequency distribution would be uninformative: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 2.2: Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

The solution to this problem is to create a grouped frequency distribution, as we saw when creating bins for histograms in Section 1.3.2. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 2.3 shows a grouped frequency distribution for these 20 times.

Table 2.3: Grouped frequency distribution

Range	**Frequency**
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

2.1.3 Shapes of Distributions

As we've already seen when graphing different data, distributions have different shapes. Some distributions are symmetric; if you folded them in the middle, the two sides would match perfectly. Figure 2.1 shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in

the negative direction. A distribution with the longer tail extending in the positive direction is said to have a **positive skew**. It is also described as "skewed to the right."

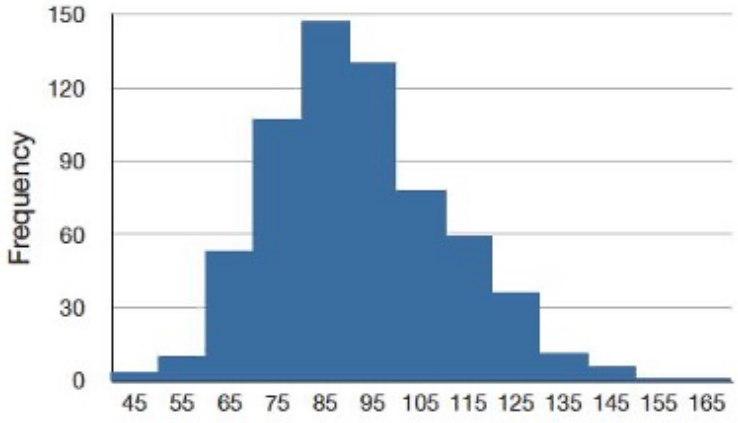

Figure 2.1: A distribution with a positive skew.

Although less common, some distributions have a **negative skew**. Figure 2.2 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the left, this distribution is skewed to the left.

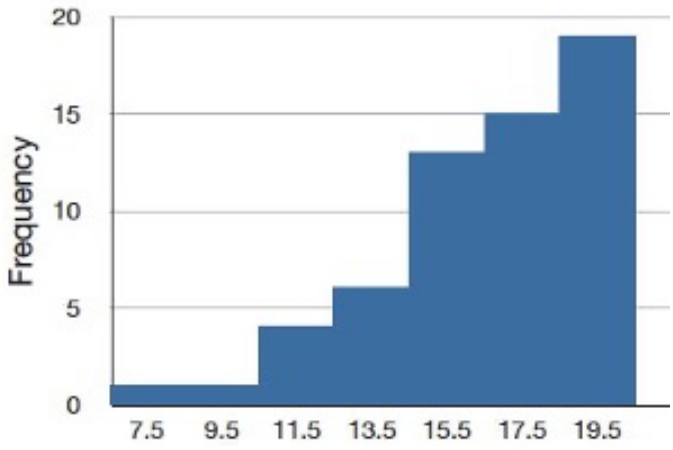

Figure 2.2: A distribution with negative skew.

The distributions shown so far all have one distinct high point or peak. The distribution in Figure 2.3 has two distinct peaks. A distribution with two peaks is called a **bimodal distribution**.

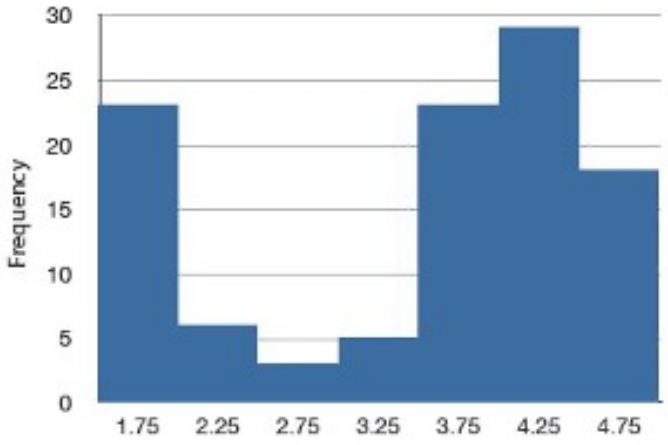

Figure 2.3: Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

2.2 Percentiles[2]

Percentiles are a helpful tool for describing distributions. Many of us have probably encountered percentiles before in the context of standardized exam testing. A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65th percentile.

2.2.1 Three Alternative Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this "Definition 1." The 65th percentile can also be defined as the smallest score that is greater than *or equal to* 65% of the scores. This we will call "Definition 2." Though these two definitions appear very similar, they can sometimes lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of

[2]This section is adapted from David M. Lane. "Percentiles." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/introduction/percentiles.html

scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5 of the scores?

A third way to compute percentiles is a weighted average of the percentiles computed according to the first two definitions. The details of computing percentiles under this third definition are a bit complicated, but fortunately, statistical software can easily do the calculations for us. Since it is unlikely you will need to compute percentiles by hand, we leave the details of these computations to the appendix appearing at the end of this chapter. Despite its complexity, the third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50th percentile. Unless otherwise specified, when we refer to "percentile," we will be referring to this third definition of percentiles.

2.3 Measures of Central Tendency[3]

2.3.1 Mean

The **mean**[4] is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. When using symbols and formulas to represent different statistics, we often distinguish between whether we are looking at a "sample" or a "population." We'll cover this distinction in more detail in Chapter 5. For now, think of a pollster who has conducted a survey with a sample of 1000 people. Even though only 1000 people responded to the survey, the pollster is actually interested in estimating the attitudes of a larger population—the entire public.

The symbol μ is used for the mean of a population. The symbol $\bar{X}$ is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum X}{N}$$

where $\sum X$ is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for $\bar{X}$ is essentially identical:

[3]This section is adapted from David M. Lane. "Measures of Central Tendency." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/summarizing_distributions/measures.html

[4]More specifically, the arithmetic mean is the most common measure of central tendency. Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

$$\bar{X} = \frac{\sum X}{n}$$

where $\sum X$ is the sum of all the numbers in the sample and n is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 2.4 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$$

Table 2.4: Number of touchdown passes.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20 20 19 19 18 18
18 18 16 15 14 14 14 12 12 9 6

2.3.2 Median

The **median** is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 2.4, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

2.3.2.1 Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5.

2.3.3 Mode

The **mode** is the most frequently occurring value. For the data in Table 2.4, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see Section 2.1.2). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2.5 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2.5: Grouped frequency distribution.

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

2.3.4 Comparing Measures of Central Tendency[5]

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean and median are equal, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Consider the positively skewed distribution we saw earlier in the chapter (Figure 2.1). Measures of central tendency for this distribution are shown in Table 2.6. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90.

Table 2.6: Measures of central tendency for the test scores.

Measure	Value
Mode	84.00
Median	90.00
Mean	91.58

[5]This section is adapted from David M. Lane. "Comparing Measures of Central Tendency." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/summarizing_distributions/comparing_measures.html

The distribution of baseball salaries (in 1994) shown in Figure 2.4 has a much more pronounced skew than the distribution in Figure 2.1.

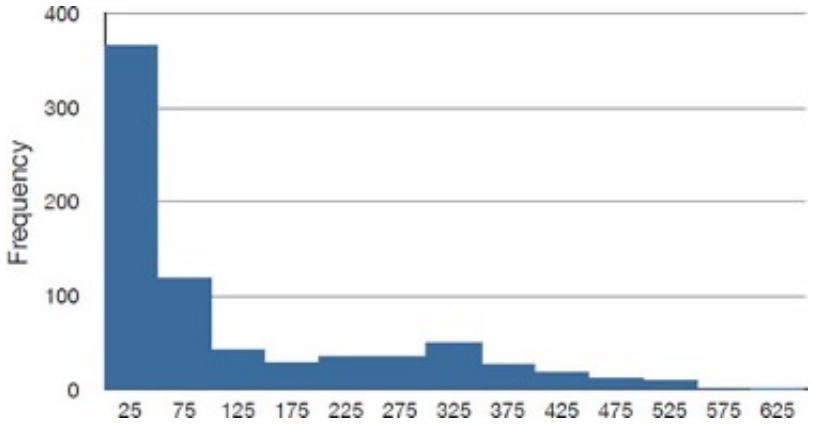

Figure 2.4: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars: 25 equals 250,000).

Table 2.7 shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central tendency is sufficient for data such as these. If you were asked the very general question: "So, what do baseball players make?" and answered with the mean of $1,183,000, you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of $250,000 or the median of $500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean and the median. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table 2.7: Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	**Value**
Mode	250
Median	500
Mean	1,183

2.4 Measures of Spread[6]

2.4.1 What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider Figure 2.5 and Figure 2.6. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

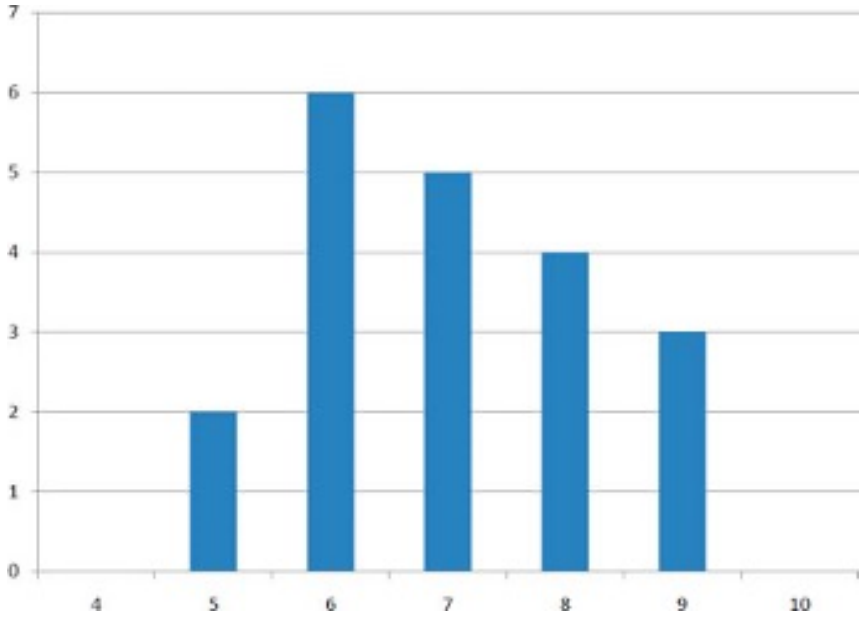

Figure 2.5: Quiz 1

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this section we will discuss measures of the variability of a distribution. There are four frequently used measures of variability: the range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

[6]This section is adapted from David M. Lane. "Measures of Variability." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/summarizing_distributions/variability.html

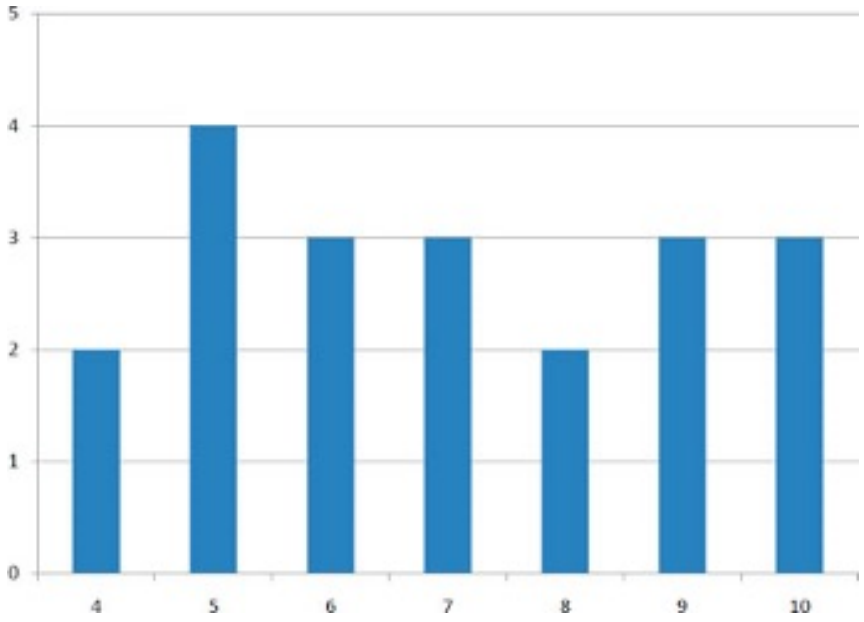

Figure 2.6: Quiz 2

2.4.2 Range

The **range** is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown above. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

Be careful, though, about using the range to make comparisons. The range can be sensitive to sample size: a larger sample offers greater opportunity for the range to expand. In the example above, both quizzes have the same number of observations (20), so using the range to make a comparison is perfectly fine. However, if we were to compare two samples with different numbers of observations, then the range is not necessarily a good way to compare the variability of the samples.

2.4.3 Interquartile Range

The **interquartile range** (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

$$IQR = \text{ 75th percentile } - \text{ 25th percentile}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. We'll see in Section 2.5 that when creating boxplots, the 75th percentile is also called the upper hinge and the 25th percentile is called the lower hinge. Thus, the interquartile range is neatly depicted by the box portion of a boxplot.

2.4.4 Variance

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the **variance** is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 2.8. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.

Table 2.8: Calculation of Variance for Quiz 1 scores.

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1

Scores	Deviation from Mean	Squared Deviation
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
	Means	
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that its variance is 6.7. The formula for the variance is:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

where $\sigma 2$ is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$$

where s^2 is the estimate of the variance and $\bar{X}$ is the sample mean.

Note that $\bar{X}$ is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used. While it is not easy to succinctly explain why we divide by $n-1$ rather than simply n, the simulation "estimating variance"[7] illustrates the bias that arises if we use n as the denominator in the formula.

Let's look at a concrete example of calculating the sample variance. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$\bar{X} = (1+2+4+5)/4 = 12/4 = 3$$

[7] https://onlinestatbook.com/2/summarizing_distributions/variance_est.html

$$s^2 = [(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2]/(4-1)$$
$$= (4+1+1+4)/3 = 10/3 = 3.333$$

2.4.5 Standard Deviation

The **standard deviation** is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. We can interpret the standard deviation of X as approximating the typical distance between a given value of X and the mean of X. For example, suppose I tell you about a prison where the prisoners have a mean age of 42 years with a standard deviation of 8 years. If I randomly select one prisoner and ask you to guess their age, you should probably guess 42 since I've told you that is the mean. But even though 42 is your best guess, you can expect your guess to be off by about 8 years since the standard deviation is 8 (meaning the typical distance between a random prisoner's age and the mean age is approximately 8). You can't say ahead of time which direction your guess is likely to be off (guessing too old versus too young), just that you are likely to miss the reality for a randomly-selected individual by about 8 years on a typical guess (though any one guess may happen to be closer or further than 8 years).

2.5 Box Plots[8]

Box plots are useful for making comparisons and identifying **outliers**, meaning unusually large or small values for a variable. We will explain box plots with the help of data from an in-class experiment. As part of the "Stroop Interference Case Study,"[9] students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve *parallel box plots*.

There are several steps in constructing a box plot. The first relies on the 25^{th}, 50^{th}, and 75^{th} percentiles in the distribution of scores. Figure 2.7 shows how these three statistics are used. For each gender, we draw a box extending from the 25^{th} percentile to the 75^{th} percentile. The 50^{th} percentile is drawn inside the box. Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile. The data for the women in our sample are shown in Table 2.9.

[8] This section is adapted from David M. Lane. "Box Plots." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/graphing_distributions/boxplots.html

[9] https://onlinestatbook.com/2/case_studies/stroop.html

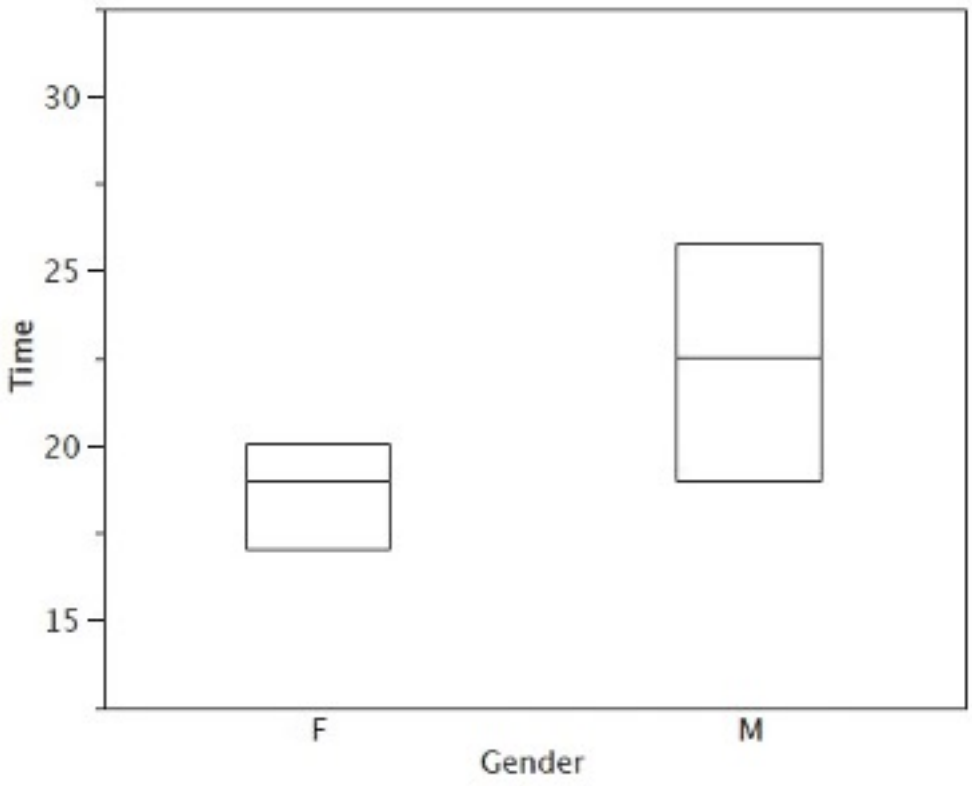

Figure 2.7: The first step in creating box plots.

Table 2.9: Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

Before proceeding, the terminology in Table 2.10 is helpful.

Table 2.10: Box plot terms and values for women's times.

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "o's" and far out values are indicated by asterisks (*). In our data, there are no far out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 2.8.

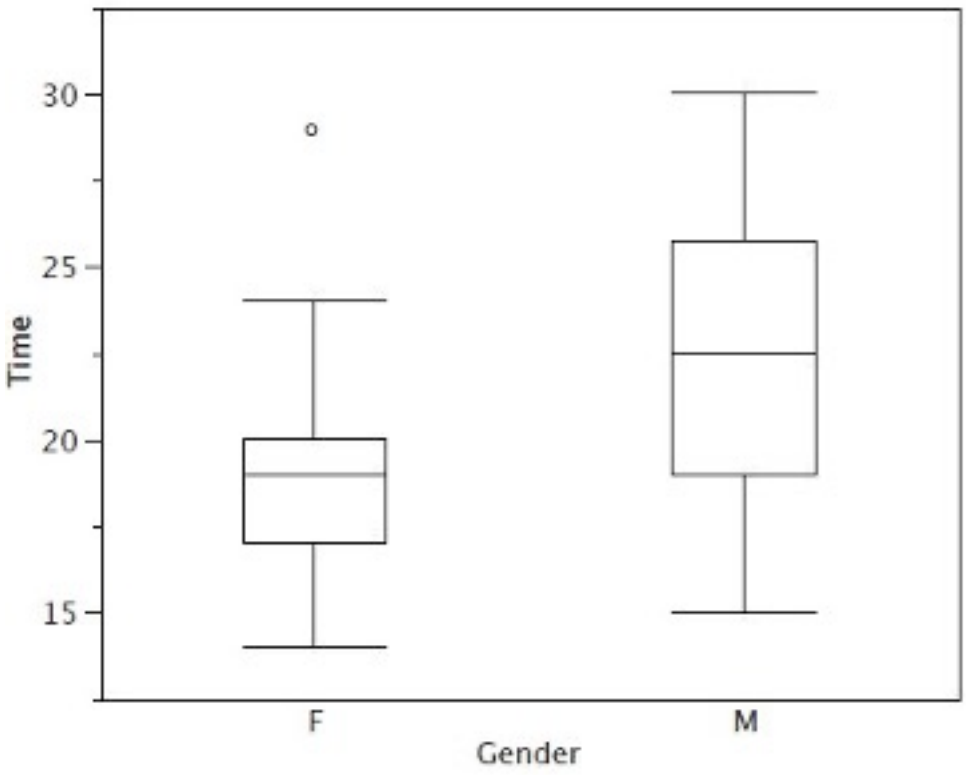

Figure 2.8: The box plots with the whiskers and outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 2.9 shows the result of adding means to our box plots.

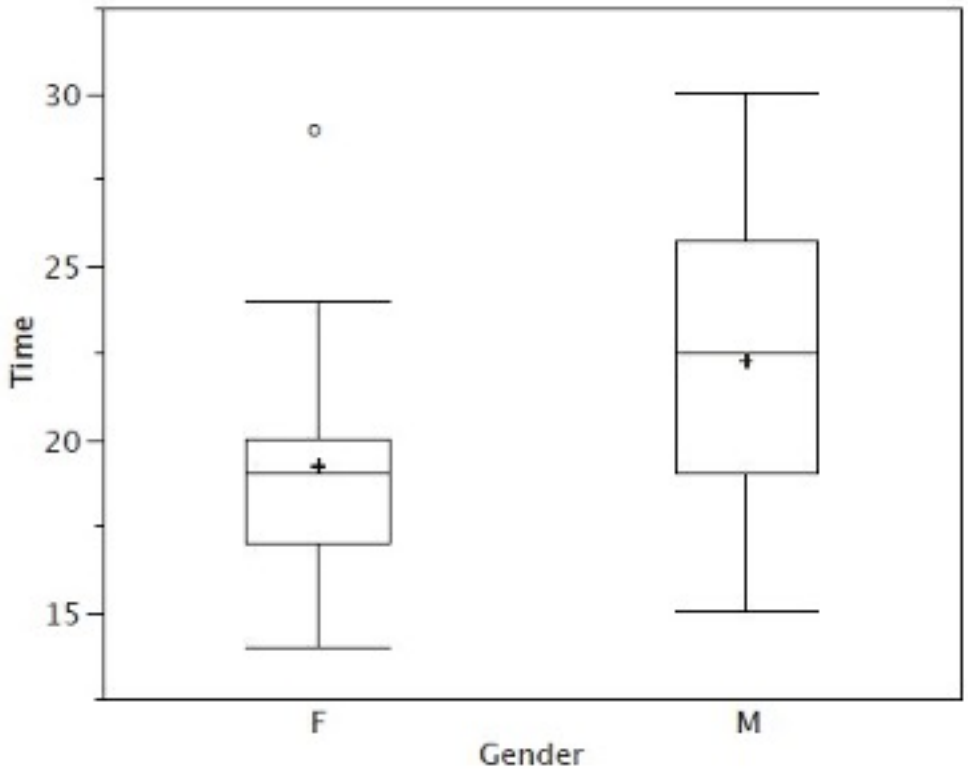

Figure 2.9: The completed box plots.

Figure 2.9 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women's times are between 17 and 20 seconds, whereas half the men's times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 2.10 shows the box plot for the women's data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram.

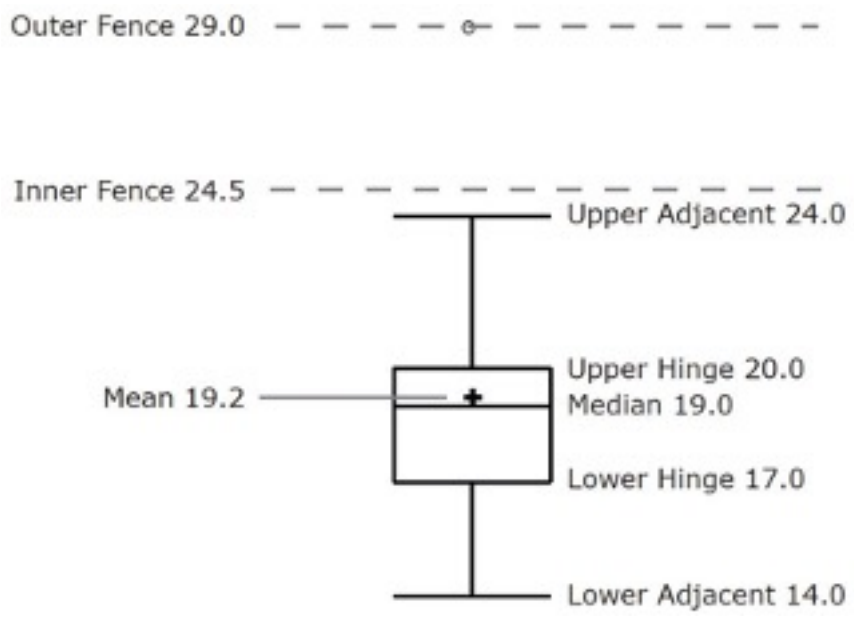

Figure 2.10: The box plot for the women's data with detailed labels.

2.5.1 Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 2.11 are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.
5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

Each dot in Figure 2.11 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to "jitter" the points. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure 2.12 shows what jittering looks like.

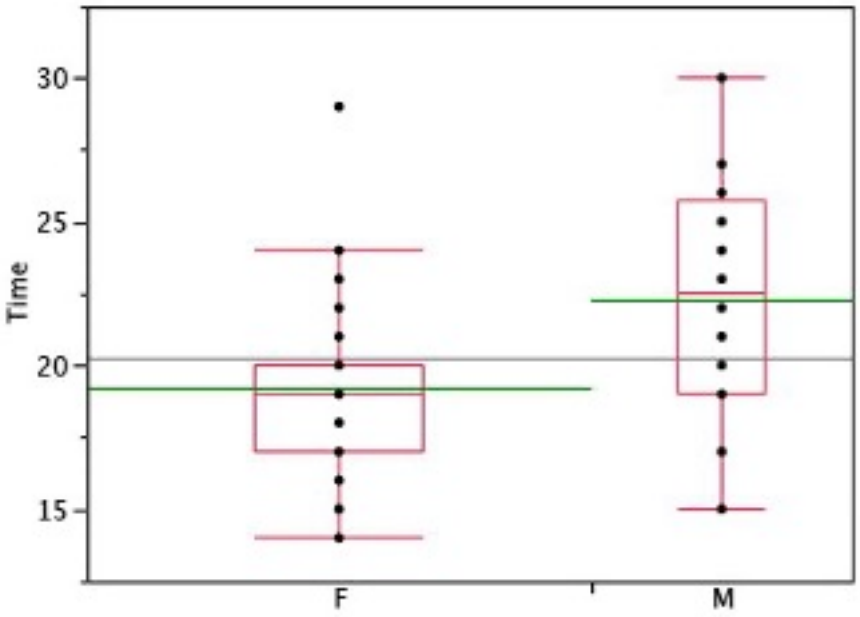

Figure 2.11: Box plots showing the individual scores and the means.

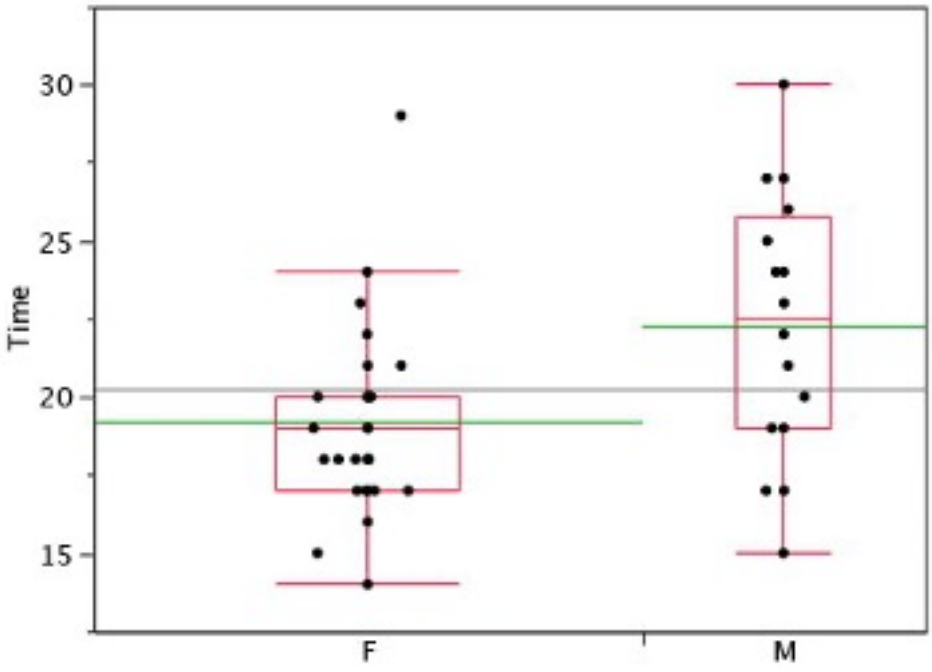

Figure 2.12: Box plots with the individual scores jittered.

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

2.6 Transforming Variables[10]

Often it is necessary to transform data from one measurement scale to another. For example, you might want to convert height measured in feet to height measured in inches. Table 2.11 shows the heights of four people measured in both feet and inches. To transform feet to inches, you simply multiply by 12. Similarly, to transform inches to feet, you divide by 12.

Table 2.11: Converting between feet and inches.

Feet	**Inches**
5.00	60
6.25	75
5.50	66
5.75	69

Some conversions require that you multiply by a number and then add a second number. A good example of this is the transformation between degrees Centigrade and degrees Fahrenheit. Table 2.12 shows the temperatures of 5 US cities in the early afternoon of November 16, 2002.

Table 2.12: Temperatures in 5 cities on 11/16/2002.

City	**Degrees Fahrenheit**	**Degrees Centigrade**
Houston	54	12.22
Chicago	37	2.78
Minneapo-	31	-0.56
lis	78	25.56
Miami	70	21.11
Phoenix		

[10]The initial part of this section is adapted from David M. Lane. "Linear Transformations." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/introduction/linear_transforms.html. There is also material adapted from David M. Lane. "Standard Normal Distribution." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/normal_distribution/standard_normal.html.

The formula to transform Centigrade to Fahrenheit is:

$$F = 1.8C + 32$$

The formula for converting from Fahrenheit to Centigrade is

$$C = 0.5556F - 17.778$$

The transformation consists of multiplying by a constant and then adding a second constant. For the conversion from Centigrade to Fahrenheit, the first constant is 1.8 and the second is 32.

Figure 2.13 shows a plot of degrees Centigrade as a function of degrees Fahrenheit. Notice that the points form a straight line. This will always be the case if the transformation from one scale to another consists of multiplying by one constant and then adding a second constant. Such transformations are therefore called **linear transformations**.

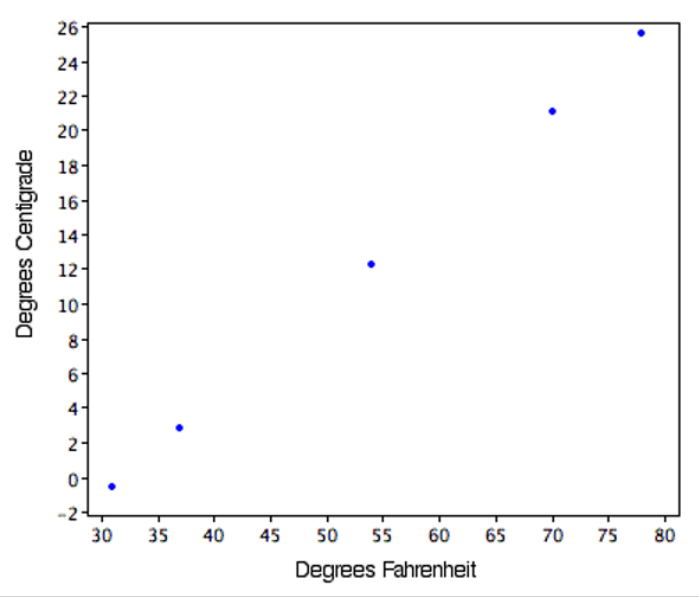

Figure 2.13: Degrees Centigrade as a function of degrees Fahrenheit.

2.6.1 Standardization (Z Scores)

So far, we've discussed transformations that are probably familiar to you. A type of transformation that may be new to you is **standardization** or creating Z scores. A value from any distribution can be transformed into a Z score using the following formula:

$$Z = \frac{(X - \mu)}{\sigma}$$

where Z is the new value, X is the value on the original distribution, μ is the mean of the original distribution, and σ is the standard deviation of the original distribution.

As a simple application, suppose you want the Z score for a value of 26 taken from a distribution with a mean of 50 and a standard deviation of 10. Applying the formula, we obtain:

$$Z = (26 - 50)/10 = -2.4$$

If all the values in a distribution are transformed to Z scores, then the new distribution will have a mean of 0 and a standard deviation of 1. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called standardizing the distribution. Sometimes it will be easier to work with a standardized version of a variable.

2.6.2 Log Transformations[11]

Sometimes it is also useful to use transformations that are not linear. For example, the log transformation can sometimes be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics (see Chapter 5).

Figure 2.14 shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel.

Chapter 2 Appendix: Calculating Percentiles Under the Third Definition

Let's begin with an example. Consider the 25th percentile for the 8 numbers in Table 2.13. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

[11]This subsection is adapted from David M. Lane. "Log Transformations." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/transformations/log.html

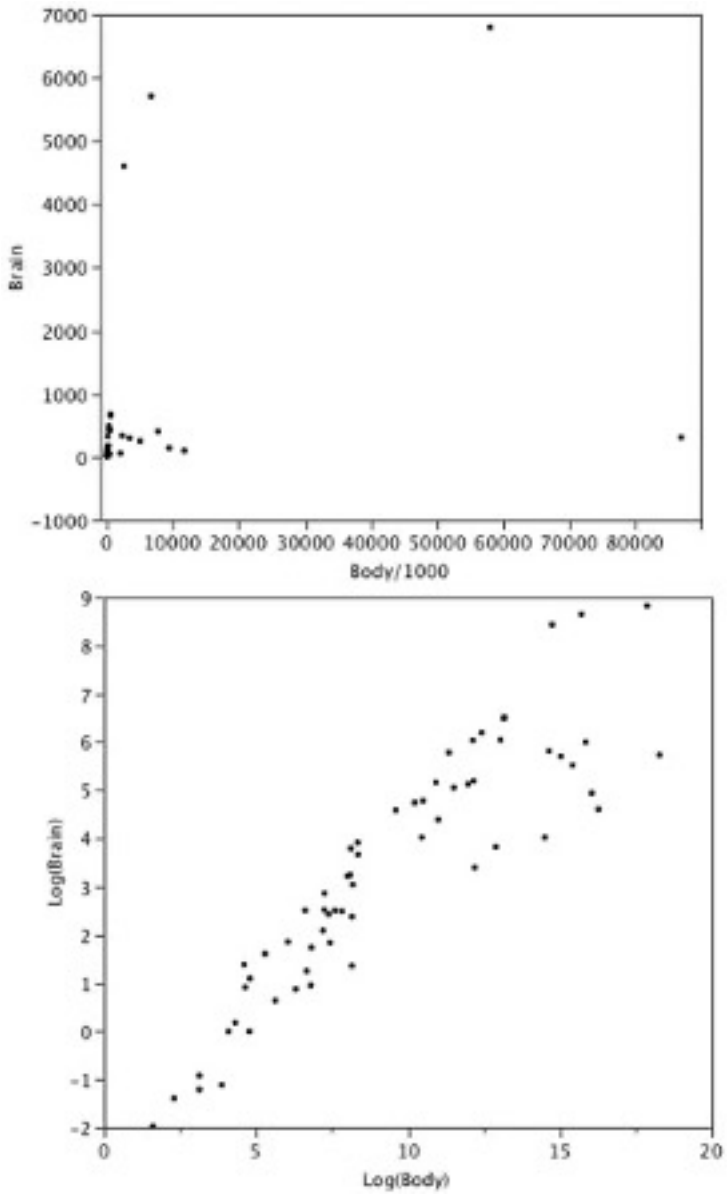

Figure 2.14: Scatter plots of brain weight as a function of body weight in terms of both raw data (upper panel) and log-transformed data (lower panel).

Table 2.13: Test Scores.

Number	**Rank**
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = P/100 \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8 + 1) = 9/4 = 2.25.$$

If R is an integer, the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, $IR = 2$.
2. Define FR as the fractional portion of R. For this example, $FR = 0.25$.
3. Find the scores with Rank IR and with Rank $IR + 1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.
4. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score. For these data, this is (0.25)(7 - 5) + 5 = 5.5.

Therefore, the 25th percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25th percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25th percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table 2.14.

Table 2.14: 20 quiz scores.

Score	**Rank**
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

We will compute the 25th and the 85th percentiles. For the 25th,

$$R = 25/100 \times (20 + 1) = 21/4 = 5.25.$$

$$IR = 5 \text{ and } FR = 0.25.$$

Since the score with a rank of IR (which is 5) and the score with a rank of $IR + 1$ (which is 6) are both equal to 5, the 25th percentile is 5. In terms of the formula:

$$\text{25th percentile } = (.25) \times (5 - 5) + 5 = 5.$$

For the 85th percentile,

$$R = 85/100 \times (20 + 1) = 17.85.$$

$$IR = 17 \text{ and } FR = 0.85$$

Caution: FR does not generally equal the percentile to be computed as it does here.

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85th percentile is:

$$(0.85)(10 - 9) + 9 = 9.85$$

Consider the 50th percentile of the numbers 2, 3, 5, 9.

$$R = 50/100 \times (4 + 1) = 2.5.$$

$$IR = 2 \text{ and } FR = 0.5.$$

The score with a rank of IR is 3 and the score with a rank of $IR + 1$ is 5. Therefore, the 50th percentile is:

$$(0.5)(5 - 3) + 3 = 4.$$

Finally, consider the 50th percentile of the numbers 2, 3, 5, 9, 11.

$$R = 50/100 \times (5 + 1) = 3.$$

$$IR = 3 \text{ and } FR = 0.$$

Whenever $FR = 0$, you simply find the number with rank IR. In this case, the third number is equal to 5, so the 50th percentile is 5. You will also get the right answer if you apply the general formula:

$$\text{50th percentile } = (0.00)(9 - 5) + 5 = 5.$$

3 Tools for Describing the Relationship Between Two Quantitative Variables

3.1 Introduction to Bivariate Data[1]

Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two quantitative variables for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. Let's begin by asking if people tend to marry other people of about the same age. Our experience tells us "yes," but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table 3.1 below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big surprise, but at least the data bear out our experiences, which is not always the case.

Table 3.1: Sample of spousal ages of 10 White American Couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

The pairs of ages in Table 3.1 are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a histogram (see Figure 3.1) and by a mean and standard deviation (See Table 3.2).

[1]This section is adapted from Rudy Guerra and David M. Lane. "Introduction to Bivariate Data." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/describing_bivariate_data/intro.html

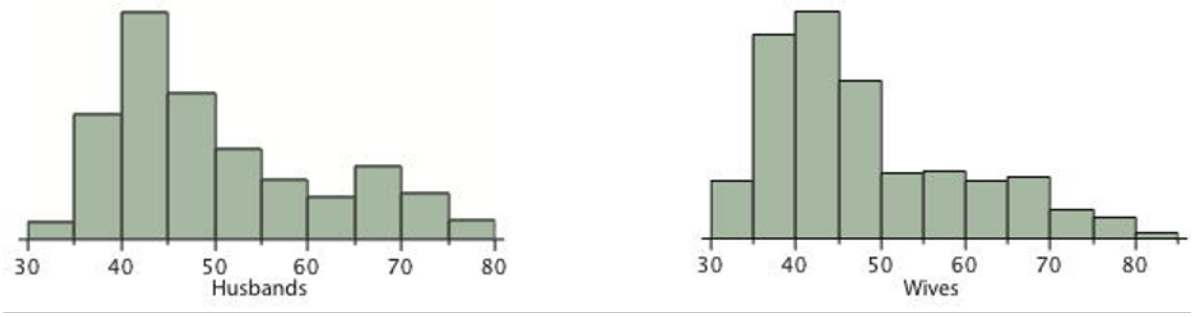

Figure 3.1: Histograms of spousal ages.

Table 3.2: Means and standard deviations of spousal ages.

	Mean	**Standard Deviation**
Husbands	49	11
Wives	47	11

Each distribution is fairly skewed with a long right tail. From Table 3.1 we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples has younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

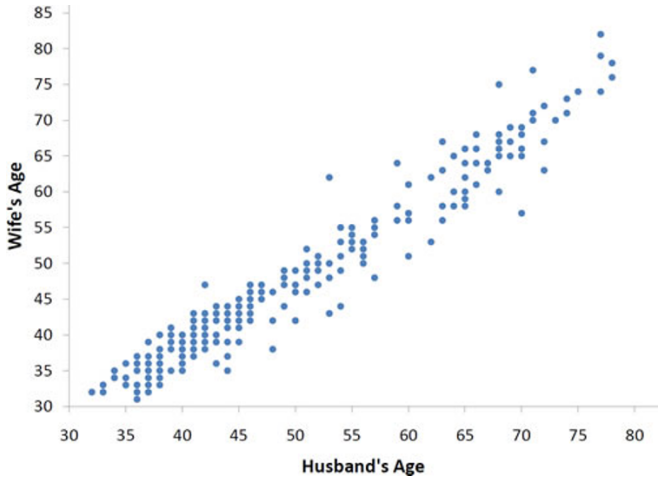

Figure 3.2: Scatter plot showing wife's age as a function of husband's age, r = 0.97.

We can learn much more by displaying the bivariate data in a graphical form that maintains the pairing. Figure 3.2 shows a scatter plot of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

There are two important characteristics of the data revealed by Figure 3.2. First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (Y) increases with the second variable (X), we say that X and Y have a **positive association**. Conversely, when Y decreases as X increases, we say that they have a **negative association**.

Second, the points cluster along a straight line. When this occurs, the relationship is called a **linear relationship**.

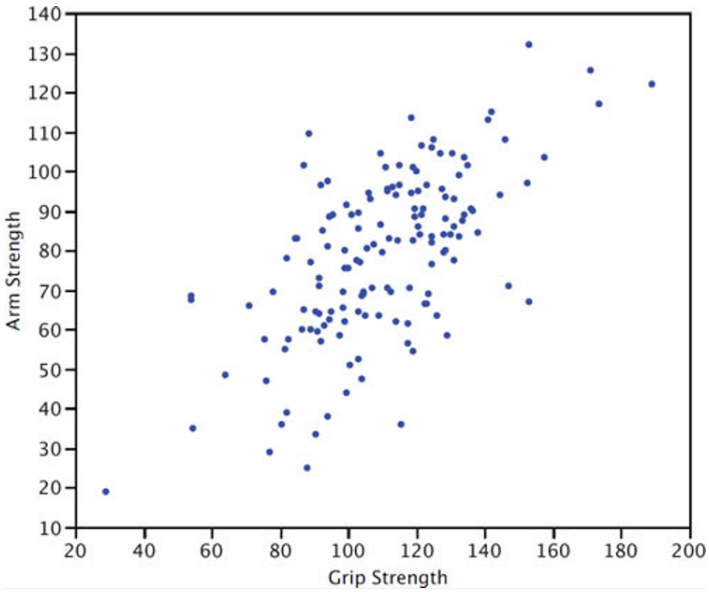

Figure 3.3: Scatter plot of Grip Strength and Arm Strength, r = 0.63.

Figure 3.3 shows a scatter plot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.

Not all scatter plots show linear relationships. Figure 3.4 shows the results of an experiment conducted by Galileo on projectile motion.[2] In the experiment, Galileo rolled balls down an incline and measured how far they traveled as a function of the release height. It is clear from Figure 3.4 that the relationship between "Release Height" and "Distance Traveled" is

[2] https://www.amstat.org/publications/jse/v3n1/datasets.dickey.html

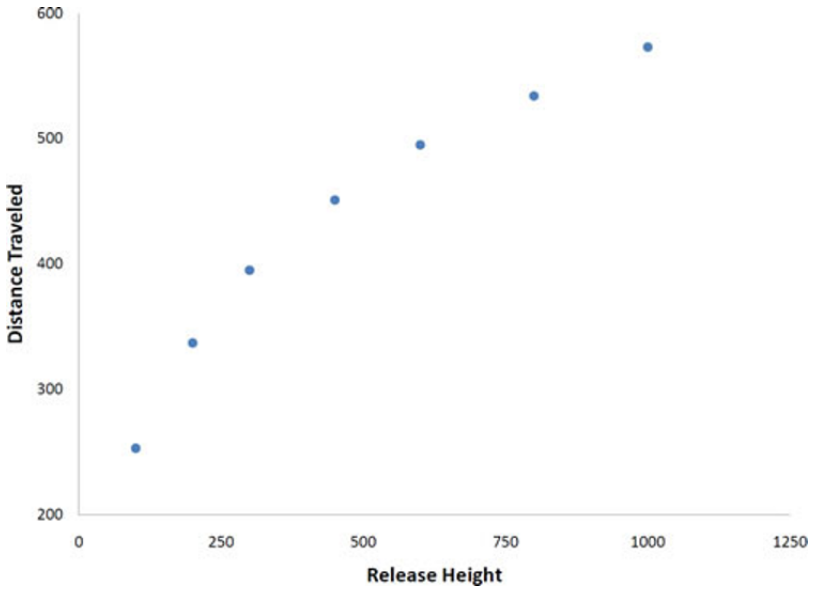

Figure 3.4: Galileo's data showing a non-linear relationship.

not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola (a type of curved line).

With large datasets, scatter plots can be difficult to interpret since observations tend to stack on top of one another (often leaving only outliers to show up distinctly). The positive association between work experience and wages in Figure 3.5 is not particularly strong, but it is a bit difficult to discern exactly what is going on because the graph is crowded with so many observations (2,246 in total).[3] Larger samples will make this problem even worse.

Fortunately, we can create what is called a binned scatter plot. Rather than plotting every observation, we can divide the x axis into bins and plot just one dot per bin. The dot will indicate the mean or median value of the y variable among observations within that bin. While binned scatter plots can bring greater clarity when working large datasets, be aware that binned scatter plots tend to make relationships look stronger than they really are. This is because aggregating data (by bin, in this case) tends to reduce the noisiness associated with idiosyncrasies in individual observations. See how clear and consistent the relationship between work experience and wages appears in the binned scatter plot (Figure 3.6), despite the relatively weak relationship in the individual-level data. Notice too that the ranges of both axes are smaller than in the original scatter plot (Figure 3.5), since the binning means that the full ranges of the variables are no longer depicted.

Scatter plots that show linear relationships between variables can differ in several ways including

[3]US National Longitudinal Survey of Young Women (NLSW), 1988 extract. Accessed via Stata.

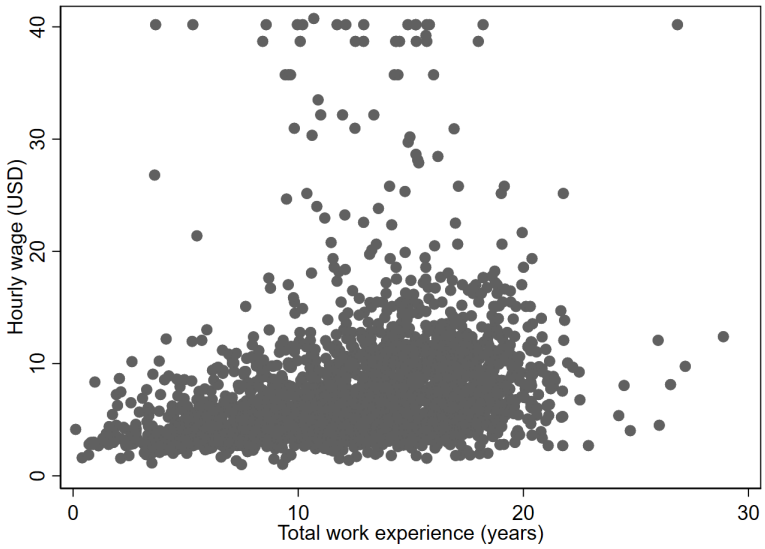

Figure 3.5: Scatter plot of 1988 wages and work experience from a sample of young women, r=0.27.

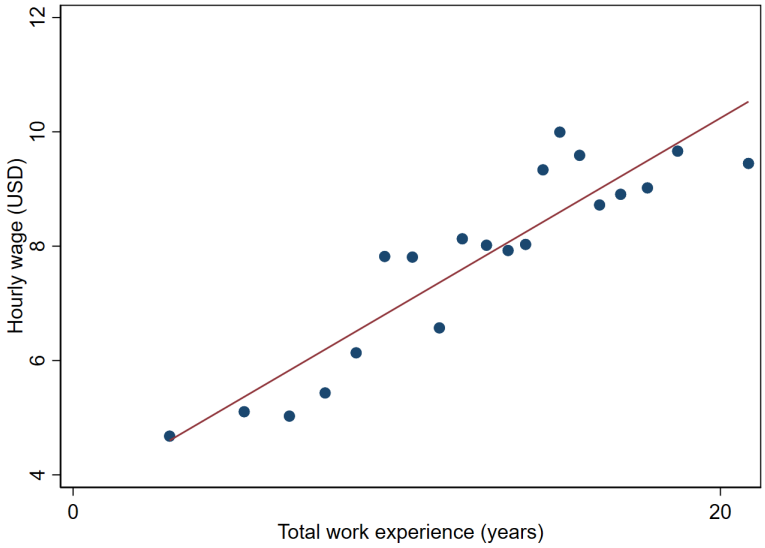

Figure 3.6: Binned scatter plot of the 1988 data for wages and work experienced among young women

the slope of the line about which they cluster and how tightly the points cluster about the line. We now turn our attention to a statistical measure of the strength of the relationship between two quantitative variables.

3.2 What is Correlation?[4]

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the **correlation coefficient**. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Figure 3.7 shows a scatter plot for which r = 1.

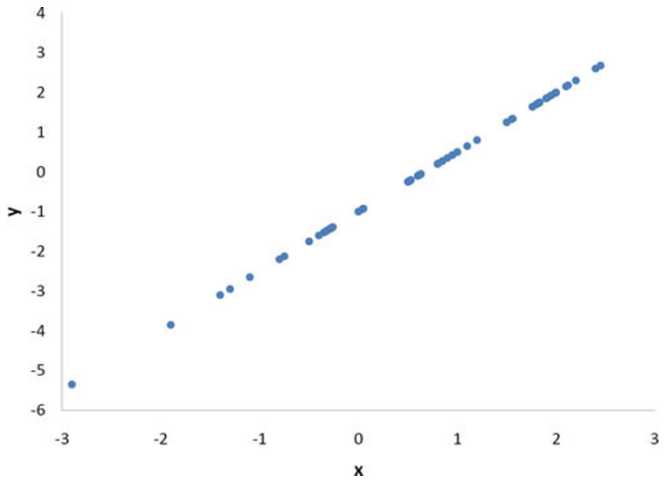

Figure 3.7: A perfect positive linear relationship, r = 1.

With real data, you would not expect to get values of r of exactly -1, 0, or 1. The data for spousal ages shown earlier in this chapter in Figure 3.2 has an r of 0.97.

[4]This section is adapted from David M. Lane. "Values of the Pearson Correlation." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/describing_bivariate_data/pearson.html

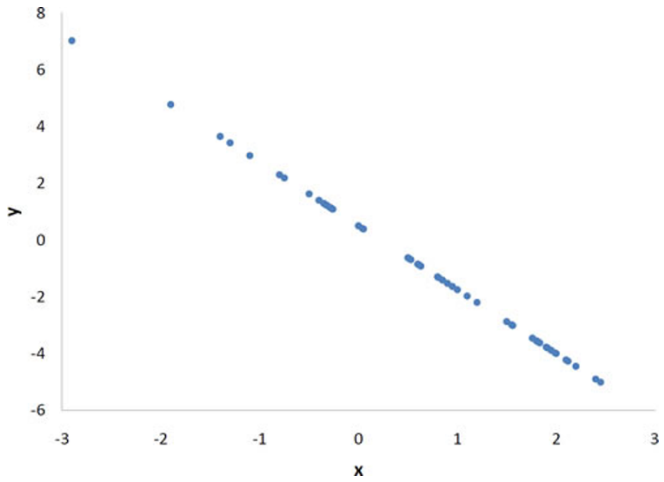

Figure 3.8: A perfect negative linear relationship, r = 1.

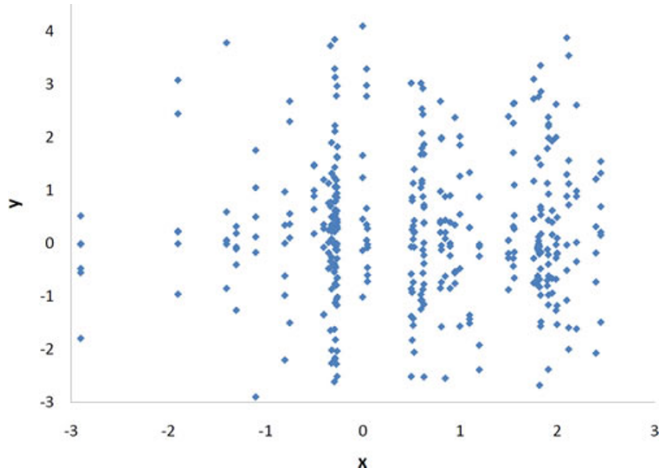

Figure 3.9: A scatter plot for which r = 0. Notice that there is no relationship between X and Y.

The relationship between grip strength and arm strength depicted in Figure 3.3 (also described in the introductory section) is 0.63.

3.3 How Correlation is Calculated[5]

There are several formulas that can be used to compute Pearson's correlation. Some formulas make more conceptual sense whereas others are easier to actually compute. We are going to begin with a formula that makes more conceptual sense.

We are going to compute the correlation between the variables X and Y shown in Table 3.3. We begin by computing the mean for X and subtracting this mean from all values of X. The new variable is called "x." The variable "y" is computed similarly. The variables x and y are said to be deviation scores because each score is a deviation from the mean. Notice that the means of x and y are both 0. Next we create a new column by multiplying x and y.

Before proceeding with the calculations, let's consider why the sum of the xy column reveals the relationship between X and Y. If there were no relationship between X and Y, then positive values of x would be just as likely to be paired with negative values of y as with positive values. This would make negative values of xy as likely as positive values and the sum would be small. On the other hand, consider Table 3.3 in which high values of X are associated with high values of Y and low values of X are associated with low values of Y. You can see that positive values of x are associated with positive values of y and negative values of x are associated with negative values of y. In all cases, the product of x and y is positive, resulting in a high total for the xy column. Finally, if there were a negative relationship then positive values of x would be associated with negative values of y and negative values of x would be associated with positive values of y. This would lead to negative values for xy.

Table 3.3: Calculation of r.

	X	**Y**	**x**	**y**	**xy**	**x^2**	**y^2**
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

[5]This section is adapted from David M. Lane. "Computing Pearson's r." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/describing_bivariate_data/calculation.html

Pearson's r is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet. To achieve this property, Pearson's correlation is computed by dividing the sum of the xy column ($\sum xy$) by the square root of the product of the sum of the x^2 column ($\sum x^2$) and the sum of the y^2 column ($\sum y^2$). The resulting formula is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

and therefore:

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

An alternative computational formula that avoids the step of computing deviation scores is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)}\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

3.4 Introduction to Linear Regression[6]

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the **dependent variable** and is referred to as Y. The variable we are basing our predictions on is called the **independent variable** and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 3.4 are plotted in Figure 3.10. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 3.4: Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30

[6]This section is adapted from David M. Lane. "Introduction to Linear Regression." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/intro.html

X	Y
4.00	3.75
5.00	2.25

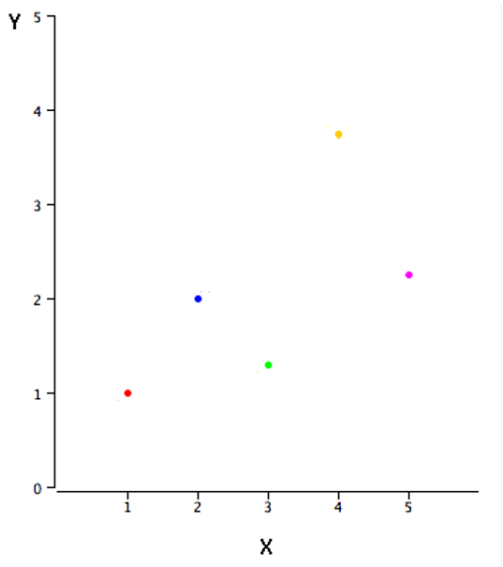

Figure 3.10: A scatter plot of the example data.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 3.11 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 3.5 shows the predicted values ($\hat{Y}$) and the errors of prediction ($Y - \hat{Y}$). For example, the first point has a Y of 1.00 and a predicted Y (called $\hat{Y}$) of 1.21. Therefore, its error of prediction is -0.21.

Table 3.5: Example data.

X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 3.11. The last column in Table 3.5 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 3.5 is lower than it would be for any other regression line.

The formula for a regression line is

$$\hat{Y} = \alpha + \beta X$$

where $\hat{Y}$ is the predicted score, α is the Y-intercept, and β is the slope of the line. The equation for the line in Figure 3.11 is

$$\hat{Y} = 0.785 + 0.425X$$

Using this equation, we can calculate predictions for Y based on the value of X. For $X = 1$,

$$\hat{Y} = 0.785 + (0.425)(1) = 1.21.$$

For $X = 2$,

$$\hat{Y} = 0.785 + (0.425)(2) = 1.64.$$

3.4.1 A Real Example

The case study "SAT and College GPA"[7] contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university grade point average (GPA) if we knew his or her high sch

Figure 3.12 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

$$\widehat{University\_GPA} = 1.097 + 0.675 \times HighSchoolGPA$$

[7] https://onlinestatbook.com/2/case_studies/sat.html

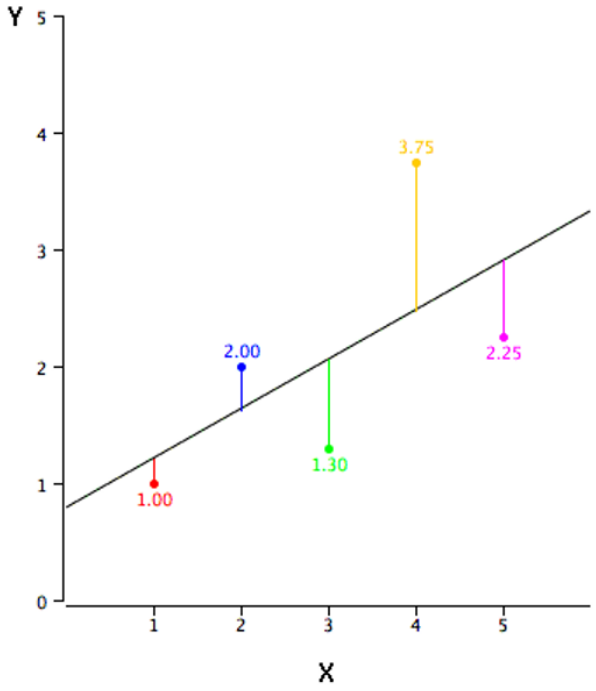

Figure 3.11: A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

$$\widehat{University\_GPA} = 1.097 + 0.675 \times (3) = 3.12.$$

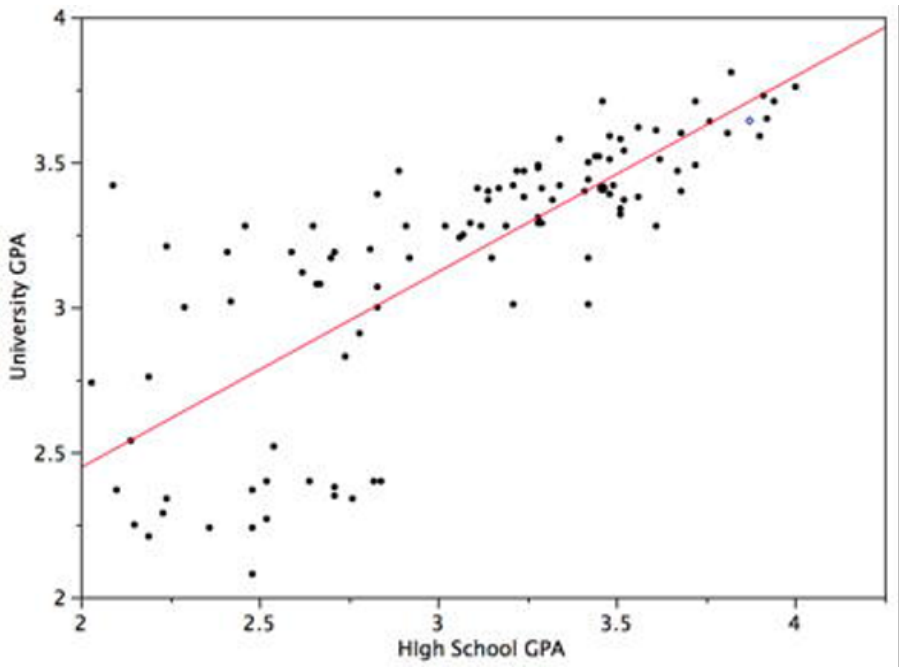

Figure 3.12: University GPA as a function of High School GPA.

3.5 Quick Guide to Interpreting Regression Results[8]

Many social science papers report their main results in the form of a regression table. It's fairly easy to get started interpreting these results using the three S's:[9]

- **Significance**: Is the relationship between the two variables strong enough (relative to the precision of the estimate) to be considered statistically reliable?[10] To assess this, check the p-value. For now, you can use the following rule-of-thumb:

[8]This section is written by Nathan Favero.
[9]Wheelan, C. (2010.) *Introduction to Public Policy.* New York: W. W. Norton & Company.
[10]In other words, can we conclude it is signal rather than noise? See: Fricker Jr, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *The American Statistician*, *73*(sup1), 374-384.

  - If $p < 0.05$: the relationship is statistically significant; proceed to evaluating sign and size
  - If $p > 0.05$: results are somewhat indeterminate; any association detected between the two variables could easily be caused by coincidence or random "noise" (so you may want to skip evaluating sign and size)

- **Sign**: Is the relationship positive or negative? Check whether the coefficient has a negative value.
  - Positive coefficient: as the independent variable *increases*, our prediction for the dependent variable *increases*
  - Negative coefficient: as the independent variable *increases*, our prediction for the dependent variable *decreases*
  - Note about odds ratios: For certain types of (non-linear) regression, odds ratios (which always take on positive values) are sometimes displayed instead of coefficients; with an odds ratio, a value greater than one indicates a positive relationship while a value smaller than one indicates a negative relationship

- **Size**: How big is the (predictive) effect? This S is often the most difficult to make sense of, and sometimes you may not have enough information to meaningfully evaluate it (e.g., if the units of measurement for a variable are not clearly explained).
  - For linear models: A one-unit increase in the independent variable means we change our prediction for the dependent variable by $\hat{\beta}_i$ (where $\hat{\beta}_i$ represents the value of the coefficient estimate)[11]
  - For non-linear models: Interpreting the size of a coefficient is typically more complicated than for a linear model; look for the authors' explanation of effect size or "magnitude" of association

Table 3.6 provides an example of regression results in a format similar to what you may encounter in many research publications. Note, however, that many publications do not list exact p-values; instead, they often use one or more asterisks (*) to denote coefficients with p-values smaller than 0.05 (sometimes also flagging p-values falling below various other thresholds).

Table 3.6: Results for a regression with computer science GPA as the dependent variable.

	Coef.	Std. err.	p-value
verb_sat	0.0017	0.0010	0.10

[11] Why $\hat{\beta}_i$ and not simply β_i? We use the hat symbol (ˆ) to indicate an estimate or prediction. So by including the hat, we are implying that our coefficient is an estimate. We'll learn more about estimation in Chapter 5.

	Coef.	Std. err.	p-value
math_sat	0.0048	0.0012	0.00014
(intercept)	-0.91	0.42	0.033
n	105		
r^2	0.487		

Up til now, we have only discussed simple regression, in which we have a single independent variable. But in Table 3.6, we find results for a regression where two independent variables representing students' university entrance exam scores—the verbal (verb_sat) and math (math_sat) sections of the SAT—are jointly used to predict students' GPA in computer science classes. It turns out that regression can easily be performed with multiple independent variables, as described in Chapter 12. When we have multiple independent variables, we evaluate each one on its own terms when working through the three S's. For the results in Table 3.6, we can interpret the results as follows:

- verb_sat: The p-value for this variable (0.10) is greater than 0.05, so this variable is not statistically significant. This means we could not establish a reliable link between verbal SAT scores and computer science GPA.[12] Maybe there is no link, or maybe we would need more data to detect the link. Since we don't find statistical significance, we don't necessarily need to interpret the sign or size.
- math_sat: The p-value (0.00014) is smaller than 0.05, so math_sat is a statistically significant predictor of computer science GPA. The coefficient (0.0048) has a positive sign, so students with higher math SAT scores are predicted to have higher computer science GPAs. When it comes to size, a one-point increase in math SAT (e.g., someone with a 501 versus someone with a 500) yields a prediction for the computer science GPA that is 0.0048 points higher. That seems very small, but a one-point increase on an SAT is barely noticeable (and not actually possible if scores are always multiples of ten). In this case, we can get a better sense of size if we consider an increase of 100 points in the math SAT, which requires multiplying the coefficient by 100. A 100-point increase in the math SAT (e.g., someone with a 600 versus someone with a 500) predicts a computer science GPA that is 0.48 points higher ($0.0048 \times 100 = 0.48$). This is nearly half a grade point higher and would be quite noticeable to most students. Thus, the size of predictive effect now seems reasonably large.

Note that we do not need to apply the three S's to the intercept (which can also be labeled the "constant") because it is not a variable. Table 3.6 also contains some additional information frquently shown in regression tables: standard errors (which we will learn more about in

[12]Note, however, that the absence of evidence is not necessarily evidence of absence. There could very well be a link between verbal SAT scores and computer science GPA—just one that we cannot reliably detect with this analysis (e.g., because our sample is too small to precisely estimate the association).

Chapter 7), the sample size (n=105), and r-squared (a statistic often used to describe how well the regression model overall explains variation in the dependent variable).

Remember that the three S's are just a starting point. But they should be enough to help you follow along a little easier when reading the results sections of many research publications. And notice that the third S (size) is helps equip us to address the third Question to Always Ask about Data ("how big is the difference?"). If you've started working with a statistical software package by now, you can run your own regression models and try using the three S's to help you understand the results.

4 Relationships with Qualitative Variables

The way that we study relationships among variables depends on the types of variables we are examining—specifically whether they are quantitative or qualitative. In the prior chapter, we learned about how to describe the relationship (or association) between two quantitative variables. In this chapter, we will consider the case when one or both variables are qualitative.[1]

Where do ordinal variables fit into this discussion? For many types of analysis, ordinal variables can be treated as either qualitative (for maximum flexibility) or as quantitative (if we are willing to assign fixed numbers to the various values, as we saw in the first example in Section 1.4). Often, analysis will be simpler if we treat ordinal variables as quantitative, so that can be a good place to start. You can always then follow up with an analysis treating the ordinal variable as qualitative, which allows you to check if the results indicate something similar to what you saw when you treated the variable as quantitative.

4.1 Describing the Relationship between Two Qualitative Variables[2]

The relationship between two qualitative variables can be examined using a contingency table (also known as "crosstabs"). For example, Table 4.1 shows the data from the Mediterranean Diet and Health case study,[3] in which 605 survivors of a heart attack were randomly assigned to follow either (1) a low-fat diet similar to one recommended at the time by the American Heart Association (AHA) or (2) a Mediterranean-type diet rich in vegetables, fruits, and grains. The percentages shown in this table are calculated by row, meaning that they are found by dividing the frequency in each cell by the total for that row, shown in the final column.

Table 4.1: Frequencies and Percentages by Row for Diet and Health Study.

	Outcome				**Total**
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	
AHA	15	24	25	239	303
	4.95%	7.92%	8.25%	78.88%	100%

[1]The first two paragraphs of this chapter are written by Nathan Favero.
[2]The initial material in this section was written by Nathan Favero.
[3]https://onlinestatbook.com/2/case_studies/diet.html

	Outcome				**Total**
Mediter-	7	14	8	273	302
ranean	2.32%	4.64%	2.65%	90.40%	100%
Total	22	38	33	512	605
	3.64%	6.28%	5.45%	84.63%	100%

Drawing conclusions based on column or row percentages takes some care; it is easy to make mistakes. For example, consider the AHA-Cancers cell, which shows 4.95%. Because this seems like a small percentage, it is tempting to conclude that being on the AHA diet is not associated with having cancer. However, that conclusion would be wrong. Since the percentages have been calculated by row, the key to finding meaningful patterns will be to compare across the percentages found within the same column. In this case, we see that the 4.95% in the AHA-Cancers cell is over twice as large as the 2.32% shown in the Mediterranean-Cancers cell right below it. To make sense of these percentages, remember that we have calculated percentages by row, meaning that the frequency in the cell is divided by the total at the far right of the table. So the 4.95% means that among all those on the AHA diet (303 individuals in total), 4.95% have cancer. By comparison, only around 2% (2.32% to be precise) of those on a Mediterranean diet have cancer. Both percentages are small, because cancer is (fortunately) a relatively rare outcome; at the bottom of the table we see that only 3.64% of participants overall have cancer. Still, since the percentage is higher among those on the AHA diet than on the Mediterranean diet, we can say that the AHA diet is positively associated with having cancer.

If we continue moving down the table column by column, we see that rates of heart disease are higher among those on the AHA diet (7.92% for fatal and 8.25% for non-fatal) than among those on the Mediterranean diet (4.64% and 2.65%). Finally, around 90% of those on the Mediterranean diet are healthy, while just under 79% of those on the AHA diet are.

We can reach an equivalent conclusion by examining percentages that are calculated by column, as shown in Table 4.2. Here, percentages are calculated by dividing the frequency in each cell by the total shown at the bottom of the column.

Table 4.2: Frequencies and Percentages by Column for Diet and Health Study.

	Outcome				**Total**
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	
AHA	15	24	25	239	303
	68.18%	63.16%	75.76%	46.68%	50.08%
Mediter-	7	14	8	273	302
ranean	31.82%	36.84%	24.24%	53.32%	49.92%

	Outcome				Total
Total	22 100%	38 100%	33 100%	512 100%	605 100%

The AHA-Cancers cell shows 68.18%, indicating that around 68% of those who have cancer are on the AHA diet. In the case of this particular dataset, the split of participants between the two diets is approximately 50-50 (as shown by the percentages in the final column), so finding that the percentage of those with cancer who are on the AHA diet is well above 50% actually does indicate that individuals on the AHA diet are overrepresented among those with cancer. But remember, you cannot generally assume that 50% is a reasonable baseline for comparison; if, for example, 80% of study participants were on the AHA diet, seeing any percentage smaller than 80% in the AHA-Cancers cell would indicate that those on the AHA diet were underrepresented among those with cancer.

As a general rule, when percentages are calculated by columns, the key to finding meaningful patterns is making comparisons across the columns (within the same row), such as noticing that the 68.18% in the AHA-Cancers cell is greater than the 50.08% in the AHA Total cell. One can also see that while people on the AHA diet make up a clear majority of those with heart disease (both fatal and non-fatal), among those who are healthy only around 47% are on the AHA diet. As a clear contrast, those on the Mediterranean diet make up around 53% of the healthy subjects but just 32% of those with cancer, 37% of those with fatal heart disease, and 24% of those with non-fatal heart disease. Thus, we see consistent indication that outcomes are better for subjects on the Mediterranean diet, as opposed to the AHA one.

Interpreting results from contingency tables is not always intuitive. You may need to practice several times before you can confidently describe the relationship between two qualitative variables based on a contingency table.

4.1.1 Visualizing a Qualitative-Qualitative Relationship[4]

Let us now visualize the relationship we have just discussed based on contingency tables. Bar charts are often an effective choice, although there are many different ways that such charts can be configured, and it may take some trial and error to find the most effective chart for any given data. In this case, because the number of individuals for each of the two diets is roughly the same (303 for the AHA diet and 302 for the Mediterranean diet), it makes it fairly easy to create a helpful chart. We clearly see from Figure 4.1 that more people on the Mediterranean diet have the healthy outcome, while the various negative outcomes are all more common among people on the AHA diet.

[4]Except for the first paragraph (written by Nathan Favero), this subsection is adapted from David M. Lane. "Graphing Qualitative Variables." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/graphing_distributions/graphing_qualitative.html

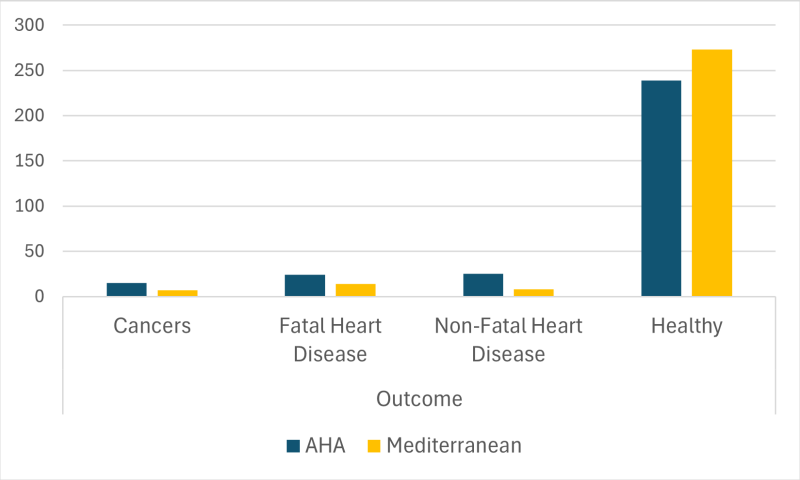

Figure 4.1: A bar chart of health outcomes by diet

More broadly, suppose we are comparing the results of different surveys, or of different conditions within the same overall survey. Because a qualitative variable allows us to divide our sample into sub-samples (based on survey or condition) and then examine distributions for each sub-sample, we can say that we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 4.2 shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

The bars in Figure 4.2 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels.

4.2 Describing the Relationship between a Qualitative and a Quantitative Variable[5]

When we have a quantitative variable and a qualitative variable, we typically begin by dividing our sample into sub-samples according to the qualitative variable. For example, if we are examining the qualitative variable sex (indicating male or female), we can divide the sample into males and females. We can then compute measures of central tendency for the sub-samples and compare them. We already saw an example of this in Section 1.4, where the relationship

[5]The initial paragraph for this section was written by Nathan Favero.

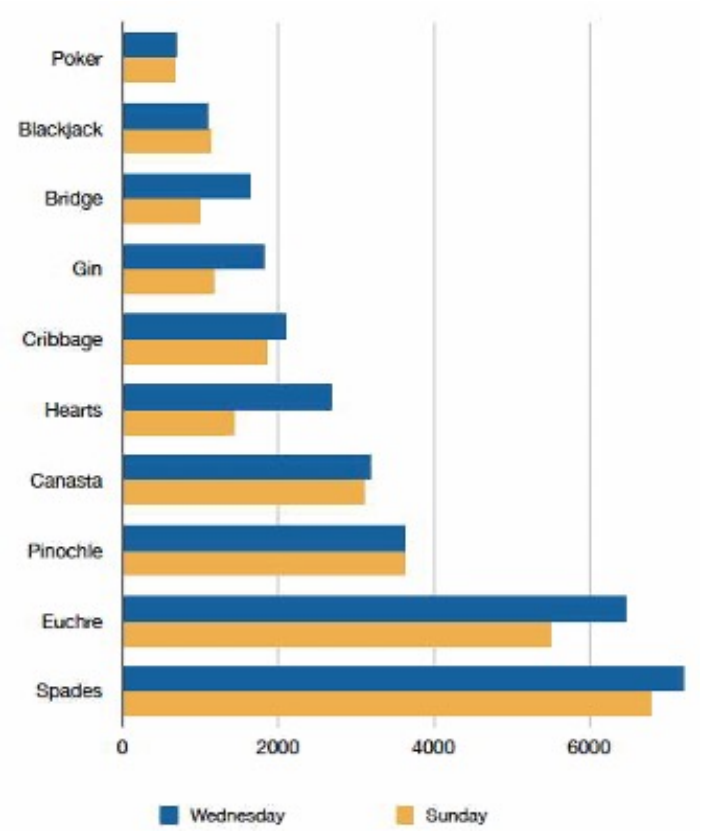

Figure 4.2: A bar chart of the number of people playing different card games on Sunday and Wednesday.

between sex and support for innovation was evaluated by comparing the means of the support-for-innovation variable among males versus females. Typically, we compare means, but we can also compare medians if we are concerned about our results being sensitive to outliers. Ultimately, we want to observe whether the means/medians appear similar across groups or if some groups have meaningfully higher means/medians than others. In the example with sex and support for innovation, the means were nearly the same (3.787 for females, 3.798 for males), suggesting that the two variables are not related (at least not to a meaningful extent).

4.2.1 Visualizing a Quantitative-Qualitative Relationship[6]

Bar charts are often used to compare the means of different experimental conditions. Figure 4.3 shows the mean time it took one of us (DL) to move the mouse to either a small target or a large target. On average, more time was required for small targets than for large ones.

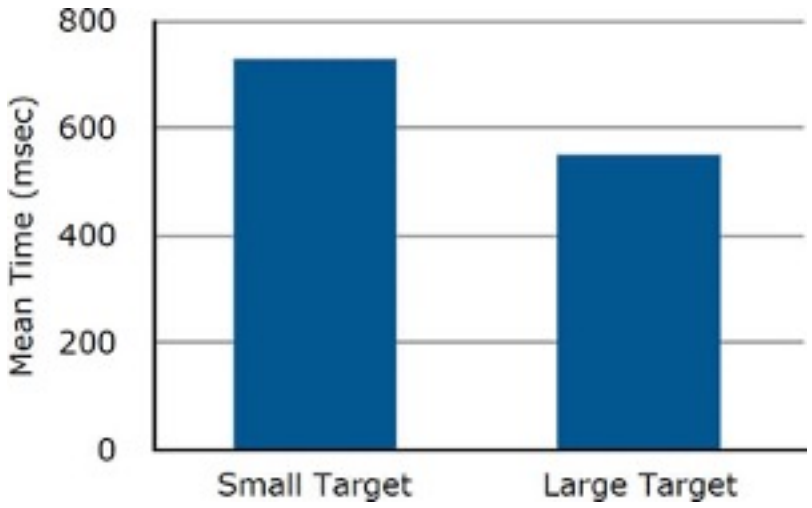

Figure 4.3: Bar chart showing the means for the two conditions.

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the mouse-movement data is shown in Figure 4.4. You can see that Figure 4.4 reveals more about the distribution of movement times than does Figure 4.3.

[6]This subsection is adapted from David M. Lane. "Bar Charts." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/graphing_distributions/bar_chart.html

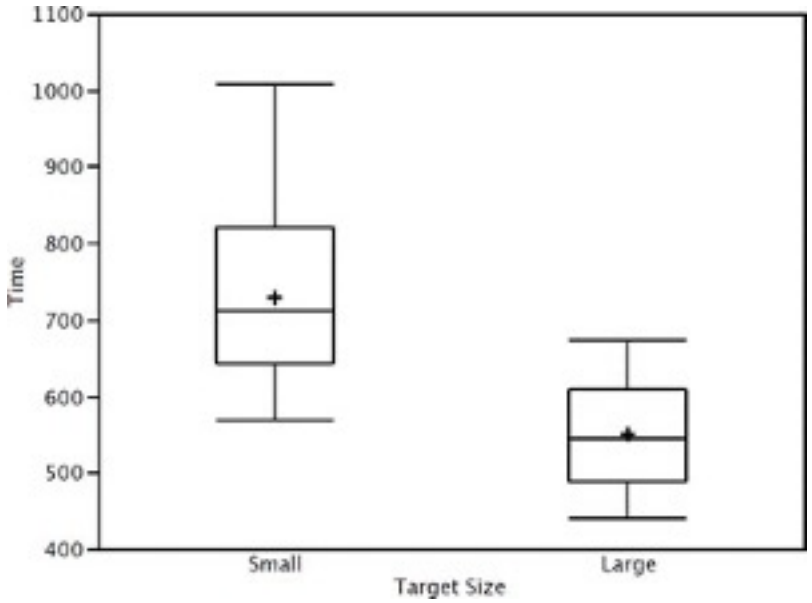

Figure 4.4: Box plots of times to move the mouse to the small and large targets.

4.2.2 Regression with a Qualitative Independent Variable[7]

Regression analysis is a very flexible set of tools that can be used to examine almost any sort of relationship among variables. For now, we will focus on understanding how it can be used to describe the relationship between a *quantitative* dependent variable and a *qualitative* independent variable.

Let's say I'm interested in studying how personality relates to gender. The most common model of personality in psychology is called the "Big Five." One common measure consists of 50 survey items (the IPIP Big-Five Factor Markers), with 10 items for each of the five personality traits from the model.[8] Figure 4.5 shows some of these questions and how they are formatted.

For now, I decide to focus on whether people are introverted or extroverted. Extroverts are outgoing and tend to enjoy interacting with others. Extroverts will tend to agree with the statement "I am the life of the party" while introverts will tend to agree with the item "I don't talk a lot."

I find a dataset that contains many responses to the Big Five personality questions as well as information on the gender of each respondent.[9] There are 10 different questions related to

[7]This subsection was written by Nathan Favero.

[8]Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological assessment*, *4*(1), 26.

[9]https://openpsychometrics.org/_rawdata/ (the file I used is called "BIG5.zip")

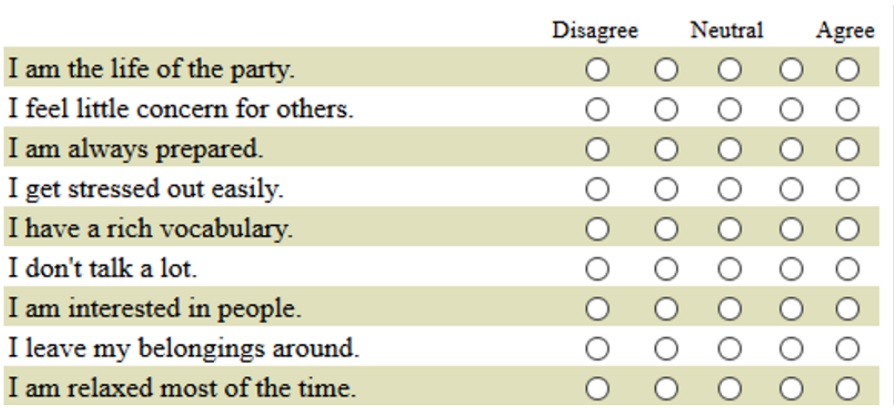

Figure 4.5: Example survey items used in the "Big Five" case study.

extraversion, and the dataset has one variable for each of these 10 questions (Figure 4.6). The variable labeled e1 shows responses to the item "I am the life of the party." A value of 1 means the respondent disagrees with this statement, while a 3 indicates neutral, and a 5 means they agree.

For all of the odd-numbered extraversion questions (e1, e3, e5, etc.), agreement indicates extraversion. For the even-numbered items (e2, e4, e6, etc.), agreement indicates introversion. To create a single extraversion variable that combines responses from all 10 survey items, I create a tally, adding up all the values for odd-numbered questions and then subtracting the responses to the even-numbered questions. An extreme extrovert will have a 5 for all the odd-numbered questions and a 1 for all of the even-numbered ones, giving them a score of 20 ($5 \times 5 - 5 \times 1 = 20$). An extreme introvert will have a -20 since they will answer 1 to all the odd-numbered questions and 5 to all the even-numbered ones ($5 \times 1 - 5 \times 5 = -20$).

As Figure 4.7 indicates, most people lie somewhere in the middle between introversion and extraversion.

Our gender variable was measured by asking respondents "What is your gender?" and they could choose from male (coded as a 1), female (2), or other (3). In a moment, we'll consider those who responded "other," but for now, let's just look at those who chose either male or female.

4.2.2.1 Predicting extraversion using gender

If I want to describe differences in extraversion by gender in this dataset, I can compute the mean value of extraversion for males and for females. It turns out that males have an average extraversion of -0.46 while females' average level of extraversion is 0.53. Thus, the average female is about 1-point more extroverted than the average male. But of course, there

Data Editor (Edit) - [Untitled]

File Edit View Data Tools

gender[1] 2

	gender	e1	e2	e3	e4	e5
1	2	1	5	1	5	2
2	2	3	2	4	2	3
3	1	1	3	4	2	4
4	2	1	5	1	5	1
5	2	5	1	5	1	5
6	2	3	1	4	2	4
7	1	2	5	2	4	2
8	2	2	2	3	3	3
9	2	2	3	4	2	4
10	2	5	1	5	1	5
11	1	4	2	5	2	5

Figure 4.6: Screenshot of some observations and variable values in the "Big Five" dataset.

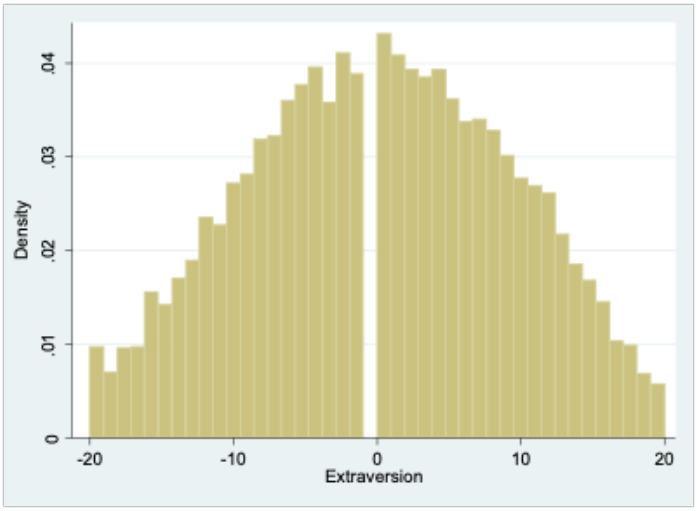

Figure 4.7: Distribution of the extraversion trait.

is abundant variation in extraversion among both groups, as seen clearly in Figure 4.8. There are plenty of females who are introverts and plenty of males who are extroverts.

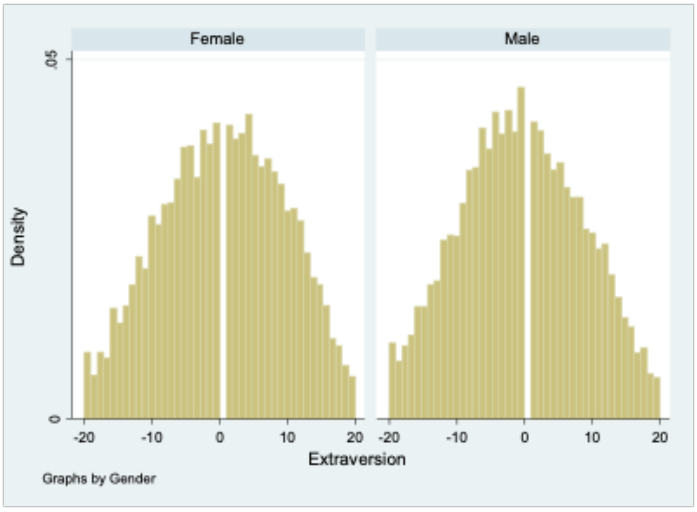

Figure 4.8: Distribution of extraversion by female/male identity.

If you asked me to guess the extraversion level of someone in this dataset and the only thing you told me about them was their gender, my best bet would probably be to guess the average extraversion level for someone of that gender. So for a female I knew nothing else about, I would guess their extraversion to be 0.53, while for a male I'd guess -0.46.

When we're working with data, sometimes it's helpful to express how I would make a guess about a dependent variable (extraversion) based on an independent variable (gender) using a mathematical formula. In fact, this is exactly what we do when we run a regression. There are many ways I could write this formula, but I'll show just two for now. First, I could write:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male \tag{4.1}$$

Notice I've added a "hat" above the name of the variable *Extraversion*; this hat means that I'm making a guess about the value of that variable (I'm guessing the level of extraversion based on gender). The equation has two other variables *Female* and *Male*, and these two variables will take on a value of 1 if the person's gender is equal to the name of the variable and will otherwise take on a value of 0. For a female, *Female* will equal 1 and *Male* will equal 0, giving us:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) = 0.53$$

So our guess for the level of extraversion ($\widehat{Extraversion}$) of a female we know nothing about is 0.53.

For a male, our guess is:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (1) = -0.46$$

There's a second way I can write my formula, which will turn out to be more useful in the future when we come to consider multiple factors at the same time that might help us predict the value of a dependent variable. Rather than having two variables to represent gender in my equation, I can just use one:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male \tag{4.2}$$

In Equation 4.2, we start from female as our baseline. Notice that the first number we see (0.53) is our guess for the value of extraversion for a female. When we're considering a female, Male=0, so:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 0 = 0.53$$

Thus, we get the right prediction for females from this equation, even though we didn't include a variable specifically for females. If we have a male, Male=1, so we get:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 1 = -0.46$$

This is the same prediction we got before. Remember, I decided to initially just analyze respondents who selected either male or female. Since we are only considering two categories (male or female), and each respondent is either a male or a female, saying $Male = 1$ lets me know that $Female = 0$. It's actually repetitive in this context to both say that $Male = 1$ and $Female = 0$. Similarly, saying $Male = 0$ implies that $Female = 1$. So I can simplify my equation by just including one variable to indicate binary gender.

Notice that in Equation 4.2, the number next to *Male* is equal to the difference between the average level of extraversion for females and the average level for males (0.53-(-0.46)=0.99). This is because Equation 4.2 starts with females as the baseline, so to get our prediction for males, we have to adjust our baseline prediction by the average difference for males.

Equation 4.2 is also typically how we will arrange our equation when we're running a regression.

4.2.2.2 Prediction with more than two categories for gender

I now move beyond the gender binary and consider the "other" category in survey responses. The average level of extraversion among those identifying as neither male nor female is -5.66. This is notably more introverted (on average) than those who identify as male or female. As with males and females, there is still considerable variation among those identifying as "other" gender (Figure 4.9).

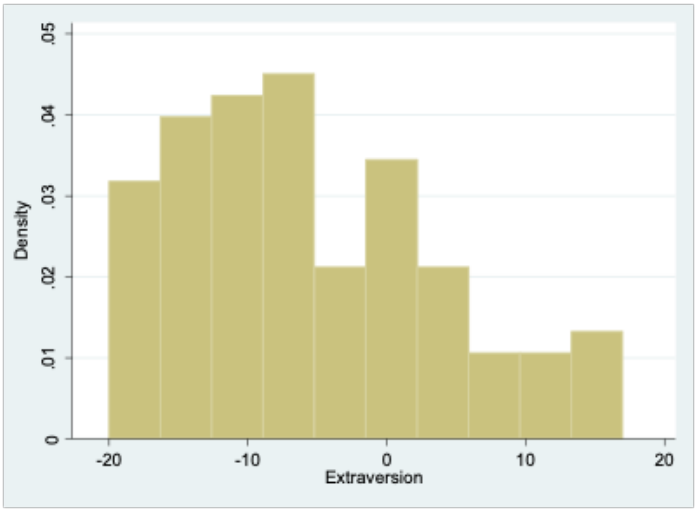

Figure 4.9: Distribution of extraversion trait among those identifying with "other" gender (neither male nor female).

The number of respondents selecting "other" is relatively small (102), so it's not terribly surprising that this histogram looks a bit choppier than the ones we saw before.

Again, if we had to make a guess about the level of extraversion of someone, and all we knew about that person was that their gender is other, we would probably want to guess the mean value among other-gender respondents (-5.66). Modifying Equation 4.1 to incorporate a third category is relatively straightforward:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male - 5.69 \times Other \tag{4.3}$$

For someone who identifies as female, we would plug in $Female = 1$, $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) - 5.66 \times (0) = 0.53$$

If someone identifies as other-gender, we would use $Female = 0$, $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (0) - 5.66 \times (1) = -5.66$$

We can also return to the format of Equation 4.2 but modify it to include the other category. This is how we will typically write our equation if we are doing a regression:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male - 6.19 \times Other \tag{4.4}$$

Now that there are three possible values for gender (female, male, and other), knowing the value of *Male* doesn't necessarily allow us to conclude what the value of female is. If $Male = 0$, the individual could identify as either female or other. So we have to include a second variable. In this case, we chose to include the variable *Other*. If we know the values of *Male* and *Other*, we can always figure out the value of *Female* by process of elimination.

For an other-gender person, we plug in $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (1) = -5.66$$

When considering a female, we use $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (0) = 0.53$$

Equation 4.3 and Equation 4.4 indicate equivalent predictions for extraversion based on gender; they just offer this information in two different formats. Equation 4.4 might be a bit trickier to understand for now, but it will become very useful in the future.

Notice that we can talk about gender either as one qualitative variable with three possible values (female, male, or other), or we can talk about it as a series of three variables (*Female*, *Male*, and *Other*) that can each take on a value of either 0 or 1 (we call such variables "binary" or "dummy" variables). This can make things a bit confusing, but the important thing to remember is that when we have a qualitative variable with more than two categories, we'll need to break out the categories into a set of dummy variables for purposes of representing the qualitative variable in an equation.

However, as Equation 4.2 and Equation 4.4 illustrate, we don't necessarily need a dummy variable for every single category. Specifically, whenever we want to create an equation with a qualitative independent variable in a format like Equation 4.2 or Equation 4.4, the number of dummy variables should be equal to the number of categories minus one. Since our gender variable can take on three possible values in this example, we included two independent variables in Equation 4.4. No dummy variable is included for female, so we call female the **omitted category**, also known as the **reference group** or **baseline category**. Remember, the first number in Equation 4.4 is 0.53, which represents our guess for females—the baseline category.

If we instead had a qualitative variable with five categories, we would include four dummy variables in our equation.

As we can see in the appendix, it doesn't really matter which group we select as the omitted category because we get equivalent predictions regardless of which one we choose.

Chapter 4 Appendix: Regression with a Qualitative Dependent Variable

This appendix will demonstrate one way to perform regression with a qualitative dependent variable, using output from the statistical software program *Stata*. Suppose I want to build a model of voting. I decide to use the 2016 American National Election Studies[^qual-associations-11] survey results to try to understand how race is associated with voting. Respondents in the 2016 survey were asked about who they voted for in 2012, and I'm going to focus on their 2012 voting patterns for now. Using the statistical software package Stata to conduct my analysis, I find the following distributions for my two main variables of interest:

```
. tab vote
PRE: RECALL OF LAST (2012) PRESIDENTAL  |
                            VOTE CHOICE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                      1. Barack Obama |      1,728       56.58       56.58
                       2. Mitt Romney |      1,268       41.52       98.10
                     5. Other SPECIFY |         58        1.90      100.00
----------------------------------------+-----------------------------------
                                Total |      3,054      100.00

. tab race
  PRE: SUMMARY - R SELF-IDENTIFIED RACE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                1. White, non-Hispanic |      3,038       71.68       71.68
                2. Black, non-Hispanic |        398        9.39       81.08
3. Asian, native Hawaiian or other Paci |        148        3.49       84.57
4. Native American or Alaska Native, no |         27        0.64       85.21
                           5. Hispanic |        450       10.62       95.82
6. Other non-Hispanic incl multiple rac |        177        4.18      100.00
----------------------------------------+-----------------------------------
                                Total |      4,238      100.00
```

Notice that my dependent variable (vote) is qualitative. It can take on three possible values: voted for Obama, voted for Romney, or voted for other. I can build a simple set of regression

models to see how race predicts vote choice. The key is to first convert each of the three categories for my dependent variable into its own dummy (or binary) variable—meaning a variable that is always equal to either 0 or 1. I can accomplish this in Stata with the following code:

```
tab vote, gen(vote_)
```

I now have several new variables in my dataset that have names starting with "vote_":

```
. tab vote_1

   vote==1. |
     Barack |
      Obama |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,326       43.42       43.42
          1 |      1,728       56.58      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_2

   vote==2. |
Mitt Romney |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,786       58.48       58.48
          1 |      1,268       41.52      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_3

   vote==5. |
      Other |
    SPECIFY |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,996       98.10       98.10
          1 |         58        1.90      100.00
------------+-----------------------------------
      Total |      3,054      100.00
```

I also convert my race variable into a set of dummy variables by running:

```
tab race, gen(race_)
```

I can then run three regressions, one for each value of my dependent variables. I will use regular linear regression (least squares) for this example, although there are arguably better and more precise models for qualitative dependent variables (e.g., various types of probit and logit regression). Nonetheless, we can get by with linear regression. When using linear regression with a binary dependent variable, we call the model a linear probability model.

Let's start by analyzing voting for Obama (vote_1) as the dependent variable:

```
. reg vote_1 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     76.29
       Model |  83.3981974         5  16.6796395   Prob > F        =    0.0000
    Residual |  662.426572     3,030  .218622631   R-squared       =    0.1118
-------------+----------------------------------   Adj R-squared   =    0.1104
       Total |  745.824769     3,035  .245741275   Root MSE        =    .46757

------------------------------------------------------------------------------
      vote_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   .4972868   .0281049    17.69   0.000     .4421802    .5523934
      race_3 |   .2078207   .0541766     3.84   0.000     .1015941    .3140472
      race_4 |   .1028423   .1353307     0.76   0.447     -.162507    .3681916
      race_5 |   .3135004    .032158     9.75   0.000     .2504466    .3765542
      race_6 |   .1042547   .0441427     2.36   0.018      .017702    .1908075
       _cons |    .480491   .0097901    49.08   0.000     .4612952    .4996868
```

Since our independent variable is qualitative, we have an omitted category. In this case, we've left category 1 (race_1) out of our regression, which indicates non-Hispanic White respondents. Our constant (or y-intercept) indicates the predicted value of the dependent variable when all independent variables are equal to zero. We can see this by writing out the regression equation:

$$\widehat{vote\_1} = .48 + .50race\_2 + .21race\_3 + .10race\_4 + .31race\_5 + .10race\_6 \qquad (4.5)$$

For non-Hispanic White respondents, race_1 equals one and all other race dummy variables equal zero, so we get:

$$\widehat{vote\_1} = .48 + .50(0) + .21(0) + .10(0) + .31(0) + .10(0) = .48$$

Remember, vote_1 is equal to zero if the respondent didn't vote for Obama, and it is equal to one if the respondent did vote for Obama. Our predicted value is neither zero nor one; instead, we get .48. This can be interpreted as indicating the probability of a one. In other words, a non-Hispanic White respondent has a .48 probability of voting for Obama. We can also convert this probability to a percentage by moving the decimal place two spots to the right: a non-Hispanic White person is estimated to have a 48% chance of voting for Obama, according to this model.

Now, let's look at the slope coefficients. The coefficient for (non-Hispanic) Black (race_2) equals .50. Thus, a one-unit increase in race_2 is associated with a .50-unit increase in vote_1. Let's break that down a bit to see if we can create a clearer interpretation. Since race_2 is a dummy variable and non-Hispanic White is the omitted category, a one-unit increase in race_2 correspondents to having a Black respondent instead of a White respondent. And since our dependent variable is binary, we should think in terms of probabilities, which can be converted to percentages: a .50-unit increase in vote_1 means a 50 percentage-point increase in the probability of voting for Obama. So putting this altogether, we'd say: (non-Hispanic) Black voters are 50 percentage points more likely to vote for Obama than (non-Hispanic) White voters, according to this model.

Similarly, Asian voters are 21 percentage points more likely to vote for Obama than (non-Hispanic) White voters. Native American voters are 10 percentage points more likely to vote for Obama than (non-Hispanic) White voters. Hispanic voters are 31 percentage points more likely to vote for Obama than non-Hispanic White voters. And voters identifying as multiracial or other race are 10 percentage points more likely to vote for Obama than (non-Hispanic) White voters. All of these differences are statistically significant, except for Native American versus White voters (probably because there are only 27 Native Americans in the sample, making the estimate of this difference very imprecise).

Let's move onto running a regression for the second category of our dependent variable:

```
. reg vote_2 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     72.35
       Model |  78.6117037         5  15.7223407   Prob > F        =    0.0000
    Residual |  658.463395     3,030  .217314652   R-squared       =    0.1067
-------------+----------------------------------   Adj R-squared   =    0.1052
       Total |  737.075099     3,035  .242858352   Root MSE        =    .46617

------------------------------------------------------------------------------
      vote_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   -.483031   .0280207   -17.24   0.000    -.5379725   -.4280895
      race_3 |  -.2002027   .0540143    -3.71   0.000     -.306111   -.0942944
```

```
      race_4 |  -.0822373   .1349253    -0.61   0.542    -.3467917     .182317
      race_5 |  -.3014791   .0320617    -9.40   0.000     -.364344   -.2386142
      race_6 |  -.1344972   .0440105    -3.06   0.002    -.2207906   -.0482038
       _cons |    .498904   .0097607    51.11   0.000     .4797657    .5180423
```

Now we're looking at predictions of voting for Mitt Romney. Our constant is .50, indicating that a non-Hispanic White voter has a 50% chance of voting for Mitt Romney. The coefficient of -.48 for race_2 indicates that (non-Hispanic) Black voters are 48 percentage points less likely to vote for Mitt Romney than (non-Hispanic) White voters. I won't go on to interpret the rest of the coefficients, but they follow the same pattern.

Finally, let's look at a regression with vote_3 as the dependent variable:

```
. reg vote_3 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |  -.0142558   .0082213    -1.73   0.083    -.0303757    .0018642
      race_3 |   -.007618   .0158479    -0.48   0.631    -.0386917    .0234557
      race_4 |   -.020605   .0395873    -0.52   0.603    -.0982258    .0570158
      race_5 |  -.0120213    .009407    -1.28   0.201     -.030466    .0064234
      race_6 |   .0302425   .0129128     2.34   0.019     .0049238    .0555611
       _cons |    .020605   .0028638     7.19   0.000     .0149898    .0262202
```

This regression provides some insights into who supported third-party candidates in the 2012 election. First, our constant indicates that a non-Hispanic White voter has a 2% chance of voting third-party. (Non-Hispanic) Black voters are one percentage point less likely to vote third-party than White voters, although this difference is only significant at the .10 level. The only other significant slope coefficient is for race_6, where we see that people who identify as multiracial or other race are estimated to be three percentage points more likely to vote third-party than (non-Hispanic) White respondents.

One final thing I want to show you is that our results will be in a slightly different format but will be in one sense equivalent if we decide to use a different category as our omitted category

when using a qualitative independent variable. Let's say we want to make Black (race_2) our reference category. Compare the following results to the previous regression:

```
. reg vote_3 race_1 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_1 |   .0142558   .0082213     1.73   0.083    -.0018642    .0303757
      race_3 |   .0066378    .017388     0.38   0.703    -.0274557    .0407313
      race_4 |  -.0063492   .0402287    -0.16   0.875    -.0852274     .072529
      race_5 |   .0022345   .0118186     0.19   0.850    -.0209387    .0254077
      race_6 |   .0444983   .0147623     3.01   0.003      .015553    .0734435
       _cons |   .0063492   .0077064     0.82   0.410    -.0087611    .0214595
```

Now, our constant tells us that a Black voter has a .6% chance of voting third-party. This is the same prediction we would get from our prior model where race_1 was the omitted category: to find our prediction for Black voters from the prior results, we would have added the coefficient for race_2 (-.014) to the constant (.021), yielding .006 or .6% (or .007 if we use the rounded numbers shown in parentheses).

The coefficient for race_1 tells us about how White voters differ from Black voters. Notice that the p-value is exactly the same as what we saw in the prior table for race_2, and the coefficient for race_1 in this table is the same as the coefficient for race_2 in the prior table, except the sign has changed. That's because comparing White to Black is the same as comparing Black to White, except that we're going in the opposite direction.

If you download the data yourself and have access to statistical software, you can go on to play around with these two sets of results more on your own if you'd like. Both regression equations will yield the same prediction for a voter of any given race. The difference lies only in the starting point, as represented by the constant. However, the p-values will usually differ because they are describing a different comparison (e.g., comparing Asian to Black in this table versus comparing Asian to White in the prior table). Thus, it doesn't really matter which category you pick as your omitted category, except that you may care more about some comparisons than others. You can also run the same regression multiple times but with different omitted categories so that you can get the p-values for a full set of comparisons across groups.

Part II

Estimation

5 Statistical Inference

In Chapters 1-4, we learned how to describe patterns in our data. In this chapter, we introduce a category of statistical tools used for something more ambitious. **Inferential statistics** allow us to use patterns in our data to draw conclusions about things not directly appearing in our dataset. For example, we might try to answer questions about a **counterfactual**: what outcomes would we see for participants in a program if they had instead not participated? Or we might try to answer questions about a broader set of data to which we do not have access: based on the 100 clients who responded to my satisfaction survey, what can I conclude about satisfaction among all clients of my organization?[1]

Example: Analyzing Data from a Randomized Experiment

Kim et. al (2025) study the experience of people trying to access food assistance benefits in the U.S. One requirement of the assistance program is to complete an interview. Working together with a nonprofit organization and a local government, the researchers designed an experiment to study whether interview completion rates could be improved by sending text messages to remind applicants of upcoming appointments and inform them of a new drop-in interview option.

For the experiment, some study participants were assigned to be part of a **control** group and thus received only a traditional mailer and no text reminders. The rest of the participants received the **treatment**, consisting of text reminders that were sent in addition to the mailer. Assignment to treatment and control groups was based on each participant's randomly-assigned case number: odd case numbers received the treatment while even case numbers received the control.

Using inferential statistics, the authors estimate how the text reminders affect the chances of completing an interview. The exact effect may vary depending on the study participant. We can game this out by imagining several distinct categories of applicants. Some applicants receiving the reminders would have completed the interview anyway, so for them the reminders had no effect. For others, the reminders were pivotal, causing them to attend an interview instead of missing it. And some missed the interview even with reminders (and possibly some *because of* the reminders, but I assume this is unlikely). Among those who did not receive the reminders, there were those who completed the interview anyway and for whom the reminders presumably would have had no effect (unless reminders somehow discouraged completion of the interview). There were also

[1]This chapter was written by Nathan Favero.

those who did not complete the interview but would have had they received the texts. Finally, there are those who did not complete the interview and wouldn't have even with reminders.

In a perfect world, we might like to know whether and how the text reminders would have changed the behavior of each person in the study. This is impossible to determine, however, since we can't ever observe the counterfactual to what actually occurred. The best we can do is estimate the likely effect of the text reminders in the aggregate—an *average treatment effect*. Based on the fact that this study's sample of participants receiving text reminders completed interviews at a rate 10.7 percentage points higher than those who received no texts, the researchers estimate that the text reminders increase the rate of interview completion by 10.7 percentage points. In other words, for 10.7% of people in the control group, it is estimated that text reminders could have caused them to complete an interview that they missed (again assuming that no one is discouraged from completing the interview based on the reminders).

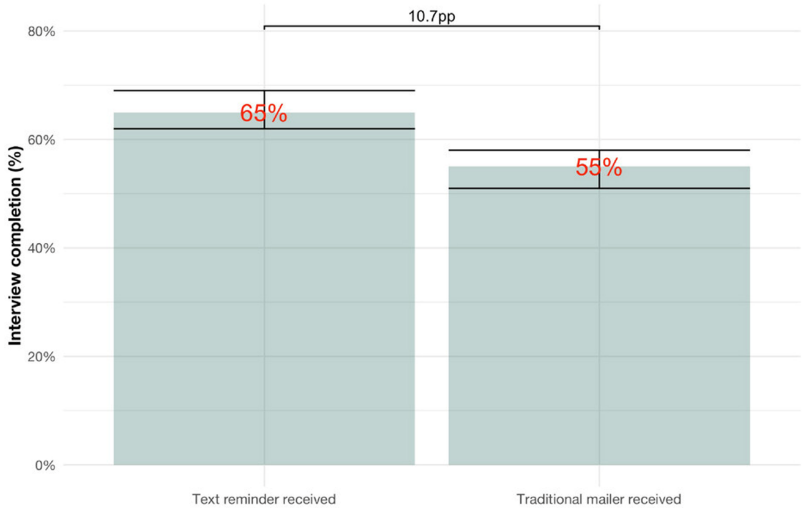

Figure 5.1: Difference in interview completion rate for treatment versus control groups, originally Figure 2 in Kim et. al (CC-BY 4.0).

Will this estimate be exactly correct? Almost certainly not. There is always noisiness in the real world that will partially distord our estimates. For example, by dumb luck, the participants in the treatment group may have had more family support on average, making it more likely that they would show up to their appointments even without the reminders. If so, the researchers would have overestimated the positive effect of the reminders. However, it is just as likely that the opposite is true: the control group may have had more favorable conditions for attending their appointments, causing the estimated effect to be smaller than the true effect. Because assignment to treatment and control groups was random, we can assume that there are not any systematic differences

between the people in the two groups. But there will still be random mismatches, meaning that estimates will almost always be wrong by at least a bit. We have no idea in which direction the estimate is off, however, and 10.7 percentage points is our best guess for what the average treatment effect was.

Because we are trying to make claims about data we do not directly observe, we are doing **estimation** when we use inferential statistics. We also have to make assumptions about how the data in our dataset were generated. Because assumptions usually cannot be (fully) tested and could end up being incorrect, there is always a risk with statistical inference that we will draw invalid conclusions. In the example above, a key assumption is that the assignment of individuals into the treatment and control groups is truly random. This assumption could be violated if there was an error such that participants with a greater ability to make a scheduled interview were systematically more likely to receive an odd case number, for example. Fortunately, it is difficult to imagine how such a violation could have occurred in practice, unless there was deliberate data manipulation by researchers or program personnel. More broadly, it is hard to imagine a system of case ID assignment by which those receiving even-numbered case IDs would be systematically different in important ways from those receiving odd-numbered case IDs, so we can feel fairly confident about meeting the assumption of random assignment in the example above. For other studies, the assumption of random assignment for an experiment may be more questionable, such as when there are concerns that participants may have been able to tamper with their own assignment or when subjects drop out mid-study and thus cannot be included in the final analysis. Assumptions of statistical models will be discussed further in future chapters.

Another important feature of inferential statistics is that it involves drawing conclusions that are expressed probabilistically. We generally cannot draw 100% definitive conclusions, but we can draw conclusions based on confidence thresholds. In most scientific work, the default confidence threshold is 95%, meaning that we are willing to accept a process for drawing conclusions with a known 5% error rate. We will see this concept demonstrated later in this chapter when discussion confidence intervals.

5.1 Key Concepts for Statistical Inference

In the context of inferential statistics, the **sample** refers to the actual units we observe (the observations in our dataset). Traditionally, this is contrasted with the **population**, meaning the full universe of units we are interested in learning about. The sample is therefore a subset of the population. If every unit in the population is included in our dataset, we have a **census**, not a sample, that we are analyzing.

5.1.1 Types of samples

Statistical models used for making inferences about a population based on a sample will almost always assume that probability sampling was used. **Probability sampling** means that the sample is made up of units that were randomly selected from the population. For example, governments may collect survey data from samples of randomly selected participants of public programs (e.g., workforce training). In a **simple random sample**, all units in the population have an equal probability of being selected. There are also more elaborate techniques, such as **stratified random sampling** whereby the population is subdivided into various "strata" (e.g., by age and gender) and then units are randomly sampled within strata.

In practice, it is often impossible to obtain a sample that is perfectly random. Even if there is a current list of everyone in a population (e.g., a client list for an organization) allowing for random selection of names to contact, there will almost always be people who refuse to provide the information required to assemble a dataset (e.g., failing to respond to a survey). One simple measure of this phenomenon commonly employed in survey research is the **response rate** or percentage of people invited to a survey who actually respond. For polling firms and others studying public opinion, persistently falling response rates over the past decades have made reliably measuring public opinion more difficult. Governments are often able to obtain higher response rates than private entities, but they too have struggled with declining willingness by the public to participate in data collection in many contexts.

Given these challenges, it is perhaps no surprise that many datasets consist of **convenience samples**, whereby units are included in the sample because of convenience rather than random selection. Perhaps the most credible type of convenience sampling is **representative sampling**, which refers to techniques (such as demographic quotas) used to ensure that certain sample characteristics will resemble the known demographics of the broader population. Users of this approach hope that matching on certain known characteristics will lead the sample to resemble the population on other characteristics where the distribution in the population is unknown, but there is no guarantee that this will be the case.

The risk for most analysts trying to draw conclusions about a population based on a sample is that they rarely have the type of purely random sample assumed in statical models used for inference. Thus, it is unclear how well the probabilistic conclusions drawn through inference methods will actually map to real-world accuracy of estimation. Still, inferential statistics provide a starting point by examining what conclusions can be drawn under the best case scenario of a sample that was perfectly random. A savvy analyst will also take into account context and any available information about the sample selection process to help inform what conclusions can reasonably be drawn from the data at hand. The more closely the sampling process appears to resemble the random selection process we assume in our statistical models, the more confidence we can have that the conclusions of our statistical models properly describe the uncertainty of our estimates.

5.1.2 Counterfactuals

We've already seen how statistical inference can be used to draw conclusions about counterfactuals, but a more precise explanation of terminology is provided here. A **counterfactual** is a hypothetical alternative to what actually occurred, where one or more independent variables takes on a different value. For example, we have seen how a treatment effect can be estimated, which represents the expected change in the dependent variable if an independent variable representing the experimental group changes from "control" to "treatment."

5.1.3 Parameters of interest

Whether we are trying to draw conclusions about the population or a counterfactual, the quantity that we are estimating is called a **parameter**. For example, the parameter might be the population mean or the average treatment effect. By contrast, a **sample statistic** directly describes the sample itself and is often used as an estimate of a parameter. For example, if the parameter of interest is the population mean, the sample mean can be our estimate of the parameter.

5.1.4 Importance of sample size

All else equal, larger samples are better. This is fairly intuitive: we are better able to estimate the attitudes of a country's population with a sample of 2000 people than with 20, all else equal.[2] Larger samples allow us to draw more precise conclusions. Since we typically assume that the imprecision of our estimates is rooted in the random and unpredictable peculiarities of individual units in our sample, a larger sample will increase our precision because the individual-level idiosyncrasies begin to matter less to the overall sample. With a sample of just 20 randomly selected people, it is not at all uncommon to end up with a very unrepresentative sample, such as one with mostly men and very few women. With a random sample of 2000, it would be extremely unlikely (though not technically impossible) to get a gender balance in the sample that deviates substantially from that of the population. One reason the tools of inferential statistics are so useful is that they can precisely show us how much the precision of an estimate increases as the sample size increases (given the assumptions of the model).

5.2 Confidence Intervals: A Key Tool for Estimation

Returning to the example at the beginning of the chapter, we saw that researchers found that applicants for a food assistance program who received text reminders were 10.7 percentage points more likely to complete an interview than those in the control group who received no

[2] Increasing sample size at the expense of other design considerations is not always wise. For example, a very large convenience sample is not necessarily better than a smaller probability sample.

text reminders. This difference in the completion rate between the treatment and control groups (the sample statistic) serves as a **point estimate** for the parameter (effect of the text reminders). As we noted, this estimate is likely to be imperfect because of random differences in who gets assigned to the treatment versus control groups. To better convey the uncertainty surrounding a point estimate, it is usually helpful to provide an **interval estimate**. Interval estimates identify a range of likely values rather than a single number representing one's best guess. For example, the researchers studying text reminders report a 95% confidence interval ranging from 5.8 to 15.5 percentage points. Thus, there is good reason to believe that the true effect size is at least 5.8 percentage points but no greater than 15.5 percentage points:

$$5.8 < \text{effect size} < 15.5$$

The range of values contained within the interval is identified by its **lower bound** (5.8 in this example) and **upper bound** (15.5). Any number between the lower and upper bounds is within the interval, so effect sizes of 6, 9, and 13 are all reasonable candidates for the true effect size (as are any other numbers within the interval).

Creating an interval estimate requires that we select a confidence level. With a 95% confidence level, we calculate an interval using formulas calibrated such that the interval should contain the true value 95% of the time, assuming all assumptions of the statistical model are met. On the flip side, 5% of the time, the interval will not contain the true value.

This property of 95% confidence intervals is demonstrated in Figure 5.2,[3] which shows the results of a simulation in which random samples were drawn again and again, to see how interval estimates behave. Suppose the true effect size for a treatment is zero and each vertical bar represents the confidence interval resulting from a different random assignment of participants to treatment and control groups. As you can see, most intervals overlap with the true parameter value of zero (on the y axis), but we also expect one out of every 20 estimates (5%) to miss the mark (the bars shown in blue). In practice, we would normally only have one sample and wouldn't know for sure if ours is one of the unlucky cases in which the confidence interval fails to include the true effect. What we do know is that we've followed a process that theoretically gives a correct answer 19 times out of 20.

The 95% confidence level is perhaps most common, but we also frequently encounter confidence levels of 90% and 99%. Whatever confidence level we choose, we are accepting that our procedure will yield a false answer some of the time. With a 95% confidence level we expect a 5% error rate, while a 99% confidence level is associated with an error rate of only 1%. There is always a tradeoff when choosing a confidence level. A more conservative level (e.g., 99%) will yield a wider interval, meaning that we are accepting a wider range of values as possible in order to obtain this higher level of confidence that we will avoid an error.

[3]This image is shared under CC-BY 4.0 and is adapted from Figure 1 of Nalborczyk, L., Bürkner, P. C., & Williams, D. R. (2019). Pragmatism should not be a substitute for statistical literacy, a commentary on Albers, Kiers, and van Ravenzwaaij (2018). *Collabra: Psychology*, *5*(1), 13. https://doi.org/10.1525/collabra.197

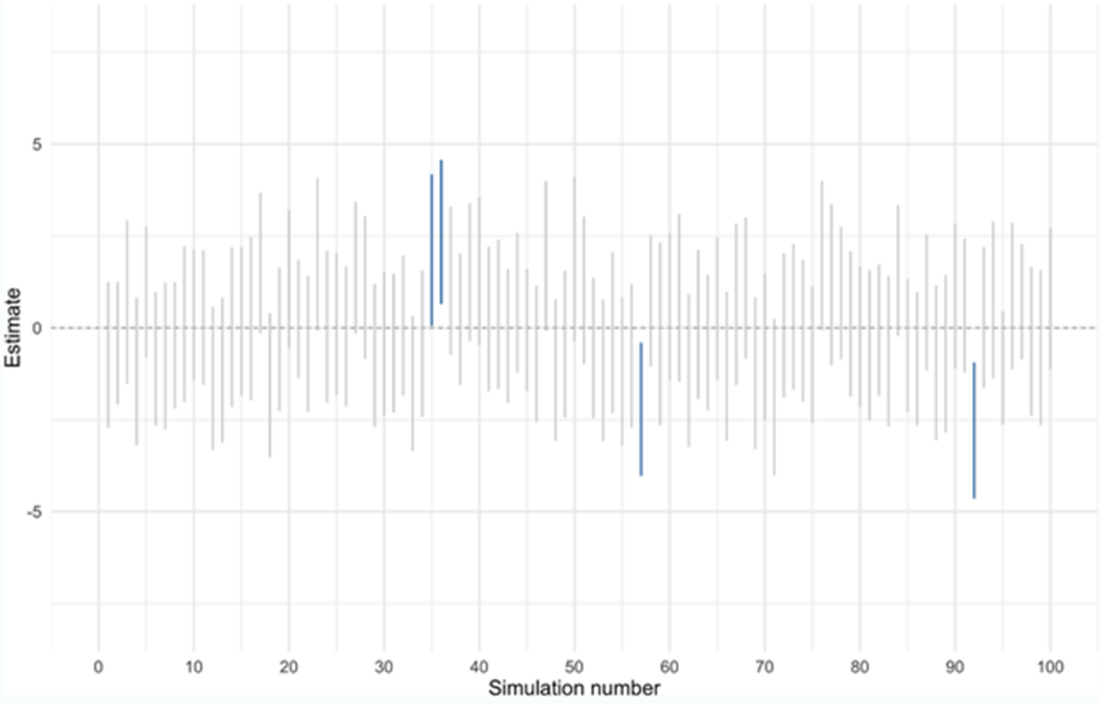

Figure 5.2: Simulation of interval estimates based on different random samples, adapted from Nalborczyk et al. (CC-BY 4.0).

What procedure is used to construct a confidence interval? We will learn the details in Chapter 7. For now, it is enough to know that there are well-established procedures for constructing confidence intervals in various settings, based on assumptions about the data. Even without knowing these exact procedures, you can hopefully begin to see the usefulness of confidence intervals from the examples in this chapter.

5.2.1 Confidence Intervals for Regression

To examine use of confidence intervals for regression, we will return to the example of predicting university grades based on an admissions exam (Section 3.5). The table is shown again here as Table 5.1.

Table 5.1: Results for a regression with computer science GPA as the dependent variable.

	Coef.	Std. err.	p-value
verb_sat	0.0017	0.0010	0.10
math_sat	0.0048	0.0012	0.00014
(intercept)	-0.91	0.42	0.033
n	105		
r^2	0.487		

Because the coefficients we see in the table are just point estimates, confidence intervals can help us better understand the precision of these estimates by providing us with a range of plausible values for the coefficients. Notice that no confidence intervals have been provided in the table (which is a situation you may frequently encounter when reading social scientific research publications). Fortunately, we can easily calculate a good approximation of a confidence interval for a coefficient estimate $\hat{\beta}_i$ as long as we also have its standard error estimate (s_{β_i}), which is provided in the table (in the column label "Std. err.").[4] We will learn exactly what a standard error is in Chapter 7, but for now, we can simply insert the standard error estimate into the following formulas:

$$\text{Lower bound} \approx \hat{\beta}_i - 2 \times s_{\beta_i}$$

$$\text{Upper bound} \approx \hat{\beta}_i + 2 \times s_{\beta_i}$$

Note that these formulas are just an approximation for a 95% confidence interval; the formulas for precise intervals are shown in Section 7.3.3. Multiplying the standard error by two yields

[4]Some publications list t scores or z scores instead of standard errors. These t (or z) scores are typically just the coefficient divided by the standard error estimate, so you can obtain the standard error estimate by dividing the coefficient ($\hat{\beta}_i$) by its t score (t_{β_i}): $s_{\beta_i} = \hat{\beta}_i / t_{\beta_i}$.

the approximate margin of error.[5] From our initial point estimate of the slope, we can then add or subtract the margin of error to identify a full range of plausible values.

Our approximation approach provides an inexact but close approximation of a 95% confidence interval as long as the sample size is reasonably large (e.g., at least 30 more observations than the number of independent variables included in the regression). In this case, there are 105 observations and only two independent variables, so we will obtain a good approximation. And an approximation is usually the best we can hope for when calculating confidence intervals by hand from a regression table, since we will usually also lose some precision due to rounding error.

For the verbal SAT scores, we find the following bounds:

$$\text{Lower bound} \approx 0.0017 - (2)(0.0010) = -0.0003$$

$$\text{Upper bound} \approx 0.0017 + (2)(0.0010) = 0.0037$$

This is very close to the precise 95% confidence interval that one finds using statistical software to compute an exact interval: [-0.0003, 0.0038]. Any values within this range can be considered plausible values for the coefficient, according to our model results.

How do we interpret this confidence interval? Remember that when it comes to interpreting size, the coefficient indicates how many units the prediction for the dependent variable changes when the independent variable increases by one unit. But with SAT scores, a 1-unit increase is so small that we found it more useful to consider a 100-point increase, which required multiplying the coefficient by 100. Doing so here, we can conclude that a 100-point increase in the verbal SAT score (e.g., comparing a student with a 600 to a student with a 500, assuming math SAT scores are equal) predicts a difference in the computer science GPA somewhere in the range of [-0.03, 0.38]. Zero is part of this range, so it's entirely plausible that there is no real association between verbal SAT score and computer science GPA (hence, the lack of statistical significance for this relationship). According to the model results, it is also plausible that a 100-point increase in verbal SAT is associated with the predicted computer science GPA decreasing by as much as 0.03, or increasing by as much as 0.38. A 0.03 decrease in GPA is tiny, so we might feel comfortable ruling out the possibility that a good verbal SAT score has any substantial negative predictive effect for computer science GPA. But a positive effect of 0.38 grade points is much more substantial, so it is plausible that the verbal SAT has a meaningfully-large positive association with computer science GPA.

What about the math SAT? Using our approximation method:

$$\text{Lower bound} \approx 0.0048 - (2)(0.0012) = 0.0024$$

[5] Two is the approximate value by which the standard error should be multiplied, but a more exact value can be found using the t distribution, as explained in Section 7.3.

$$\text{Upper bound} \approx 0.0048 + (2)(0.0012) = 0.0072$$

Multiplying these two values by 100, we find that a 100-point increase in the math SAT is plausibly associated with an increase of between 0.24 and 0.72 points in predicted computer science GPA.

5.2.2 Interpreting Confidence Intervals Correctly

Suppose we want to estimate the average poverty rate for a city based on a sample of residents who have completed a survey. We might calculate a confidence interval and obtain the values [13.4%, 15.1%]. A common mistake made by students and scientists alike when describing this 95% confidence interval is to say "this means there is a 95% chance that the true poverty rate is between 13.4% and 15.1%." One reason this interpretation is too simplistic is that there may be other relevant evidence about the poverty rate beyond the data used to compute our confidence interval. If several other high-quality studies of the city have recently estimated the poverty rate and produced results in the range of 17-20%, we would probably conclude that our own interval estimate has a good chance of being wrong (much greater than 5%).

So what is the correct interpretation of a confidence interval? You can make the following statement any time you encounter a 95% confidence interval (of the form [A, B]):

> Using a process with 95% accuracy (in theory), it is estimated that the parameter lies between A and B.

I realize this interpretation is a bit indirect; it is difficult to provide a technically-accurate and meaningful interpretation, despite the fact that confidence intervals have demonstrated great practical value to researchers and analysts.[6] What makes interpretation difficult is the fact that the "% confidence" in a "95% confidence interval" refers to the accuracy of the *process* of creating a confidence interval—not the probability that a specific confidence interval we encounter will contain the true population parameter. If this distinction seems confusing, it is!

Fortunately, even if you miss the precise details, you will still probably get something useful out of confidence intervals.[7] Nonetheless, let's try to set the record straight.

An analogy may help. Suppose you are interacting with a chatbot that is truthful 95% of the time and lies the other 5%.[8] For each statement, will you always conclude it has a 95%

[6] Stephens, M. (2023). The Bayesian lens and Bayesian blinkers. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220144.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *26*(1), 1.

[7] Anderson, A. A. (2019). Assessing statistical results: magnitude, precision, and model uncertainty. *The American Statistician*, *73*(sup1), 118-121.

[8] This example is adapted from Behar, R., Grima, P., & Marco-Almagro, L. (2013). Twenty-five analogies for explaining statistical concepts. *The American Statistician*, *67*(1), 44-48.

chance of being true? Not necessarily. If the chatbot discusses a topic you already know a lot about, you will probably be able to pick out the lies from the true statements with fairly high confidence. Some things the bot says will be things you know to be true, so you can be nearly 100% sure they are true. Other statements will be things you're quite sure are wrong, so you will conclude that the probability they are true is close to 0%. If you wanted to be very systematic, you could even use the mathematical formula known as Bayes' theorem[9] to combine your prior knowledge of a statement's probability of being true with the fact that a 95%-accurate bot claimed the statement was true, allowing you to precisely quantify how confident you should be about the statement's truth in the end.

Now imagine you ask this same bot to start telling you about a topic you know nothing about. Absent any prior insights into which statements are likely to be true, it would now be reasonable to conclude that each statement the bot makes has a 95% chance of being true.

In the same way, it turns out that *absent any other information,* a 95% confidence interval is often a good approximation for a range of values that contains the population parameter with 95% probability.[10] Thus, I think it is quite reasonable that many of us, when we see a mean estimate with a 95% confidence interval ranging from A to B, assume there is a 95% chance the population mean does indeed lie between A and B. But technically, that is not a direct interpretation of the confidence interval; instead, this statement about plausible values of the population mean is a subjective conclusion that I can draw based on the confidence interval. Another person might see the same confidence interval and reasonably decide—drawing on their own prior knowledge of the topic—that the confidence interval contains values that are highly implausible, and thus they would reach a different conclusion from me about how likely the interval is to contain the true population mean.

If you want to elaborate on how the 95% confidence interval [A, B] can inform our practical understanding, you might add the following to our earlier interpretation:

> Assuming no additional information and an appropriate statistical model, this result usually suggests that we can be about 95% confident the parameter lies between A and B.

[9] See https://onlinestatbook.com/2/glossary/bayes.html or https://onlinestatbook.com/2/probability/bayes_demo.html.

[10] Kass, R. E. (2011). Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *26*(1), 1.

Albers, C. J., Kiers, H. A., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 31.

Greenland, S., & Poole, C. (2013). Living with p values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, *24*(1), 62-68.

6 Probabilistic models

Policy makers, managers, and researchers all face uncertainty in the world around them. One way we can describe uncertainty about a particular matter is to list out alternative possibilities for what could happen and then describe how likely we think each alternative is to occur. In doing so, we would essentially be creating a probability distribution.[1]

6.1 Probability distributions

Probability distributions are precise descriptions of all possible values that could be obtained from a random process as well as the probability of each value occurring. One of the simplest probability distributions describes a coin flip. The two possible values are "heads" and "tails," and each value has a 50% chance—or .5 probability—of occurring. Probability distributions are very useful in statistics because they allow us to build statistical models that account for randomness or uncertainty.

> Example: Anticipating How Many Visitors to Expect
>
> Suppose you run an organization and want to plan for how many customers to expect the next day. If you are operating in a stable environment, you might be able to use data on past customer volume to make reasonable predictions about the future.
> To demonstrate how past data might help predict the future, we will examine some data on hospital emergency room (ER) visits. This data comes from 2018 and indicates the number of daily visits across 30 different hospitals in Seoul, South Korea.

[1] This chapter was written by Nathan Favero.

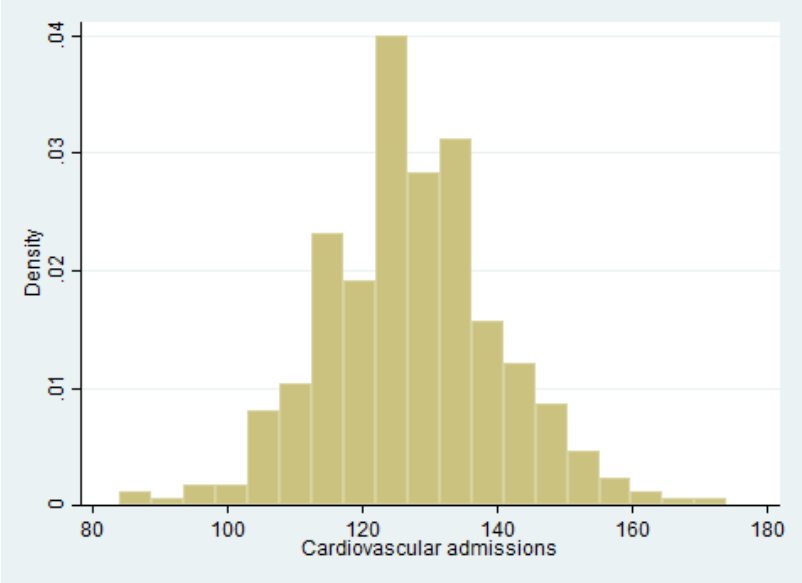

Figure 6.1: Histogram of daily volume of cardiovascular ER admissions; data from Hwang et al. (CC-BY 4.0).

When making a prediction for how many cardiovascular admissions there will be tomorrow, we could provide a guess in the form of a single number. In that case, we would probably want to pick a number that appears to be near the center of the distribution, such as 130. However, if we want to give a more sophisticated (and comprehensive) prediction, we could list out many possible values and indicate each one's probability of occurring. In doing so, we would be creating a probability distribution. We could present this information either graphically or in a table. One very simple approach to creating a probability distribution based on past experience would be to use the exact proportion of times each value was observed in the past year as the probability we assign to that value for the future. If we recreate Figure 6.1 with a bin width of 1 so that each bar displays only a single value, we get a precise visual depiction of this distribution.

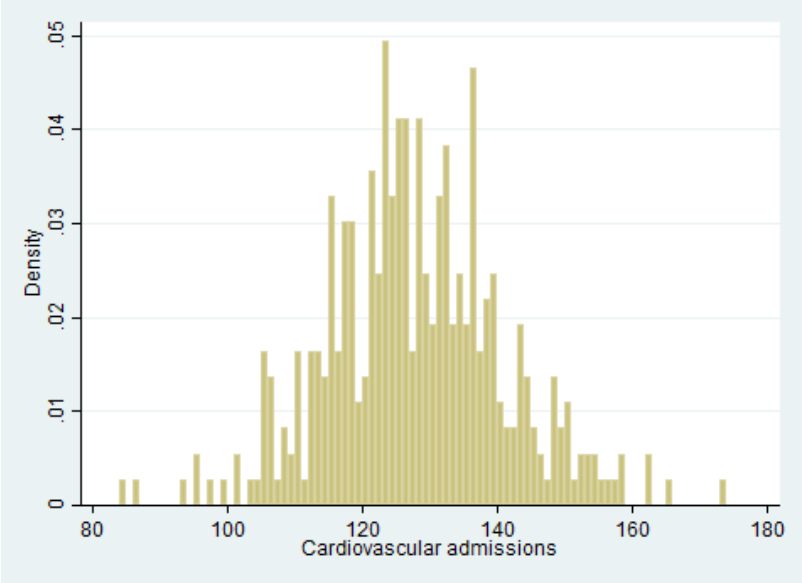

Figure 6.2: Histogram of daily cardiovascular ER admissions with a bin width of 1

From Figure 6.2, we might begin to spot some problems with the simple approach of using last year's proportions as direct probabilities for forecasting. The data looks a bit "choppy," with some bars noticeably taller or shorter than those surrounding them. In the tails, the issue is particularly pronounced. For example, some values (e.g., 91, 92) were never observed in the last year, but that doesn't necessarily mean they have 0 probability of occurring in the future.

A savvier approach to making a prediction might be to take the general shape of this distribution of past occurrences but then smooth it out since we have no reason to believe that the probability of one number should differ much from the numbers immediately around it. One way to accomplish this is to use one of the many well-known distributions that statisticians have developed to describe data generated under various assumptions. We can pick a distribution that resembles the general shape we see here. Let's try a normal distribution,[2] since it is the most widely-used distribution.

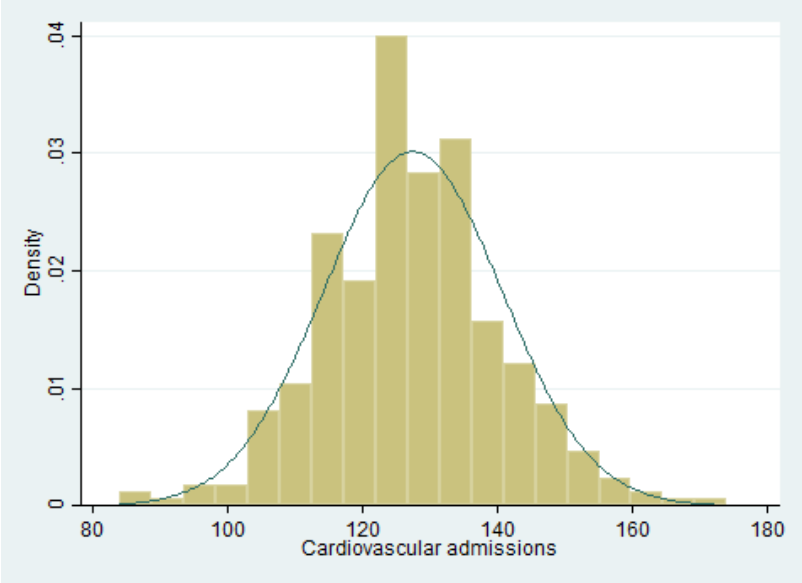

Figure 6.3: Normal distribution (solid line) overlaid on histogram of daily cardiovascular ER admissions

We can see in Figure 6.3 how the normal distribution provides a smooth curve that generally matches the shape of the distribution of observations from 2018. The height of the normal distribution indicates the probability associated with draws from that area of the curve, so values between 120 and 140 are more likely to occur than values between 140 and 160, according to this probability distribution. There are ways we could further refine our prediction for future visits, such as by differentiating among days of the week (more cardiovascular ER admissions occur on Mondays and Fridays) and accounting for the fact that we probably observe more outliers than the normal distribution assumes. (Technically, a Poisson distribution would also be more appropriate for this data since the number of admissions will always be a whole number, and the normal distribution is more appropriate for situations where we measure a variable to many digits.) Still, you can see how the normal distribution could provide a reasonable basis for making predictions about the relative probability of observing different values for the number of visitors in the future. More generally, this example demonstrates how a probability distribution can be used to think about possible values that may occur for a variable.

We learned in Chapter 2 that a distribution describes how frequently every possible value for a variable occurs in a dataset. In contrast to this sort of distribution, a *probability* distribution doesn't describe data we've collected; instead, probability distributions are used to describe a (theoretical) *process* and indicate how likely we are to obtain different possible values from this

[2]Specifically, we use a normal distribution with mean and standard deviation set equal to the mean and standard deviation of the 2018 sample we're examining.

random process.

This distinction between a distribution of data versus a probability distribution is subtle but important. For example, if we write out the (theoretical) probabilities for each possible outcome of a die roll (1, 2, 3, etc.), we're talking about a probability distribution. If we roll a die 20 times and record the results, we're looking at a distribution of data. Because there is plenty of variability from one player's die rolls to the next's (some will be luckier than others), the distribution of one player's results will not necessarily provide a close match to the theoretical probabilities associated with a die roll (where each of the six values on a six-sided die have a probability of 1/6 each).

Another way to think about this distinction is that for a distribution of data we've collected, each observation was (theoretically) drawn from the probability distribution. In our example from the box above, if we have a distribution that describes how likely we think it is that we get various numbers of patients visiting the ER on a given day, we're looking at a probability distribution. If we're looking at data (e.g., a histogram) of past daily totals for the number of patients who visited the ER, we're looking at a distribution of data (not a probability distribution).

Many of the same statistics and words we use to describe data that's been collected can also be used (with a bit of adaptation) to describe probability distributions. For example, a probability distribution will (often) have a mean (also called the expected value) and a variance. A probability distribution can be skewed or symmetric. It can be bimodal or unimodal.

Probability distributions can be described using (usually complex) equations, but in this text, we'll mainly depict probability distributions graphically. We generally depict complex (continuous) distributions by plotting what is called a probability density function (PDF). The normal distribution depicted in the box above for ER visits is an example of a PDF. Statisticians have developed PDFs for distributions with many different shapes.

You can basically interpret a graph of a PDF like a kernel density plot (or even a histogram): for values where the graph is taller (indicating greater "density"), those values are more likely to occur. But kernel density plots and histograms are used for data that we've already collected; PDF graphs depict a probability distribution from which data can be theoretically "drawn" (you can't necessarily tell the difference between a kernel density plot and a PDF plot from just looking at the graph). A precise interpretation for PDFs is a bit tricky since any value can be measured to infinite digits (and therefore has infinitely small probability of being selected) when we are discussing continuous distributions like the normal distribution. The total area under a PDF will always equals 1. To calculate actual probabilities, we need to identify a range of values: for example, we can use software (or various online calculators) to determine the probability of drawing a value between 0 and 0.7 for a well-known continuous distribution. This probability will be equal to the area under the line within that range (which can be found using calculus). Our statistics software will often be relying on calculus behind the scenes based on these PDFs to provide relevant statistical results.

A **random variable** is a term used to describe a variable generated through draws from a probability distribution.

6.1.1 Normal distributions

Normal distributions are one of the most important sets of probability distributions we use in statistics. As can be seen in Figure 6.4, a normal distribution has a symmetrical shape called a bell curve. Many variables in nature appear to approximately follow a normal distribution. For example, if you measure the heights of a population of adult humans belonging to a single sex (male or female), the distribution should look similar to a normal distribution. We often encounter normal distributions because of a principle called the central limit theorem, which states that any variable that results from adding up many small, independent factors will approximately follow a normal distribution.

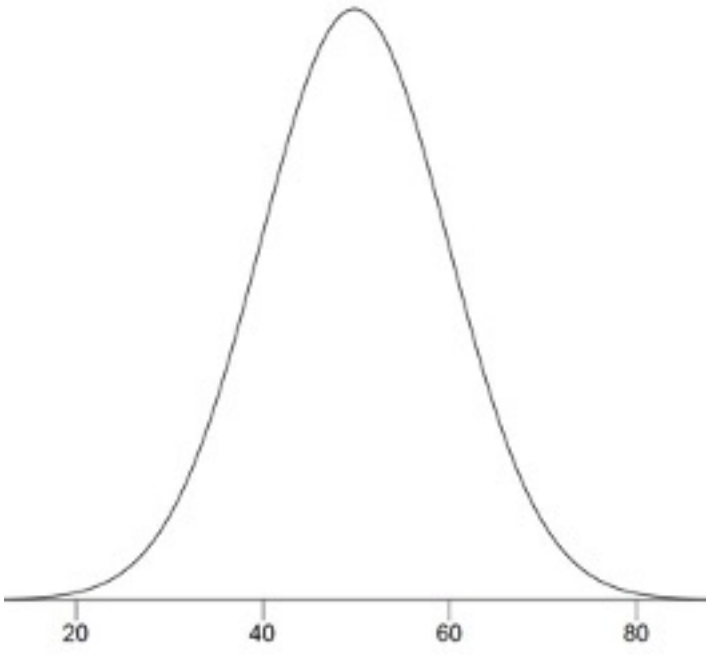

Figure 6.4: An example of the normal distribution

All normal distributions have the basic "bell curve" shape seen in Figure 6.4, but we can get different versions of the normal distribution by shifting this curve to the left or right, and by squishing or expanding the width of the curve. Figure 6.5 illustrates this by showing three different versions of the normal curve in one graph. Any normal distribution can be uniquely identified by its two parameters: the mean (μ) and standard deviation (σ). Because the normal curve is unimodal and symmetric, the mean, median, and mode are all equal to one another. Thus, the mean can be identified visually as the tallest point (the mode) on the curve.

Changing the mean will move the curve to the left or to the right; in Figure 6.5, the mean must be smallest for the green curve since it is furthest to the left (although no axis labels are provided). Changing the standard deviation will expand or narrow the width of the curve; the black curve in Figure 6.5 has the largest standard deviation since it is the widest curve. As noted above, all probability density functions have a total area under the curve equal to one, so the narrower versions of the normal distribution (e.g., the green one in Figure 6.5) have to be taller than the wider curves in order to maintain that same area of 1 under each curve.

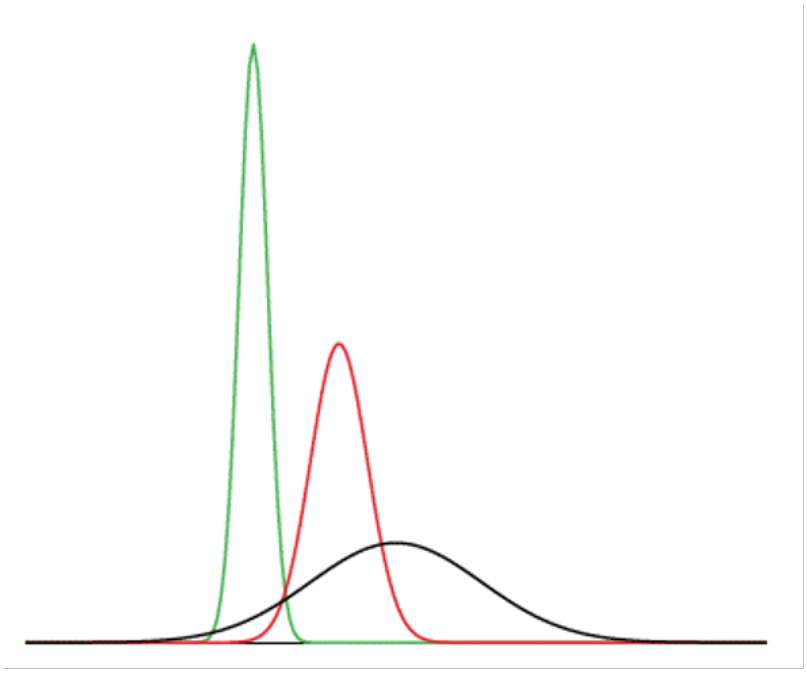

Figure 6.5: Three different examples of normal distributions

There are some numbers you can memorize to help describe the probabilities associated with any normal distribution:

- 68% of the area under a normal curve is within approximately *one* standard deviation of the mean
- 95% of the area under a normal curve is within approximately *two* standard deviations of the mean[3]
- 99.7% of the area under a normal curve is within approximately *three* standard deviations of the mean

Let's consider the normal distribution with a mean of 100 and a standard deviation of 20. To find the range describing one standard deviation from the mean, we first subtract the value of

[3]As we will see in Chapter 7, a more precise value is 1.96 standard deviations from the mean.

the standard deviation from the mean to find the lower bound (100 - 20 = 80), and then we add the standard deviation to the mean to find the upper bound (100 + 20 = 120). We can then say that there is a 68% chance that a draw from this distribution will yield a number between 80 and 120. Figure 6.6 shows this visually, with the blue highlighted region representing an area of 0.68 (compared to a total area of 1 under the entire cure).

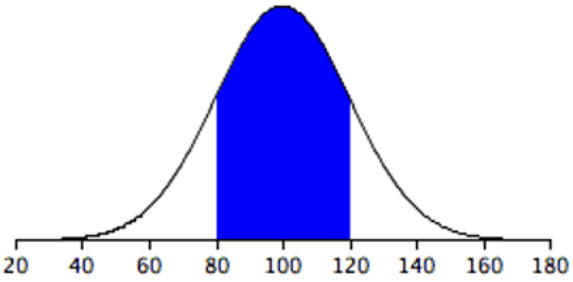

Figure 6.6: A normal distribution with mean of 100 and standard deviation of 20; the region within one standard deviation of the mean is highlighted

To consider two standard deviations from the mean, we would first multiply the standard deviation by two ($20 \times 2 = 40$). For the lower bound, we subtract this value from the mean (100 - 40 = 60), and for the upper bound we add (100 + 40 = 140). Thus, there is a 95% chance that a draw from this normal distribution yields a value between 60 and 140.

Using a normal distribution calculator (available online[4] or in statistical software), we can easily determine the areas for other ranges of values. For example, still looking at this same normal distribution (mean=100, standard deviation=20), we could discover that the area for values less than 65 is 0.04, meaning there is a 4% chance a draw will yield a value less than 65.

The ***standard* normal distribution** refers to a normal distribution with a mean of 0 and a standard deviation of 1. This name is related to the term *standardization* we encountered in Section 2.6.1; recall that standardizing a variable means transforming it so that its mean is 0 and its standard deviation is 1. The standard normal distribution, and other closely related distributions, are often utilized in statistical tests. Sometimes, variables or individual values are standardized (turned into Z scores) and then compared to the standard normal distribution.

6.2 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather app tells me the percent chance of rain for the day.

[4]For example: https://onlinestatbook.com/2/calculators/normal_dist.html

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Consider this simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{6.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: ε. This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term (ε) represents everything else besides income that affects happiness. It is a formal acknowledgement that if all we know about someone is their income, we will have uncertainty about their exact level of happiness. An error term (ε) in the model makes clear that we only claim to have a partial understanding of happiness, not a complete one.

The first part of our model that appears on the right side of the equation ($3.0 + 2.3 \times income$) is the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we knew about someone was their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be at least a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term (ε).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly drawn from a probability distribution.

6.2.1 Assumptions About Error Terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

As one example, we might assume the following things about the error term (ε):

1. The values of the error term (ε) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term (ε)

What do these two assumptions mean?

First, if the error term (ε) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term (ε), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term (ε) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two

standard deviations away from zero. Suppose that the standard deviation of the error term (ε) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term (ε) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood directly causes (on average) both higher incomes and greater happiness in adulthood,[5] the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty. This example also illustrates how problems of causality can often be conceptualized as violations of assumptions about the error term; in Chapter 10, we will see that we can label the problem posed for our analysis by the effects of childhood stability a "third-variable problem," but here we have shown how it can also be understood as problematic correlation between the dependent variable and the error term.

If you want to explore in more detail how equations can be used to describe statistical models, this chapter's appendix provides a more formal presentation of some ideas from this section.

6.2.2 Models and Probabilistic Thinking

Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to build only *deterministic* models, meaning models that are supposed to predict with 100% accuracy. Little (if anything) about the social world follows absolute laws, so deterministic models are arguably not well suited to social scientific study. Instead, the best we can hope for is a *probabilistic* model, indicating the conditions under which particular outcomes are more or less likely. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about

[5] By "directly cause" greater happiness in adulthood, I mean that a stable childhood environment causes greater adult happiness by means other than increasing income (which in turn may increase happiness).

the world in probabilistic terms, as we've done throughout this chapter. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?

Appendix: Expected Values and Conditional Probabilities

Let us now practice using equations to more formally describe some of what we discussed in the main chapter. Given our assumptions and the equation representing our model of happiness, we can use the notation of expected value to express the predictions we previously made:

$$\begin{aligned}\mathbb{E}[happiness|income = 4] &= \mathbb{E}[(3.0 + 2.3 \times (4) + \varepsilon)|income = 4] \\ &= 12.2 + \mathbb{E}[\varepsilon|income = 4] = 12.2\end{aligned}$$

We use $\mathbb{E}$ to indicate an expected value and the symbol | indicates "conditional on," meaning that we want to know the expected value of happiness conditional on income being equal to 4. Given that the error term (ε) is independent of income (our second assumption), $\mathbb{E}[\varepsilon|income = 4]$ simplifies to $\mathbb{E}[\varepsilon]$ and since ε is drawn from a normal distribution with a mean of 0, $\mathbb{E}[\varepsilon] = 0$.

We can also apply the notion of conditionals to probabilities. For example, we might want to say something about the probability of happiness being greater than some value. To make the math simpler, I will choose values that correspond to 0, 1, 2, or 3 standard deviations from the center of the distribution of ε since that will make it easy to do the math by hand using the proportions of the normal distribution we learned in the main part of this chapter. A normal distribution calculator could be used to find probabilities for other values (e.g., greater than 1.47 standard deviations above the mean).

First, let's consider the probability of happiness being greater than our prediction of 12.2 when income is 4:

$$\begin{aligned}Pr(happiness > 12.2|income = 4) &= Pr((3.0 + 2.3 \times (4) + \varepsilon) > 12.2) \\ &= Pr(12.2 + \varepsilon > 12.2) = Pr(\varepsilon > 0) = .5\end{aligned}$$

From any normal distribution, half of the area under the curve will be above the mean, while half of the area will be below the mean. Since we assume ε is drawn from a normal distribution

with a mean of 0, the probability of a value greater than 0 is .5. In other words, there is a 50% chance that the actual value of happiness will exceed the predicted value. Similarly, there is a 50% chance the true happiness will fall below the predicted value:

$$\begin{aligned} Pr(happiness < 12.2 | income = 4) &= Pr(12.2 + \varepsilon < 12.2) \\ &= Pr(\varepsilon < 0) = .5 \end{aligned}$$

Let us again assume for the moment that the standard deviation of the normal distribution from which ε is drawn is three (in real-world analysis, we can estimate the standard deviation of this normal distribution based on the data we observe). Given this value, we can now calculate other conditional probabilities. We can calculate the probability of happiness exceeding 15.2 when income is 4:

$$\begin{aligned} Pr(happiness > 15.2 | income = 4) &= Pr(12.2 + \varepsilon > 15.2) \\ &= Pr(\varepsilon > 3) = .16 \end{aligned}$$

For the final step, we rely on the fact that for any normal distribution, 68% of the area under the curve falls within one standard deviation of the mean. Remember, we assumed ε is drawn from a normal distribution with a standard deviation of three (and mean of 0), so there is a .68 probability of drawing a value between -3 and 3. Thus, the probability of drawing a value outside this range must be .32 ($1 - .68 = .32$). Half of this .32 will belong to the lower tail (values less than -3) and half to the upper tail (values greater than 3). Thus, the probability that ε is greater than 3 is .16.

7 Sampling Distributions

This is perhaps the most difficult chapter in the whole book. That is because sampling distributions are a very abstract concept that many students struggle to grasp. And yet sampling distributions are the core of how we conduct statistical inference. You may need to reread this chapter a few times before it makes much sense, but doing so will be well worth your time if you want to understand applied statistics.

7.1 Introduction to Sampling Distributions[1]

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall from Chapter 5 that inferential statistics concern generalizing from a sample to a population (or to a counterfactual, but for this chapter we will focus on inferring about a population). A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter being estimated. (In this example, the sample statistics are the sample means and the population parameter is the population mean.) As the later portions of this chapter show, these determinations are based on sampling distributions.

7.1.1 Discrete Distributions

We will illustrate the concept of sampling distributions with a simple example. Figure 7.1 shows three pool balls, each with a number on it. Two of the balls are selected randomly (with replacement) and the average of their numbers is computed.

All possible outcomes are shown below in Table 7.1.

[1]This subsection is adapted from David M. Lane. "Introduction to Sampling Distributions." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/sampling_distributions/intro_samp_dist.html

Figure 7.1: The pool balls.

Table 7.1: All possible outcomes when two balls are sampled with replacement.

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in Table 6-2. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Table 7.2: Frequencies of means for n = 2.

Mean	Frequency	Relative Frequency
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

Figure 7.2 shows a relative frequency distribution of the means based on Table 7.2. This distribution is also a probability distribution since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.

The distribution shown in Figure 7.2 is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 (n = 2). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions. The pool balls have only the values 1, 2, and 3, and a sample mean can have one of only five values shown in Table 7.2.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement) and the mean of the two balls is computed and recorded. Then this process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean computed for each, a relative frequency distribution is drawn. The more

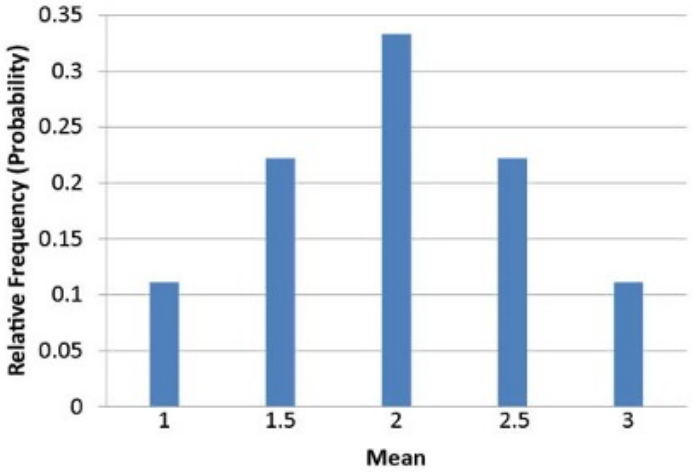

Figure 7.2: Distribution of means for n = 2.

samples, the closer the relative frequency distribution will come to the sampling distribution shown in Figure 7.2. As the number of samples approaches infinity, the relative frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a relative frequency distribution based on a very large number of samples. To be strictly correct, the relative frequency distribution approaches the sampling distribution as the number of samples approaches infinity.

It is important to keep in mind that every statistic, not just the mean, has a sampling distribution. For example, Table 7.3 shows all possible outcomes for the range of two numbers (larger number minus the smaller number). Table 7.4 shows the frequencies for each of the possible ranges and Figure 7.3 shows the sampling distribution of the range.

Table 7.3: All possible outcomes when two balls are sampled with replacement.

Outcome	**Ball 1**	**Ball 2**	**Range**
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

Table 7.4: Distribution of ranges for n = 2.

Range	Frequency	Relative Frequency
0	3	0.333
1	4	0.444
2	2	0.222

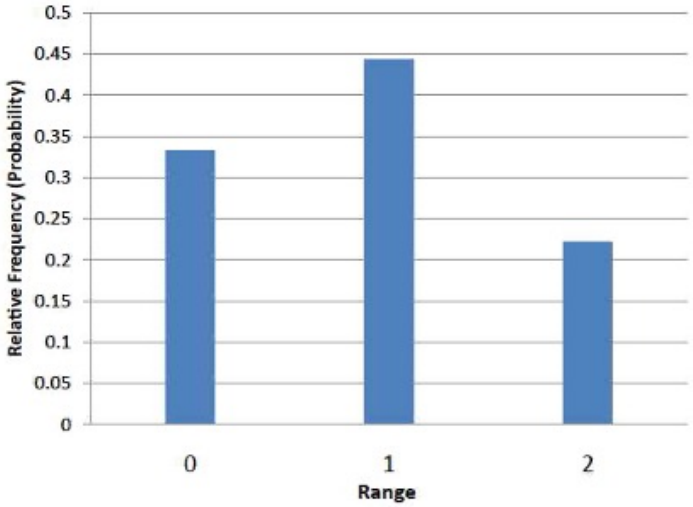

Figure 7.3: Distribution of ranges for n = 2.

It is also important to keep in mind that there is a sampling distribution for various sample sizes. For simplicity, we have been using n = 2. The sampling distribution of the range for n = 3 is shown in Figure 7.4.

7.1.2 Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? (Although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two balls. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this example since there are 1,000,000 possible outcomes (1,000 for the first ball x 1,000 for the second). Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency

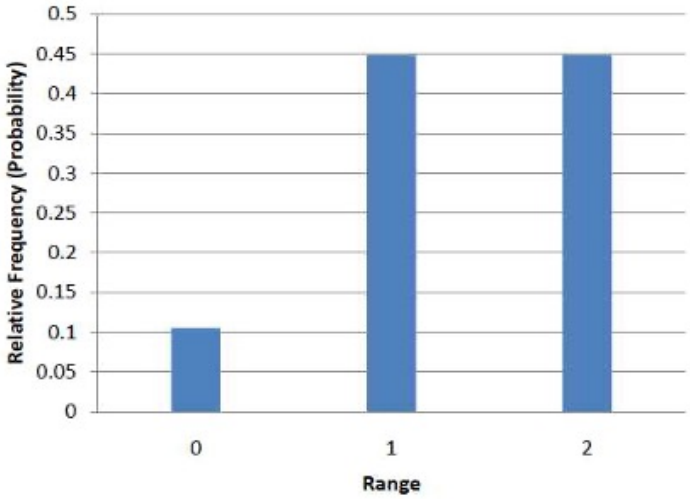

Figure 7.4: Distribution of ranges for n = 3.

distributions. Specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous distributions, the probability of obtaining any single value is zero. Therefore, these values are called probability densities rather than probabilities.

7.1.3 Sampling Distributions and Inferential Statistics

As we stated in the beginning of this chapter, **sampling distributions** are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: you collect sample data and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the **standard error** of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Keep in mind that all statistics have sampling distributions, not just the mean. For example, later in this chapter we will construct confidence intervals relying on the sampling distribution for a regression slope coefficient.

7.2 Sampling Distribution of the Mean[2]

As we learned in the prior section, the sampling distribution of the mean refers to the probability distribution describing all possible values of the sample mean we could obtain in repeated sampling. This section goes over some important properties of the sampling distribution of the mean.

7.2.1 Mean

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean ($\bar{X}$) is also μ. The symbol $\mu_{\bar{X}}$ is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_{\bar{X}} = \mu$$

7.2.2 Variance

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

[2]This subsection is adapted from David M. Lane. "Sampling Distribution of the Mean." *Online Statistics Education: A Multimedia Course of Study*. https://onlinestatbook.com/2/sampling_distributions/samp_dist_mean.html

That is, the variance of the sampling distribution of the mean is the population variance divided by n, the sample size (the number of scores used to compute a mean).[3] Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

As noted previously, the **standard error** of the mean is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The standard error is represented by a σ because it is a standard deviation. The subscript ($\bar{X}$) indicates that the standard error in question is the standard error of the (sample) mean.

7.2.3 Central Limit Theorem

The central limit theorem states that:

> Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/n as n, the sample size, increases.

The expressions for the mean and variance of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as n increases. Figure 7.5 shows the results of the simulation for n = 2 and n = 10. The parent population was a uniform distribution. You can see that the distribution for n = 2 is far from a normal distribution. Nonetheless, it does show that the scores are denser in the middle than in the tails. For n = 10 the distribution is quite close to a normal distribution. Notice that the means of the two distributions are the same, but that the spread of the distribution for n = 10 is smaller.

Figure 7.6 shows how closely the sampling distribution of the mean approximates a normal distribution even when the parent population is very non-normal. If you look closely you can see that the sampling distributions do have a slight positive skew. The larger the sample size, the closer the sampling distribution of the mean would be to a normal distribution.

[3] This expression can be derived very easily from the variance sum law. Let's begin by computing the variance of the sampling distribution of the sum of three numbers sampled from a population with variance σ^2. The variance of the sum would be $\sigma^2 + \sigma^2 + \sigma^2$. For n numbers, the variance would be $n\sigma^2$. Since the mean is $1/n$ times the sum, the variance of the sampling distribution of the mean would be $1/n^2$ times the variance of the sum, which equals σ^2/n.

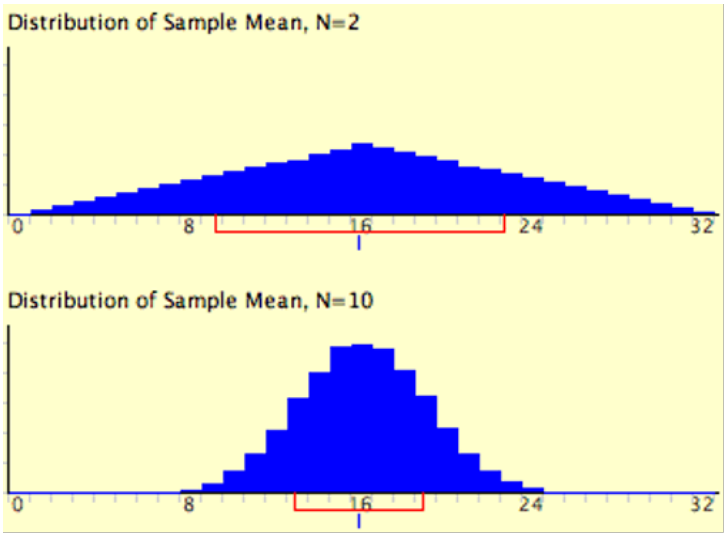

Figure 7.5: A simulation of a sampling distribution. The parent population is uniform. The blue line under "16" indicates that 16 is the mean. The red line extends from the mean plus and minus one standard deviation.

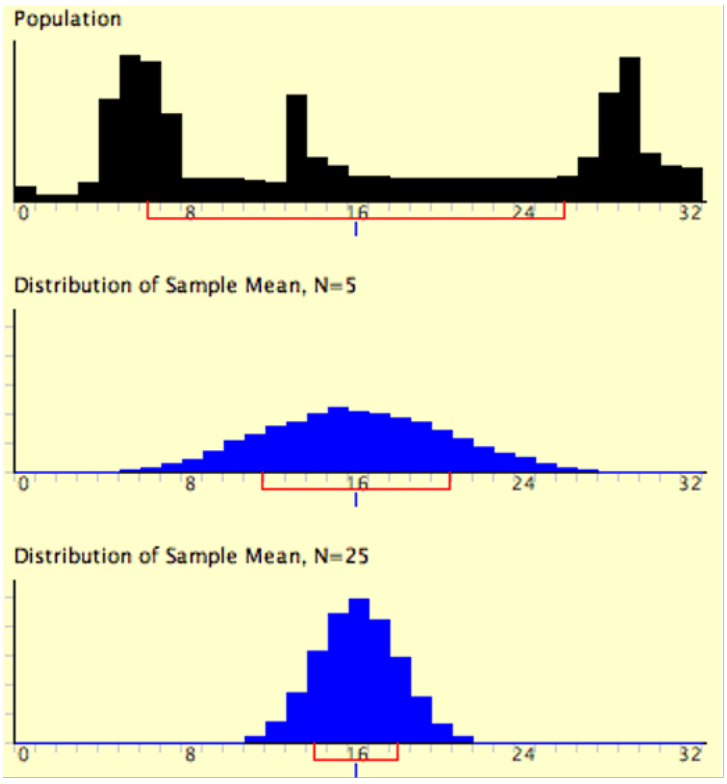

Figure 7.6: A simulation of a sampling distribution. The parent population is very non-normal.

7.3 Calculating Confidence Intervals

7.3.1 Confidence Intervals for the Mean[4]

When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean, there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

7.3.1.1 An Artificial Example: Using the Normal Distribution

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from Section 7.2 that the mean of the sampling distribution is μ and the standard error of the mean is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of $36/3 = 12$. Note that the standard deviation of a sampling distribution is its standard error. Figure 7.7 shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48 to 113.52. These limits were computed by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

$$90 - (1.96)(12) = 66.48$$

$$90 + (1.96)(12) = 113.52$$

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean (in Chapter 6, we used 2 as a close approximation, but 1.96 is a more precise value); 12 is the standard error of the mean given our sample size of 9.

Figure 7.7 shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is

[4]This subsubsection is adapted from David M. Lane. "Confidence Interval on the Mean." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/estimation/mean.html

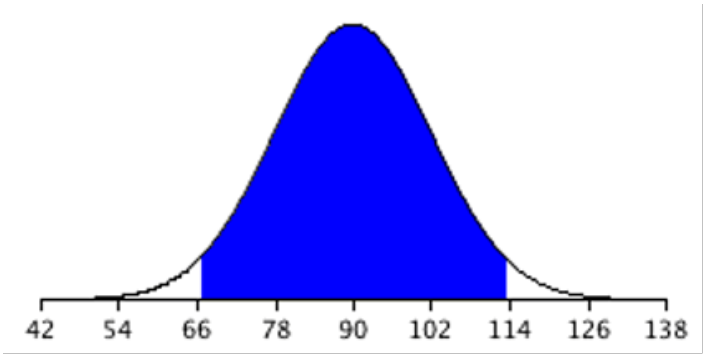

Figure 7.7: The sampling distribution of the mean for n=9. The middle 95% of the distribution is shaded.

within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the mean ($\bar{X}$) from a sample (drawing a new random sample each time), and create an interval ranging from $\bar{X} - 23.52$ to $\bar{X} + 23.52$, this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

$$\text{Lower limit} = \bar{X} - Z_{.95} \times \sigma_{\bar{X}}$$
$$\text{Upper limit} = \bar{X} + Z_{.95} \times \sigma_{\bar{X}}$$

where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area (always equal to 1.96) and $\sigma_{\bar{X}}$ is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the standard deviation (σ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when σ is known is easier than when σ has to be estimated, and serves a pedagogical purpose. Later in this section we will show how to compute a confidence interval for the mean when σ has to be estimated.

Suppose the following five numbers were sampled from a normal distribution with a standard deviation of 2.5: 2, 3, 5, 6, and 9. To compute the 95% confidence interval, start by computing the mean and standard error:

$$\bar{X} = (2 + 3 + 5 + 6 + 9)/5 = 5.$$

$$\sigma_{\bar{X}} = \frac{2.5}{\sqrt{5}} = 1.118.$$

$Z_{.95}$ can be found using the normal distribution calculator[5] and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points. As shown in Figure 7.8, the value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been 2.58.

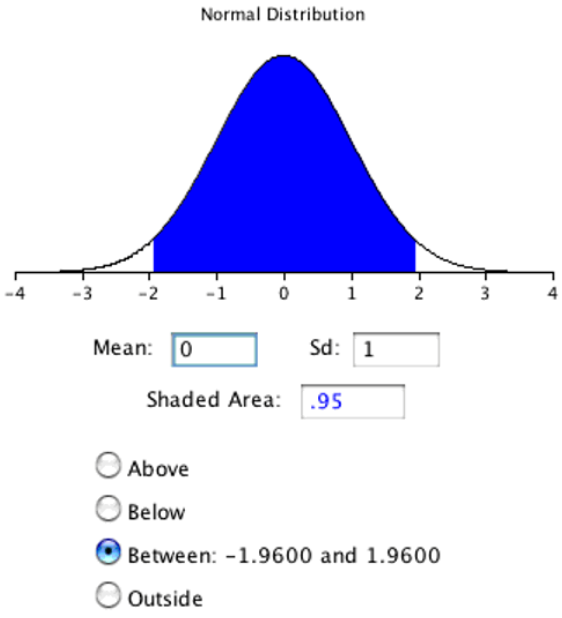

Figure 7.8: 95% of the area is between -1.96 and 1.96.

The confidence interval can then be computed as follows:

$$\text{Lower limit} = 5 - (1.96)(1.118) = 2.81$$

$$\text{Upper limit} = 5 + (1.96)(1.118) = 7.19$$

7.3.1.2 The Realistic Case: Using the T Distribution

You should use the t distribution rather than the normal distribution when the variance is not known and has to be estimated from sample data. You will learn more about the t distribution in the next section. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall

[5] https://onlinestatbook.com/2/calculators/normal_dist.html

that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78 standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution, a small version of which is provided in the following section. You can also use an "inverse t distribution" calculator[6] to find the t values to use in confidence intervals. With either approach, the t values will vary depending upon what is called the degrees of freedom (df). For confidence intervals on the mean, df is equal to n - 1, where n is the sample size.

Assume that the following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9 and that the standard deviation is not known. The first steps are to compute the sample mean and variance:

$$\bar{X} = 5$$

$$s^2 = 7.5$$

The next step is to estimate the standard error of the mean. If we knew the population variance, we could use the following formula:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Instead we compute an estimate of the standard error ($s_{\bar{X}}$). Note that since we previously computed the sample variance ($s^2 = 7.5$), we can take the square root of this to obtain the sample standard deviation (s):

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{7.5}}{\sqrt{5}} = 1.225$$

The next step is to find the value of t. As shown in Table 7.6 of the following section, the value for the 95% interval for df = n - 1 = 4 is 2.776. The confidence interval is then computed just as it is when $\sigma_{\bar{X}}$. The only differences are that $s_{\bar{X}}$ and t rather than $\sigma_{\bar{X}}$ and Z are used.

$$\text{Lower limit} = 5 - (2.776)(1.225) = 1.60$$

$$\text{Upper limit} = 5 + (2.776)(1.225) = 8.40$$

More generally, the formula for the 95% confidence interval on the mean is:

[6] https://onlinestatbook.com/2/calculators/inverse_t_dist.html

$$\text{Lower limit} = \bar{X} - (t_{CL})(s_{\bar{X}})$$
$$\text{Upper limit} = \bar{X} + (t_{CL})(s_{\bar{X}})$$

where $\bar{X}$ is the sample mean, t_{CL} is the t for the confidence level desired (0.95 in the above example), and $s_{\bar{X}}$ is the estimated standard error of the mean.

We will finish our discussion of confidence intervals for the mean with an analysis of the Stroop Data.[7] Specifically, we will compute a confidence interval on the mean difference score. As mentioned in Section 2.5, the study involved 47 subjects naming the color of ink that words were written in. An additional detail that is now relevant to us is that subjects completed similar naming tasks multiple times under different conditions. In the "interference" condition, the names conflicted so that, for example, they would name the ink color of the word "blue" written in red ink. The correct response is to say "red" and ignore the fact that the word is "blue." In a second condition, subjects named the ink color of colored rectangles.

Table 7.5: Response times in seconds for 10 subjects.

Naming Colored Rectangle	**Interference**	**Difference**
17	38	21
15	58	43
18	35	17
20	39	19
18	33	15
20	32	12
20	45	25
19	52	33
17	31	14
21	29	8

Table 7.5 shows the time difference between the interference and color-naming conditions for 10 of the 47 subjects. The mean time difference for all 47 subjects is 16.362 seconds and the standard deviation is 7.470 seconds. The standard error of the mean is 1.090. A t table shows the critical value of t for 47 - 1 = 46 degrees of freedom is 2.013 (for a 95% confidence interval). Therefore the confidence interval is computed as follows:

$$\text{Lower limit} = 16.362 - (2.013)(1.090) = 14.17$$
$$\text{Upper limit} = 16.362 + (2.013)(1.090) = 18.56$$

[7] https://onlinestatbook.com/2/case_studies/stroop.html

Therefore, the interference effect (difference) for the whole population is likely to be between 14.17 and 18.56 seconds.

7.3.2 More about the T Distribution[8]

In the introduction to normal distributions it was shown that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean. Therefore, if you randomly sampled a value from a normal distribution with a mean of 100, the probability it would be within 1.96σ of 100 is 0.95. Similarly, if you sample n values from the population, the probability that the sample mean ($\bar{X}$) will be within 1.96 $\sigma_{\bar{X}}$ of 100 is 0.95.

Now consider the case in which you have a normal distribution but you do not know the standard deviation. You sample n values and compute the sample mean ($\bar{X}$) and estimate the standard error of the mean ($\sigma_{\bar{X}}$) with $s_{\bar{X}}$. What is the probability that $\bar{X}$ will be within 1.96 $s_{\bar{X}}$ of the population mean (μ)? This is a difficult problem because there are two ways in which $\bar{X}$ could be more than 1.96 $s_{\bar{X}}$ from μ: (1) $\bar{X}$ could, by chance, be either very high or very low and (2) $s_{\bar{X}}$ could, by chance, be very low. Intuitively, it makes sense that the probability of being within 1.96 standard errors of the mean should be smaller than in the case when the standard deviation is known (and cannot be underestimated). But exactly how much smaller? Fortunately, the way to work out this type of problem was solved in the early 20th century by W. S. Gosset who determined the distribution of a mean divided by an estimate of its standard error. This distribution is called the *Student's t distribution* or sometimes just the t distribution. Gosset worked out the t distribution and associated statistical tests while working for a brewery in Ireland. Because of a contractual agreement with the brewery, he published the article under the pseudonym "Student." That is why the t test is called the "Student's t test."

The **t distribution** is very similar to the normal distribution when the estimate of variance is based on a large sample, but the t distribution has relatively more scores in its tails when there is a small sample. When working with the t distribution, sample size is expressed in what are called **degrees of freedom**. Degrees of freedom indicate the number of independent pieces of information on which an estimate is based; a more complete discussion of the concept is provided in Appendix I at the end of this chapter. As we noted in the prior section, when we are estimating the standard error for a sample mean, the degrees of freedom is simply equal to the sample size minus one (n-1).

Figure 7.9 shows t distributions with 2, 4, and 10 degrees of freedom and the standard normal distribution. Notice that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t distribution approaches the normal distribution as the degrees of freedom increase.

[8]This section is adapted from David M. Lane. "t Distribution." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/estimation/t_distribution.html

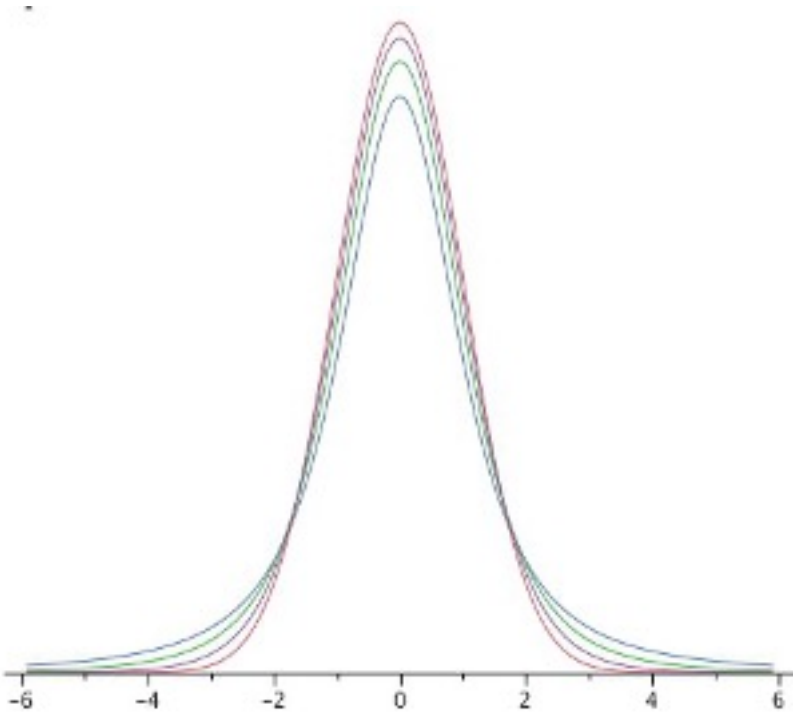

Figure 7.9: A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the highest is the standard normal distribution.

Since the t distribution has more area in the tails, the percentage of the distribution within 1.96 standard deviations of the mean is less than the 95% for the normal distribution. Table 7.6 shows the number of standard deviations from the mean required to contain 95% and 99% of the area of the t distribution for various degrees of freedom. These are the values of t that you use in a confidence interval. The corresponding values for the normal distribution are 1.96 and 2.58 respectively. Notice that with few degrees of freedom, the values of t are much higher than the corresponding values for a normal distribution and that the difference decreases as the degrees of freedom increase. The values shown in Table 6-7 can be obtained from statistical software or an online calculator.[9]

Table 7.6: Abbreviated t table.

df	**0.95**	**0.99**
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Returning to the problem posed at the beginning of this section, suppose you sampled 9 values from a normal population and estimated the standard error of the mean ($\sigma_{\bar{X}}$) with $s_{\bar{X}}$. What is the probability that $\bar{X}$ would be within $1.96 s_{\bar{X}}$ of μ? Since the sample size is 9, there are n - 1 = 8 df. From Table 7.6, you can see that with 8 df the probability is 0.95 that the mean will be within 2.306 $s_{\bar{X}}$ of μ. The probability that it will be within 1.96 $s_{\bar{X}}$ of μ is therefore lower than 0.95.

As shown in Figure 7.10, a t distribution calculator[10] can be used to find that 0.086 of the area of a t distribution is more than 1.96 standard deviations from the mean, so the probability that $\bar{X}$ would be less than $1.96 s_{\bar{X}}$ from μ is 1 - 0.086 = 0.914.

As expected, this probability is less than 0.95 that would have been obtained if $\sigma_{\bar{X}}$ had been known instead of estimated.

[9] https://onlinestatbook.com/2/calculators/inverse_t_dist.html
[10] https://onlinestatbook.com/2/calculators/t_dist.html

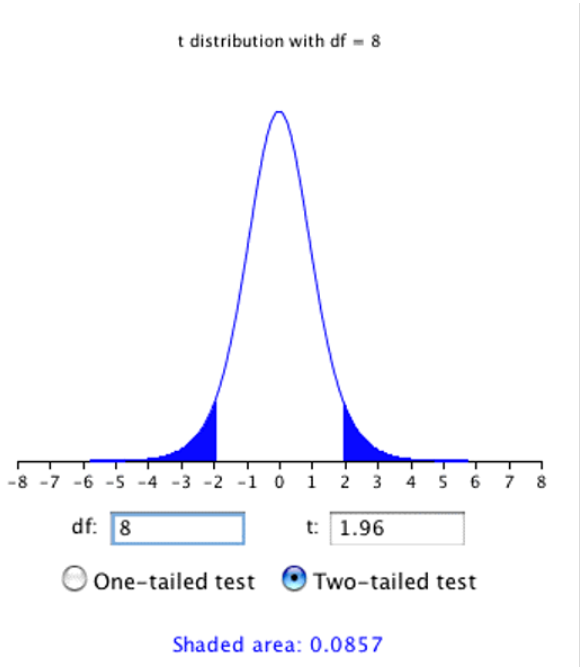

Figure 7.10: Area more than 1.96 standard deviations from the mean in a t distribution with 8 df. Note that the two-tailed button is selected so that the area in both tails will be included.

7.3.3 Confidence Intervals for a Regression Slope Coefficient [11]

The method for computing a confidence interval for the population slope in a simple linear regression is very similar to methods for computing other confidence intervals. For the 95% confidence interval, the formula is:

$$\text{Lower limit} = \hat{\beta} - (t_{.95})(s_\beta)$$

$$\text{Upper limit} = \hat{\beta} + (t_{.95})(s_\beta)$$

where $\hat{\beta}$ is the slope coefficient estimate, $t_{.95}$ is the value of t for 95% (2-tailed) confidence, and s_β is the standard error for the slope estimate. As before, the t value can be found from a table or an inverse t distribution calculator based on the degrees of freedom.

We illustrate generating a confidence interval using the same data as in the example from Section 3.4, depicted again here as Figure 7.11.

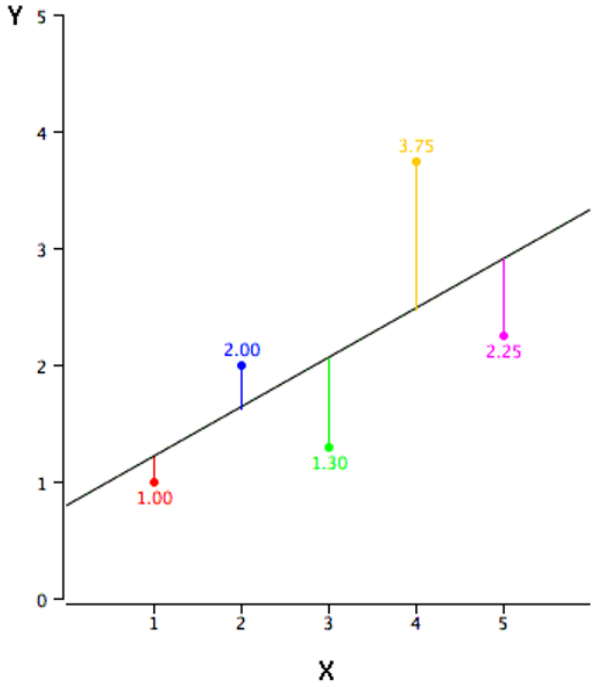

Figure 7.11: A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

[11]This section is adapted from David M. Lane. "Inferential Statistics for b and r." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/inferent

When conducting statistical inference for linear regression coefficients, the degrees of freedom is equal to the number of observations minus the number of coefficients being estimated (usually one for the intercept plus one for each independent variable). In the case of simple regression, there is just one independent variable plus a y-intercept, so the number of degrees of freedom is:

$$df = n - 2$$

where n is the number of pairs of scores (number of observations in the sample).

As we saw in Section 3.4, the estimated regression slope coefficient is 0.425 with this data. An n of 5 yields 3 degrees of freedom (5 - 2 = 3), which means the critical t value (at 95% confidence) is 3.182. Finally, the estimated standard error for this slope coefficient (the calculative of which is shown in this chapter's Appendix II) is 0.305 with this data.

Applying these formulas we obtain a confidence interval with the following lower and upper limits:

$$\text{Lower limit} = 0.425 - (3.182)(0.305) = -0.55$$

$$\text{Upper limit} = 0.425 + (3.182)(0.305) = 1.40$$

Chapter 7 Appendix I: Degrees of Freedom

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5. The **degrees of freedom (df)** of an estimate is the number of independent pieces of information on which the estimate is based.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean, $(8\text{-}6)^2 = 4$, is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df. If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance, $(5\text{-}6)^2 = 1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on

two independently and randomly selected Martians. The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

As you are probably thinking, it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean (μ) with the sample mean ($\bar{X}$). The process of estimating the mean affects our degrees of freedom as shown below.

Returning to our problem of estimating the variance in Martian heights, let's assume we do not know the population mean and therefore we have to estimate it from the sample. We have sampled two Martians and found that their heights are 8 and 5. Therefore $\bar{X}$, our estimate of the population mean, is

$$\bar{X} = (8 + 5)/2 = 6.5.$$

We can now compute two estimates of variance:

$$\text{Estimate 1} = (8 - 6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5 - 6.5)^2 = 2.25$$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of $\bar{X}$. Since the first Martian's height of 8 influenced $\bar{X}$, it also influenced Estimate 2. If the first height had been, for example, 10, then $\bar{X}$ would have been 7.5 and Estimate 2 would have been (5-7.5)^2 = 6.25 instead of 2.25. The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be 13-5 = 8.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest (σ^2). Therefore, the estimate of variance has 2 - 1 = 1 degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to n - 1, where n is the number of observations.

Recall from Section 2.4.4 that the formula for estimating the variance in a sample is:

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

The denominator of this formula is the degrees of freedom.

So far, we've mainly seen examples where the degrees of freedom is equal to n - 1. But as we saw in Section 7.3.3, the degrees of freedom can also be equal to other values such as n - 2. In the case of constructing a confidence interval for a coefficient from a multiple regression with four independent variables, the degrees of freedom will generally be equal to n - 5. The exact formula for degrees of freedom depends on what we are estimating.

Chapter 7 Appendix II: Estimating the Standard Error of a Regression Slope

This appendix shows how to compute the estimated standard error for the slope of a simple linear regression. The estimated standard error of β is computed using the following formula:

$$s_\beta = \frac{s_{est}}{\sqrt{SSX}}$$

where s_β is the estimated standard error of β, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X. SSX is calculated as:

$$SSX = \sum (X - \bar{X})^2$$

where $\bar{X}$ is the mean of X. The standard error of the estimate can be calculated as:

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{n - 2}}$$

where r is the correlation between X and Y, and SSY is the sum of squared deviations of Y from the mean of Y.

These formulas are illustrated with the data shown in Table 7.7. These data are reproduced from Section 3.4. The column X has the values of the independent variable and the column Y has the values of the dependent variable. The third column, x, contains the differences between the values of column X and the mean of X. The fourth column, x^2, is the square of the x column. The fifth column, y, contains the differences between the values of column Y and the mean of Y. The last column, y^2, is simply square of the y column.

Table 7.7: Example data.

	X	Y	x	x^2	y	y^2
	1.00	1.00	-2.00	4	-1.06	1.1236
	2.00	2.00	-1.00	1	-0.06	0.0036
	3.00	1.30	0.00	0	-0.76	0.5776
	4.00	3.75	1.00	1	1.69	2.8561
	5.00	2.25	2.00	4	0.19	0.0361
Sum	15.00	10.30	0.00	10.00	0.00	4.5970

SSY is the sum of squared deviations from the mean of Y. It is, therefore, equal to the sum of the y^2 column and is equal to 4.597. The correlation (r) between X and Y is 0.6268, and there are 5 observations (n=5). Thus, the standard error of the estimate is:

$$s_{est} = \sqrt{\frac{(1-(0.6268)^2)(4.597)}{5-2}} = \sqrt{\frac{2.791}{3}} = 0.964$$

SSX can be found as the sum of the x^2 column and is equal to 10.

We now have all the information to compute the standard error of β:

$$s_\beta = \frac{0.964}{\sqrt{10}} = 0.305$$

8 Theory of Hypothesis Testing

8.1 Introduction to Hypothesis Testing[1]

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from what is called the binomial distribution, and a binomial distribution calculator[2] shows it to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions[3] sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

[1] This section is adapted from David M. Lane. "Introduction." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/logic_of_hypothesis_testing/intro.html

[2] https://onlinestatbook.com/2/calculators/binomial_dist.html

[3] https://onlinestatbook.com/2/case_studies/weight.html

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for average-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the two groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in Section 9.1, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

8.1.1 The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial calculator, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the **probability value (p value)** is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and, therefore, are difficult to apply in some contexts.

8.1.2 The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the **null hypothesis**. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average}$$

or as

$$\mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a *population parameter* is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The **alternative hypothesis** is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{obese} = \mu_{average}$$

is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average}$$

$$\mu_{obese} > \mu_{average}$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960)[4] showed how it is justified to draw a conclusion about the direction of the difference.

8.2 Steps in Hypothesis Testing[5]

There's much to learn about hypothesis testing, but before going any further, here's an overview of the four basic steps of any hypothesis test. Some of the details won't make sense yet, but we'll explain them in more detail in the following sections.

1. The first step is to ***specify the null hypothesis***. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu_1 - \mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is $\mu_2 - \mu_1 \geq 0$. This is equivalent to $\mu_1 \leq \mu_2$.

[4]Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

[5]This section is adapted from David M. Lane. "Steps in Hypothesis Testing." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/logic_of_hypothesis_testing/steps.html

2. The second step is to ***specify the*** α level which is also known as the significance level. Typical values are 0.05 and 0.01.

3. The third step is to ***compute the probability value*** (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.

4. Finally, ***compare the probability value with the*** α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

8.3 One- and Two-Tailed Tests[6]

In the James Bond case study,[7] Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure 8.1 shows a graph of the binomial distribution. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a "one-tailed probability."

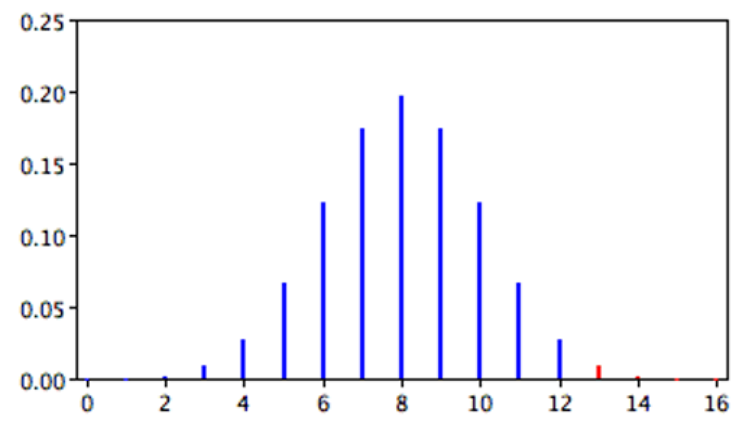

Figure 8.1: The binomial distribution. The upper (right-hand) tail is red.

[6]This section is adapted from David M. Lane. "One- and Two-Tailed Tests." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/logic_of_hypothesis_testing/tails.html

[7]https://onlinestatbook.com/2/case_studies/bond.html

A slightly different question can be asked of the data: "What is the probability of getting a result as extreme or more extreme than the one observed?" Since the chance expectation is 8/16, a result of 3/16 is equally as extreme as 13/16. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously. Therefore, p = 0.0212. A probability calculated in both tails of a distribution is called a "two-tailed probability" (see Figure 8.2).

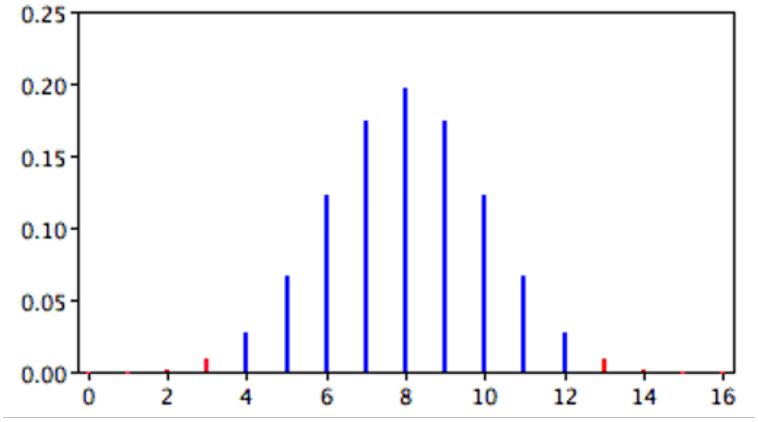

Figure 8.2: The binomial distribution. Both tails are red.

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond's performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is $\pi = 0.5$. By contrast, the null hypothesis for the one-tailed test is $\pi \leq 0.5$.[8] Accordingly, we reject the two-tailed hypothesis if the sample proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected

[8]Some sources write the null hypothesis of the one-tailed test identically to the two-tailed test ($\pi = 0.5$). While this alternative notation does not preserve the intuitive logic of the null hypothesis being the strict reverse of the alternative hypothesis, it does hint at how the p-value in a one-tailed test is calculated, since a distribution with $\pi = 0.5$ is used to determine the p-value (as shown in Figure 8.1).

only if the sample proportion is much greater than 0.5. The alternative hypothesis in the two-tailed test is $\pi \neq 0.5$. In the one-tailed test it is $\pi > 0.5$.

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of a treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified (according to the test) in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

8.4 Significance Testing[9]

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α (alpha) level or simply α. It is also called the significance level.

When the null hypothesis is rejected, the effect is said to be **statistically significant**. For example, in the Physicians' Reactions case study,[10] the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means in contexts outside of statistics. When an effect is statistically significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

[9] This section is adapted from David M. Lane. "Significance Testing." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/logic_of_hypothesis_testing/significance.html

[10] https://onlinestatbook.com/2/case_studies/weight.html

Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher), a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis.[11] If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

[11] See also: Goodman, W. M., Spruill, S. E., & Komaroff, E. (2019). A proposed hybrid effect size plus p-value criterion: empirical evidence supporting its use. *The American Statistician*, *73*(sup1), 168-185.

8.5 Testing a Single Mean[12]

The way we calculate the probability (p) value for a hypothesis test depends on what type of statement is made in our null hypothesis. Normally, statistical software will automatically compute a p value behind the scenes, but we still want to learn a bit about how the software comes up with this value. To illustrate what these calculations can look like, this section will focus on what to do if we want to test a null hypothesis stating that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures, and for each pair they are asked to select one. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50 (the number we would expect if subliminal messages have no effect). In other words, the null hypothesis is that the population mean (μ) is 50. The (hypothetical) data are shown in Table 8.1. The data in Table 8.1 have a sample mean ($\bar{X}$) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table 8.1: Distribution of scores.

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from μ by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As we learned in the prior chapter, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_{\bar{X}} = \mu$$

and

[12]This section is adapted from David M. Lane. "Testing a Single Mean." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/tests_of_means/single_mean.html

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

respectively. It is clear that if the null hypothesis is true, $\mu_{\bar{X}} = 50$. In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation (σ).

The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know σ and therefore you would use s, the sample estimate of σ. However, it is instructive to see how the probability is computed if σ is known before proceeding to see how it is calculated when σ is estimated.

For the current example, if the null hypothesis is true, then based on a well-established formula for the binomial distribution, one can compute that the variance of the number correct is

$$\sigma^2 = N\pi(1-\pi) = 100(0.5)(1-0.5) = 25$$

where N is the number of times a subject makes a selection between two pictures. Therefore, $\sigma = 5$ (since $\sigma = \sqrt{\sigma^2} = \sqrt{25} = 5$). For a σ of 5 and an n of 9, the standard deviation of the sampling distribution of the mean is $5/\sqrt{9} = 1.667$. Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or greater assuming the null hypothesis is true. If the null hypothesis is true, the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute the relevant probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use a normal distribution calculator as shown in Figure 8.3.

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area above 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results from a normal distribution calculator are shown in Figure 8.4.

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure 8.3.

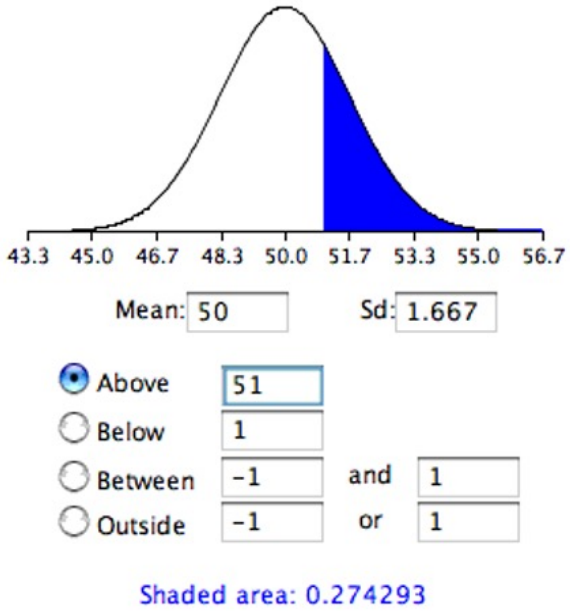

Figure 8.3: Probability of a sample mean being 51 or greater.

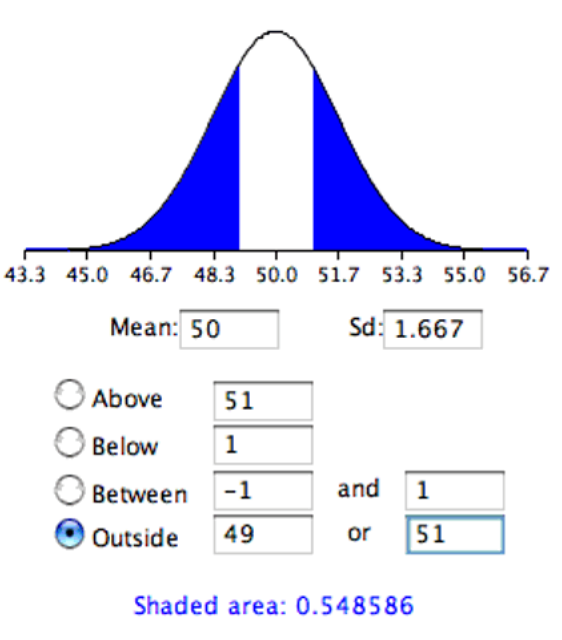

Figure 8.4: Probability of a sample mean being less than or equal to 49 or greater than or equal to 51.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by computing Z based on the formula

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

where Z is the value on the standard normal distribution, $\bar{X}$ is the sample mean, μ_0 is the hypothesized value of the mean (under the null hypothesis),[13] and $\sigma_{\bar{X}}$ is the standard error of the mean. For this example, $Z = (51\text{-}50)/1.667 = 0.60$. Use a normal calculator, with a mean of 0 and a standard deviation of 1, as shown below.

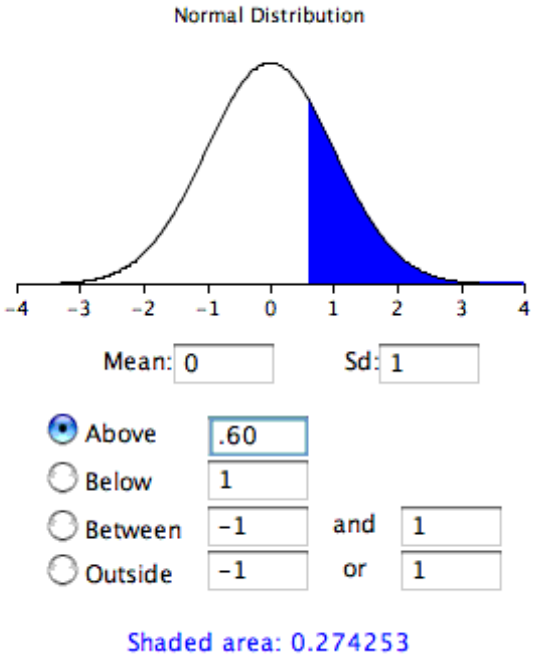

Figure 8.5: Calculation using the standardized normal distribution.

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analyses it is very rare that you would know σ and wish to estimate μ. Typically σ is not known and is estimated in a sample by s, and $\sigma_{\bar{X}}$ is estimated by $s_{\bar{X}}$. For our next example, we will consider the data in the "ADHD Treatment" case study.[14] These

[13] The subscript $_0$ in μ_0 (the population mean according to the null hypothesis) corresponds to how we typically represent the null hypothesis: H_0.

[14] https://onlinestatbook.com/2/case_studies/adhd.html

data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table 8.2 shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled "Diff" that shows the difference in performance between the 0.6 mg (D60) and the 0 mg (D0) conditions. These difference scores are positive for children who performed better in the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

Table 8.2: DOG scores as a function of dosage.

D0	**D60**	**Diff**
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

To test this null hypothesis, we compute what we call a t statistic (as opposed to a z statistic) because we will compare this value to the t distribution—a distribution which allows for accurate inferences when σ is estimated rather than known (see Section 7.3.2). We compute t using a special case of the following formula:

$$\text{t} = \frac{\text{statistic} - \text{hypothesized value}}{\text{estimated standard error of the statistic}}$$

The special case of this formula applicable to testing a single mean is

$$\text{t} = \frac{\bar{X} - \mu_0}{s_{\bar{X}}}$$

where t is the value we compute for the significance test, $\bar{X}$ is the sample mean, μ_0 is the hypothesized value of the population mean, and $s_{\bar{X}}$ is the estimated standard error of the mean. Notice the similarity of this formula to the formula for Z we saw before.

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean ($\bar{X}$) of the n = 24 difference scores is 4.958, the hypothesized value of μ is 0, and the standard deviation (s) is 7.538. The estimate of the standard error of the mean is computed as:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{7.5382}{\sqrt{24}} = 1.54$$

Therefore, t = 4.96/1.54 = 3.22. The probability value for t depends on the degrees of freedom. The number of degrees of freedom is equal to n - 1 = 23. As shown below, a t distribution calculator finds that the probability of a t less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.

In order to conduct this hypothesis test, we made the following *assumptions*:

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

Now that we've filled in more of the details of hypothesis tests, you may want to go back and review Section 8.2 to see whether you can follow the succinct overview of the hypothesis testing approach. Once you can follow that description, it is a good indication that you have understood the key concepts essential to every hypothesis test.

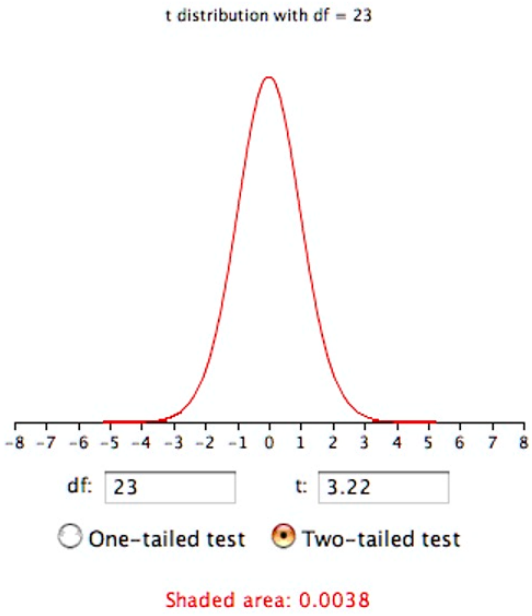

Figure 8.6: Calculation using the t distribution.

8.6 Type I and Type II Errors[15]

In the Physicians' Reactions case study,[16] the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a **Type I error** occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level. As discussed in the section on significance testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a

[15]This section is adapted from David M. Lane. "Type I and Type II Errors." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/logic_of_hypothesis_testing/errors.html

[16]https://onlinestatbook.com/2/case_studies/weight.html

reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the α level: the lower the α level, the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a **Type II error**. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The value of this probability β will be affected by the sample size (larger sample sizes make it less likely a false null hypothesis will fail to be rejected), but the exact formula for calculating β depends on the particular type of hypothesis test being conducted. The probability of correctly rejecting a false null hypothesis equals $1 - \beta$ and is called *statistical power*. When researchers say that a study is "well-powered," they mean that the sample size is large enough to reject a false null hypothesis with fairly high probability under certain assumptions (such as a reasonably large effect size).

8.7 Significance Test for a Regression Slope Coefficient[17]

To conclude this chapter, let's briefly revisit the very first type of hypothesis test we encountered in this textbook (though we did not call it that at the time): testing for the significance of a regression slope coefficient (Section 3.5). Now that we know more about hypothesis testing, let's fill in some of the details of how to calculate the p-values we rely upon to determine the significance of these coefficients.

The appropriate type of significance test in the case of the regression coefficients we have learn about is a t test. Recall the general formula for a t test:

$$t = \frac{\text{statistic - hypothesized value}}{\text{estimated standard error of the statistic}}$$

[17]This section is adapted from David M. Lane. "Inferential Statistics for b and r." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/inferent

As applied to the case of the slope in a simple regression, the statistic is the sample value of the slope coefficient ($\widehat{\beta}$). Generally, the hypothesized value is 0, meaning that we want to test a null hypothesis of no relationship between the independent and dependent variables.

Just as when we generated a confidence interval for the slope coefficient in Section 7.3.3, the degrees of freedom for this t test is n-2. We also use the same calculation for the estimated standard error as when calculating a confidence interval, so refer back to Chapter 7 (specifically Appendix II) if you would like to review how we calculate it.

With the data example we used when learning precise confidence interval calculations (Section 7.3.3), we had a sample slope coefficient ($\widehat{\beta}$) of 0.425, a standard error (s_{β}) of 0.305, and a sample size of 5. Given these numbers and a hypothesized value of 0:

$$t = \frac{0.425 - 0}{0.305} = 1.39$$
$$df = n - 2 = 5 - 2 = 3.$$

With these values of t and df, the p value for a two-tailed t test is 0.26. Therefore, the slope is not significantly different from 0 under this example.

9 Hypothesis Testing in Practice

Throughout this chapter, we will build on the tools we previously examined in Chapter 4 (comparisons of means and contingency tables). Specifically, we will learn about hypothesis tests that can be used in conjunction with previously discussed tools, allowing us to determine whether associations we observe are statistically significant.

9.1 Comparing Means (One Qualitative and One Quantitative Variable)

There are actually many different types of significance tests that can be used when comparing means or medians. In this section, we will introduce some of the most common ones.

9.1.1 Difference between Two Means[1]

This section covers how to test for differences between means from two separate groups of subjects, using an independent-groups t test.

We take as an example the data from the "Animal Research" case study.[2] In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 9.1.

Table 9.1: Means and Variances in Animal Research study.

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

[1]This section is adapted from David M. Lane. "Difference between Two Means (Independent Groups)." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/tests_of_means/difference_means.html

[2]https://onlinestatbook.com/2/case_studies/animal_research.html

The female mean is 1.47 units higher than the male mean. This is just the difference in our sample, however, and we wish to draw an inference about the difference in the *population* means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently[3] from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent.

One could look at these assumptions in much more detail, but suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

In practice, most researchers use software to automate calculation with all formulas we encounter in this chapter. Nonetheless, your ability to understand the output of the software may improve if you have some idea of what's happening under the hood. As we saw in the previous chapter, the following general formula is used for significance testing based on the t distribution:

$$\text{t} = \frac{\text{statistic} - \text{hypothesized value}}{\text{estimated standard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0 because the null hypothesis states that the difference between population means is 0.

We continue to use the data from the "Animal Research" case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males.

The first step is to compute the statistic, which is simply the difference between means.

$$\bar{X}_1 - \bar{X}_2 = 5.3529 - 3.8824 = 1.4705$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

[3]Two variables are said to be independent if the value of one variable provides no information about the value of the other variable. In this case, if knowing the value of the X variable for one observation could help us predict the value of X for another observation, the two values of X are not independent. For example, if there is clustered sampling, such that selecting an individual into the sample implies that neighbors with similar X values are also likely to be in the sample, the observations are not independent.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means, so the estimated standard error of the statistic is s_{diff} (which can also be written as $s_{\bar{X}_1 - \bar{X}_2}$). The formula for the standard error of the difference between means is:[4]

$$\sigma_{\text{diff}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

where σ_1^2 and n_1 are the variance and sample size of the first group, and σ_2^2 and n_2 are the variance and sample size of the second group. Note that since we assumed $\sigma_1^2 = \sigma_2^2$ (as our first of the three assumptions listed above), we can represent both of these variances as simply σ^2. Likewise, when $n_1 = n_2$ (as in our example with equal numbers of females and males), it is conventional to use "n" to refer to the sample size of each group.

Because the value of σ^2 is unknown, we estimate it by averaging our two sample variances, relying again on our assumption that the two population variances are the same (and thus each sample's variance should be an equally valid estimate of σ^2). This estimate of variance is can be written as follows:

$$\text{MSE} = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2. In this example,

$$\text{MSE} = (2.743 + 2.985)/2 = 2.864.$$

We can now estimate σ_{diff} with s_{diff}, substituting in MSE where we previously saw σ^2 in the formula for σ_{diff}. Since n (the number of scores in each group) is 17,

$$s_{\text{diff}} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805.$$

The next step is to compute t by plugging these values into the formula:

$$t = \frac{1.4705}{0.5805} = 2.533.$$

Finally, we compute the probability of getting a t as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of

[4] For a more detailed discussion, see https://onlinestatbook.com/2/sampling_distributions/samplingdist_diff_means.html

freedom is the number of independent estimates of variance on which MSE is based. This is equal to (n_1 - 1) + (n_2 - 1), and for this example, $n_1 = n_2 = 17$. Therefore, the degrees of freedom is 16 + 16 = 32.

Once we have the degrees of freedom, we can use a t distribution calculator[5] to find the probability. Figure 9.1 shows that the probability value (p) for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. As shown in Figure 9.1, it is the probability of a t < -2.533 or a t > 2.533.

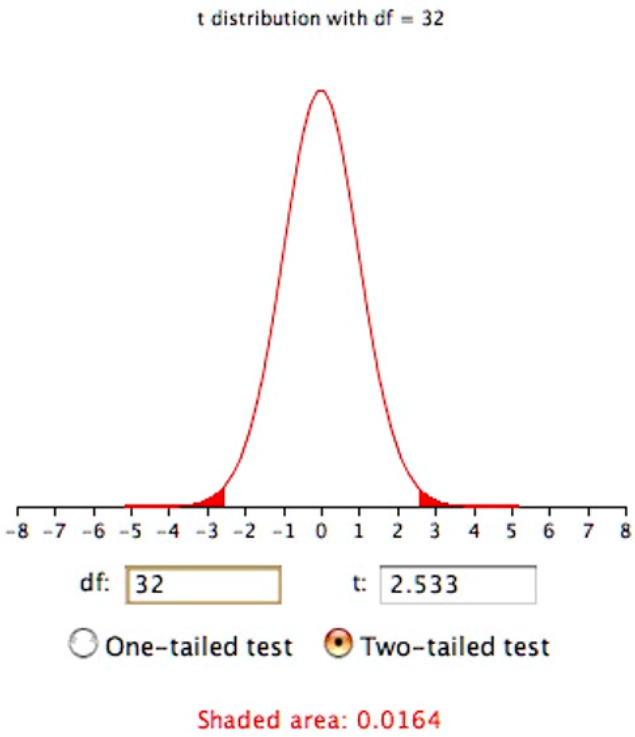

Figure 9.1: The two-tailed probability.

The results of a one-tailed test are shown in Figure 9.2. As you can see, the probability value of 0.0082 is half the value for the two-tailed test.

9.1.1.1 Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 9.2.

[5] https://onlinestatbook.com/2/calculators/t_dist.html

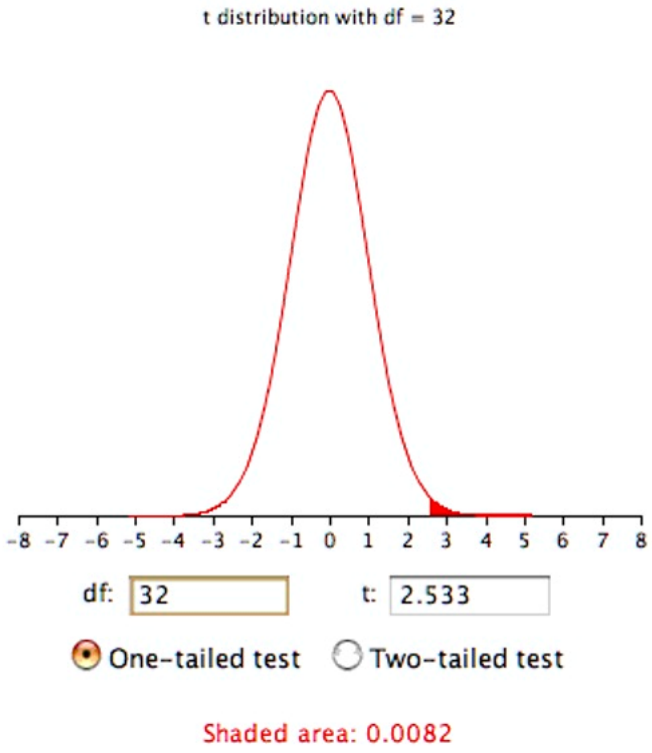

Figure 9.2: The one-tailed probability.

Table 9.2: Example data in "wide" form.

Group 1	Group 2
3	2
4	6
5	8

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 9.2 is shown in Table 9.3. We sometimes describe the original format as "wide" form and the reformatted data as "long" form.

Table 9.3: Reformatted data (now in "long" form).

Group	Y
1	3
1	4
1	5
2	2
2	6
2	8

Using statistical software, we'd find that the t value is -0.718, the df = 4, and p = 0.512.

9.1.2 Pairwise Comparisons Among Multiple Means[6]

Many experiments are designed to compare more than two conditions. We will take as an example the case study "Smiles and Leniency."[7] In this study, the effect of different smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, and miserable) were shown. "Type of Smile" is the independent variable, and the dependent variable is a leniency rating given by the subject to a fictional student (depicted with one of the four smiles) in an academic misconduct case. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table 9.4.

[6]This section is adapted from David M. Lane. "All Pairwise Comparisons Among Means." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/tests_of_means/pairwise.html

[7]https://onlinestatbook.com/2/case_studies/leniency.html

Table 9.4: Six Comparisons among Means.

false vs. felt		
false vs. miserable		
false vs. neutral		
felt vs. miserable		
felt vs. neutral		
miserable vs. neutral		

You can certainly conduct a series of six t tests in this manner. However, one potential problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05.[8] The more

[8]When discussing probability of Type I errors, we assume all null hypotheses are true, since a Type I error can't occur if the null hypothesis is false.

means that are compared, the more the Type I error rate is inflated. Figure 9.3 shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

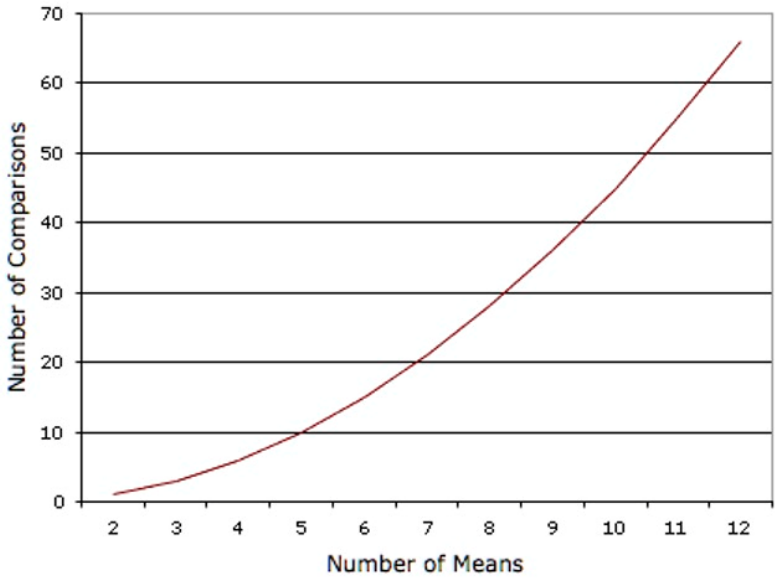

Figure 9.3: Number of pairwise comparisons as a function of the number of means.

Figure 9.4 shows the probability of a Type I error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or Tukey HSD for short. The Tukey HSD test is one example of a multiple comparison test, but several alternatives are frequently used, such as the Bonferroni correction. Regardless of the exact method used for a multiple comparison test, the interpretation of results is similar. The Tukey HSD is based on a variation of the t distribution that takes into account the number of means being compared. This distribution is called the studentized range distribution.

Normally, statistical software will make all the necessary calculations for you in the background. But to illustrate what sorts of calculations the software is relying on, let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. For our example, they are shown in Table 9.5.

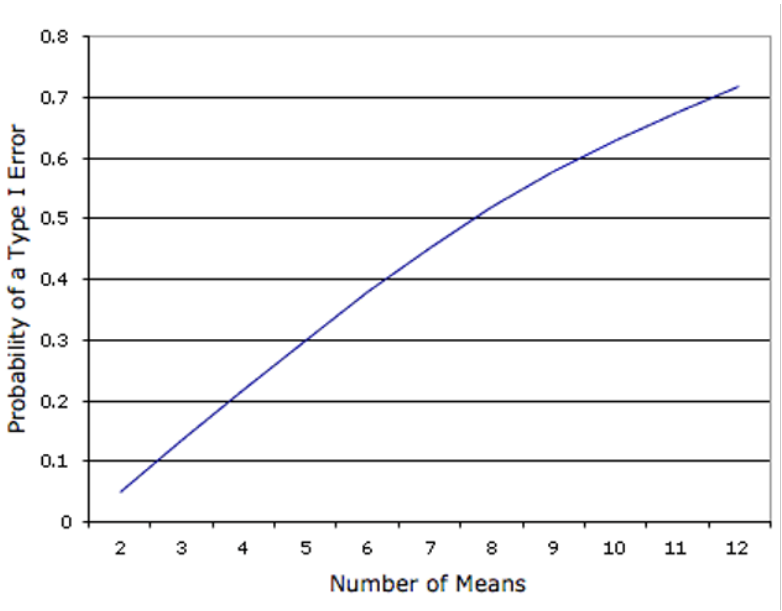

Figure 9.4: Probability of a Type I error as a function of the number of means.

Table 9.5: Means and Variances from the "Smiles and Leniency" Study.

Condition	Mean	Variance
False	5.37	3.34
Felt	4.91	2.83
Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.

3. Compute Q (using the formula below) for each pair of means, where $\bar{X}_i$ is one mean, $\bar{X}_j$ is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

$$Q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MSE}{n}}}$$

4. Compute p for each comparison using a Studentized Range Calculator.[9] The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, df = 136 - 4 = 132.

[9] https://onlinestatbook.com/2/calculators/studentized_range_dist.html

The tests for these data are shown in Table 9.6.

Table 9.6: Six Pairwise Comparisons.

Comparison	$\bar{X}_i - \bar{X}_j$	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

The only significant comparison is between the false smile and the neutral smile.

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in-between.

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

9.1.2.1 Computer Analysis

For most computer programs, you should format your data the same way you do for an independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

9.1.3 Analysis of Variance (ANOVA)[10]

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." The name is appropriate because inferences about means are made by analyzing variance, as outlined in this chapter's appendix.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example, so we will continue considering the data on leniency and smiles we examined in the prior section on the Tukey HSD test.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is,

$$\mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral}$$

in our example. More generally, the null hypothesis tested by ANOVA is that the population means for all conditions are the same. For whatever data is being examined, this can be written as:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

where H_0 is the null hypothesis and k is the number of conditions (k = 4 in our example).

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the Tukey HSD test. The Tukey HSD is therefore preferable to ANOVA in this situation.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analyses that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is one of the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

[10]The initial material in this subsection (until the header indicating otherwise) is adapted from David M. Lane. "Introduction." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/analysis_of_variance/intro.html

9.1.3.1 The Critical Step: Calculating an F Ratio[11]

There are many types of ANOVA, but for our example, we will use what is called a one-factor between-subjects design. Other types of ANOVA are beyond the scope of what is covered in this text.

More details are provided in this chapter's appendix, but the critical step in an ANOVA is comparing what is called the mean square error (MSE) to the mean square between (MSB). MSB estimates a larger quantity than MSE only when the population means are not equal, so finding a larger MSB than an MSE is a sign that the population means are not equal. But since MSB could be larger than MSE by chance even if the population means are equal, MSB must be much larger than MSE in order to justify the conclusion that the population means differ. But how much larger must MSB be? For the "Smiles and Leniency" data, the MSB and MSE are 9.179 and 2.649, respectively. Is that difference big enough? To answer, we would need to know the probability of getting that big a difference or a bigger difference if the population means were all equal. The mathematics necessary to answer this question were worked out by the statistician R. Fisher. Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

For these data, the F ratio is

$$F = \frac{9.179}{2.649} = 3.465.$$

Therefore, the MSB is 3.465 times higher than MSE. Would this have been likely to happen if all the population means were equal? That depends on the sample size. With a small sample size, it would not be too surprising because results from small samples are unstable. However, with a very large sample, the MSB and MSE are almost always about the same (assuming the null hypothesis is true), and an F ratio of 3.465 or larger would be very unusual. Figure 9.5 shows the sampling distribution of F for the sample size in the "Smiles and Leniency" study. As you can see, it has a positive skew.

From Figure 9.5, you can see that F ratios of 3.465 or above are unusual occurrences. The area to the right of 3.465 represents the probability of an F that large or larger and is equal to 0.018. In other words, given the null hypothesis that all the population means are equal, the probability value (p) is 0.018 and therefore the null hypothesis can be rejected. The conclusion that at least one of the population means is different from at least one of the others is justified.

The shape of the F distribution depends on the sample size. More precisely, it depends on two degrees of freedom (df) parameters: one for the numerator (MSB) and one for the denominator

[11]This subsection and the following are adapted from David M. Lane. "One-Factor ANOVA (Between Subjects)." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/analysis_of_variance/one-way.html

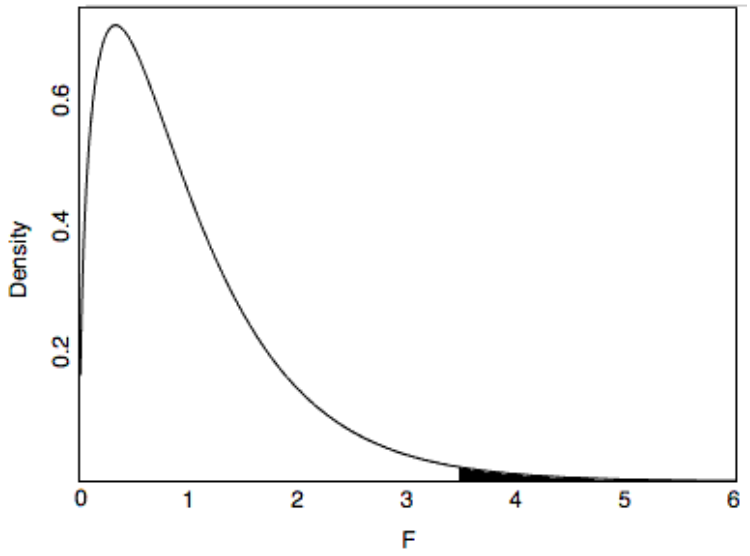

Figure 9.5: Distribution of F.

(MSE). Recall that the degrees of freedom for an estimate of variance is equal to the number of observations minus one. Since the MSB is the variance of k means (where k is the number of groups), it has k - 1 df. The MSE is an average of k variances, each with n - 1 df. Therefore, the df for MSE is k(n - 1) = N - k, where N is the total number of observations, n is the number of observations in each group, and k is the number of groups. To summarize:

$$df_{\text{numerator}} = k - 1$$
$$df_{\text{denominator}} = N - k$$

For the "Smiles and Leniency" data,

$$df_{\text{numerator}} = k - 1 = 4 - 1 = 3$$
$$df_{\text{denominator}} = N - k = 136 - 4 = 132$$
$$F = 3.465$$

An F distribution calculator[12] shows that p = 0.018. Again, because this value is less than 0.05, one would generally reject the null hypothesis and conclude that average leniency varies depending on type of smile. The p-value from an ANOVA is sometimes reported in a larger table of summary results such as Table 9.7.

[12] https://onlinestatbook.com/2/calculators/F_dist.html

Table 9.7: ANOVA Summary Table.

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.465	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

9.1.3.2 Relationship to T Tests and Regression

Since an ANOVA and an independent-groups t test can both test the difference between two means, you might be wondering which one to use. Fortunately, it does not matter since the results will always be the same. When there are only two groups, the following relationship between F and t will always hold:

$$F(1, dfd) = t^2(df)$$

where dfd is the degrees of freedom for the denominator of the F test and df is the degrees of freedom for the t test. dfd will always equal df. And because of how their probability distributions are constructed, these values of F and t will yield identical p-values for the (two tailed) null hypothesis of no difference between the two means.

There is also a third equivalent way to compare two means: using linear regression, as described in Section 4.2.2. More generally, linear regression and ANOVA are two sides of the same coin and will yield equivalent results (assuming the same data/assumptions), even when testing for differences among more than two means. Statistical software will generally include a model F statistic among the results shown for a regression, and in the case of a model a single qualitative independent variable, the regression model F statistic will be the same F ratio used in an ANOVA. Because of this equivalence, whether one reports results as an ANOVA or regression is usually a matter of habit and familiarity. In some social science literatures, ANOVA results are rarely reported because researchers typically default to using regression instead.

9.2 Chi Square Tests for Contingency Tables (Two Qualitative Variables)[13]

We previously learned how to describe the relationship between two qualitative variables with a contingency table or bar chart bar chart (Chapter 4). To make a comparison that includes a significance test, we will need to use a distribution called the Chi Square distribution. As

[13]This section is adapted from David M. Lane. "Contingency Tables." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/chi_square/contingency.html

such, our significance test will be called a Chi Square test. A description of the Chi Square distribution itself can be found in @appendix-chi-square.

To demonstrate the Chi Square test, we again look to data from the Mediterranean Diet and Health case study,[14] in which heart attack survivors were randomly assigned to follow one of two diets. Looking to Table 9.8, we want to know whether there is a *significant relationship* between diet and outcome.

Table 9.8: Frequencies for Diet and Health Study.

	Out-come				Total
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	
AHA	15	24	25	239	303
Mediter-ranean	7	14	8	273	302
Total	22	38	33	512	605

As with all other hypothesis tests in this chapter, the null hypothesis indicates no relationship between the two values. And once again, software can calculate a p-value for us to evaluate this hypothesis. But if we are wondering what's going on under the hood, the first step is to compute the expected frequency for each cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the AHA Diet-Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome (as the null hypothesis states), then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect $(0.0364)(303) = 11.02$ cancers on the AHA diet. Similarly, we would expect $(0.0364)(302) = 10.98$ cancers on the Mediterranean diet. In general, the expected frequency for a cell in the ith row and the jth column is equal to

$$E_{ij} = \frac{T_i T_j}{T}$$

where E_{ij} is the expected frequency for cell i, j, T_i is the total for the ith row, T_j is the total for the jth column, and T is the total number of observations. For the AHA Diet-Cancers cell, $i = 1$, $j = 1$, $T_i = 303$, $T_j = 22$, and $T = 605$. Table 9.9 shows the expected frequencies (in parenthesis) for each cell in the experiment.

The significance test is conducted by computing Chi Square as follows.

[14] https://onlinestatbook.com/2/case_studies/diet.html

$$\chi^2_3 = \sum \frac{(E-O)^2}{E} = -16.55$$

The degrees of freedom is equal to $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. For this example, the degrees of freedom is $(2-1)(4-1) = 3$. A Chi Square calculator[15] can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is equal to 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

Table 9.9: Observed and Expected Frequencies for Diet and Health Study.

	Outcome			
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)
Mediter-ranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)
Total	22	38	33	512

9.2.1 Drawing Substantive Conclusions from a Contingency Table[16]

Now that we know there is a statistically significant relationship between diet and health outcome (since we rejected the null hypothesis of no relationship), we will naturally wonder what kind of relationship exists. Specifically, which diet is associated with better health? To answer this, we must look to the specific values within the cells and describe the patterns we observe. We already covered how to do this in Chapter 4 when percentages are provided by column or row. An even more straightforward way to interpret the association is to make note of where the observed frequency differs notably from the expected frequency.

From Table 9.9, we can see that for those on the AHA diet, the frequencies for cancers, fatal heart disease, and non-fatal heart disease are all higher than expected. At the same time, the frequency for "healthy" is lower than expected on the AHA diet. Results for the Mediterranean diet are the exact opposite: cancers and both types of heart disease occur less frequently than expected, and the healthy outcome occurs more than expected. Thus, the Mediterranean diet is unambiguously associated with better outcomes than the AHA diet.

Though it is easy to make sense of results in this manner based on a comparison of observe to expected frequencies, the typical contingency table you encounter will probably not display

[15] https://onlinestatbook.com/2/calculators/chi_square_prob.html
[16] This subsection is written by Nathan Favero.

expected frequencies. Instead, it is common to include percentages by row or by column, which is why we focused on interpreting such tables in Section 4.1.

The good news is that regardless of whether we examine (1) expected versus observed outcomes, (2) percentages by row, or (3) percentages by column, we reach the same conclusion: health outcomes are uniformly better for those on the Mediterranean diet. All three approaches are equally valid ways of evaluating associations from a contingency table, and you normally need use only one. We briefly mention all three here for learning purposes since each approach is one you may encounter in your own analysis or in a research report.

Chapter 9 Appendix I: More about ANOVA

Terminology for Various Designs[17]

There are many types of experimental designs that can be analyzed by ANOVA. This section discusses many of these designs and defines several key terms used.

Factors and Levels

In describing an ANOVA design, the term factor is a synonym of independent variable. Therefore, in the case study "Smiles and Leniency," "Type of Smile" is the factor in this experiment. Since four types of smiles were compared, the factor "Type of Smile" has four levels.

An ANOVA conducted on a design in which there is only one factor is called a one-way ANOVA. If an experiment has two factors, then the ANOVA is called a two-way ANOVA. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

Between- and Within-Subjects Factors

In the "Smiles and Leniency" study, the four levels of the factor "Type of Smile" were represented by four separate groups of subjects. When different subjects are used for the levels of a factor, the factor is called a between-subjects factor or a between-subjects variable. The term "between subjects" reflects the fact that comparisons are between different groups of subjects.

In the "ADHD Treatment" study,[18] in which every subject was tested with each of four dosage

[17]This subsection is adapted from David M. Lane. "Analysis of Variance Designs." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/analysis_of_variance/anova_designs.html

[18]https://onlinestatbook.com/2/case_studies/adhd.html

levels (0, 0.15, 0.30, 0.60 mg/kg) of a drug. Therefore there was only one group of subjects, and comparisons were not between different groups of subjects but between conditions within the same subjects. When the same subjects are used for the levels of a factor, the factor is called a within-subjects factor or a within-subjects variable. Within-subjects variables are sometimes referred to as repeated-measures variables since there are repeated measurements of the same subjects.

Multi-Factor Designs

It is common for designs to have more than one factor. For example, consider a hypothetical study of the effects of age and gender on reading speed in which males and females from the age levels of 8 years, 10 years, and 12 years are tested. There would be a total of six different groups as shown in Table 9.10.

Table 9.10: Gender x Age Design.

Group	Gender	Age
1	Female	8
2	Female	10
3	Female	12
4	Male	8
5	Male	10
6	Male	12

This design has two factors: age and gender. Age has three levels and gender has two levels. When all combinations of the levels are included (as they are here), the design is called a *factorial design*. A concise way of describing this design is as a Gender (2) x Age (3) factorial design where the numbers in parentheses indicate the number of levels. Complex designs frequently have more than two factors and may have combinations of between- and within-subjects factors.

Details of One-Factor ANOVA (Between Subjects)[19]

This section shows how ANOVA can be used to analyze a one-factor between-subjects design.

Analysis of variance is a method for testing differences among means by analyzing variance. The test is based on two estimates of the population variance (σ^2). One estimate is called the mean square error (MSE) and is based on differences among scores within the groups. MSE

[19]This subsection is adapted from David M. Lane. "One-Factor ANOVA (Between Subjects)." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/analysis_of_variance/one-way.html

estimates σ^2 regardless of whether the null hypothesis is true (the population means are equal). The second estimate is called the mean square between (MSB) and is based on differences among the sample means. MSB only estimates σ^2 if the population means are equal. If the population means are not equal, then MSB estimates a quantity larger than σ^2. Therefore, if the MSB is much larger than the MSE, then the population means are unlikely to be equal. On the other hand, if the MSB is about the same as MSE, then the data are consistent with the null hypothesis that the population means are equal.

Before proceeding with the calculation of MSE and MSB, it is important to consider the assumptions made by ANOVA:

1. The populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the values are not independent; to accomodate such data, one must use within-subjects ANOVA (a type of ANOVA which is easily implemented but which lies beyond the scope of this text).

These assumptions are the same as for a t test of differences between groups (Section 9.1.1) except that they apply to two or more groups, not just to two groups.

Sample Sizes

As in the main part of the chapter, we will use as our example the "Smiles and Leniency" case study. The first calculations in this section all assume that there is an equal number of observations in each group (unequal sample size calculations are shown later in this appendix). We will refer to the number of observations in each group as n and the total number of observations as N. For these data there are four groups of 34 observations. Therefore, n = 34 and N = 136.

Computing MSE

Recall that the assumption of homogeneity of variance states that the variance within each of the populations (σ^2) is the same. This variance, σ^2, is the quantity estimated by MSE and is computed as the mean of the sample variances. For these data, the MSE is equal to 2.6489.

Computing MSB

The formula for MSB is based on the fact that the variance of the sampling distribution of the mean is

$$\sigma_{\mu}^2 = \frac{\sigma^2}{n}$$

where n is the sample size of each group. Rearranging this formula, we have

$$\sigma^2 = n\sigma_{\mu}^2.$$

Therefore, if we knew the variance of the sampling distribution of the mean, we could compute σ^2 by multiplying it by n. Although we do not know the variance of the sampling distribution of the mean, we can estimate it with the variance of the sample means. For the leniency data, the variance of the four sample means is 0.270. To estimate σ^2, we multiply the variance of the sample means (0.270) by n (the number of observations in each group, which is 34). We find that MSB = 9.179.

To sum up these steps:

1. Compute the means.
2. Compute the variance of the means.
3. Multiply the variance of the means by n.

Recap

If the population means are equal, then both MSE and MSB are estimates of σ^2 and should therefore be about the same. Naturally, they will not be exactly the same since they are just estimates and are based on different aspects of the data: The MSB is computed from the sample means and the MSE is computed from the sample variances.

If the population means are not equal, then MSE will still estimate σ^2 because differences in population means do not affect variances. However, differences in population means affect MSB since differences among population means are associated with differences among sample means. It follows that the larger the differences among sample means, the larger the MSB. *In short, MSE estimates σ^2 whether or not the population means are equal, whereas MSB estimates σ^2 only when the population means are equal and estimates a larger quantity when they are not equal.*

As shown in Section 9.1.3.1, we compare the MSE to the MSB by way of an F ratio in order to determine a p-value for the null hypothesis that the population means are all equal.

One-Tailed or Two?

Is the probability value from an F ratio a one-tailed or a two-tailed probability? In the literal sense, it is a one-tailed probability since, as you could see in Figure 9.5 earlier in the chapter, the probability is the area in the right-hand tail of the distribution. However, the F ratio is sensitive to any pattern of differences among means. It is, therefore, a test of a two-tailed hypothesis and is best considered a two-tailed test.

Sources of Variation

Why do scores in an experiment differ from one another? Consider the scores of two subjects in the "Smiles and Leniency" study: one from the "False Smile" condition and one from the "Felt Smile" condition. An obvious possible reason that the scores could differ is that the subjects were treated differently (they were in different conditions and saw different stimuli). A second reason is that the two subjects may have differed with regard to their tendency to judge people leniently. A third is that, perhaps, one of the subjects was in a bad mood after receiving a low grade on a test. You can imagine that there are innumerable other reasons why the scores of the two subjects could differ. All of these reasons except the first (subjects were treated differently) are possibilities that were not under experimental investigation and, therefore, all of the differences (variation) due to these possibilities are unexplained. It is traditional to call unexplained variance error even though there is no implication that an error was made. Therefore, the variation in this experiment can be thought of as being either variation due to the condition the subject was in or due to error (the sum total of all reasons the subjects' scores could differ that were not measured).

One of the important characteristics of ANOVA is that it partitions the variation into its various sources. In ANOVA, the term sum of squares (SSQ) is used to indicate variation. The total variation is defined as the sum of squared differences between each score and the mean of all subjects. The mean of all subjects is called the grand mean and is designated as GM. (When there is an equal number of subjects in each condition, the grand mean is the mean of the condition means.) The total sum of squares is defined as

$$SSQ_{\text{total}} = \sum (X - GM)^2$$

which means to take each score, subtract the grand mean from it, square the difference, and then sum up these squared values. For the "Smiles and Leniency" study, $\text{SSQ}_{\text{total}} = 377.19$.

The sum of squares condition is calculated as shown below.

$$SSQ_{\text{condition}} = n\left[(\bar{X}_1 - GM)^2 + (\bar{X}_2 - GM)^2 + ... + (\bar{X}_k - GM)^2\right]$$

where n is the number of scores in each group, k is the number of groups, $\bar{X}_1$ is the mean for Condition 1, $\bar{X}_2$ is the mean for Condition 2, and $\bar{X}_k$ is the mean for Condition k. For the Smiles and Leniency study, the values are:

$$SSQ_{\text{condition}} = 34\left[(5.37-4.83)^2+(4.91-4.83)^2+(4.91-4.83)^2+(4.12-4.83)^2\right]$$
$$= 27.5$$

If there are unequal sample sizes, the only change is that the following formula is used for the sum of squares condition:

$$SSQ_{\text{condition}} = n_1(\bar{X}_1-GM)^2+n_2(\bar{X}_2-GM)^2+...+n_k(\bar{X}_k-GM)^2$$

where n_i is the sample size of the ith condition. $\text{SSQ}_{\text{total}}$ is computed the same way as shown above.

The sum of squares error is the sum of the squared deviations of each score from its group mean. This can be written as

$$SSQ_{\text{error}} = \sum(X_{i1}-\bar{X}_1)^2+\sum(X_{i2}-\bar{X}_2)^2+...+\sum(X_{ik}-\bar{X}_k)^2.$$

where X_{i1} is the ith score in group 1 and $\bar{X}_1$ is the mean for group 1, X_{i2} is the ith score in group 2 and $\bar{X}_2$ is the mean for group 2, etc. For the "Smiles and Leniency" study, the means are: 5.368, 4.912, 4.912, and 4.118. The SSQ_{error} is therefore:

$$(2.5-5.368)^2+(5.5-5.368)^2+...+(6.5-4.118)^2 = 349.65$$

The sum of squares error can also be computed by subtraction:

$$SSQ_{\text{error}} = SSQ_{\text{totoal}} - SSQ_{\text{condition}}$$
$$SSQ_{\text{error}} = 377.189 - 27.535 = 349.65$$

Therefore, the total sum of squares of 377.19 can be partitioned into $SSQ_{\text{condition}}$ (27.53) and SSQ_{error} (349.66).

Once the sums of squares have been computed, the mean squares (MSB and MSE) can be computed easily. The formulas are:

$$MSB = \frac{SSQ_{\text{condition}}}{dfn}$$

where dfn is the degrees of freedom numerator and is equal to k - 1 = 3.

$$MSB = \frac{27.535}{3} = 9.18$$

which is the same value of MSB obtained previously (except for rounding error). Similarly,

$$MSE = \frac{SSQ_{\text{error}}}{dfd}$$

where dfd is the degrees of freedom for the denominator and is equal to N - k.

$$dfd = 136 - 4 = 132$$
$$MSE = 349.66/132 = 2.65$$

which is the same as obtained previously (except for rounding error). Note that the dfd is often called the dfe for degrees of freedom error.

As we saw in the main portion of the chapter, SSQ and MSB/MSE can be reported alongside the F statistic and p-value for an ANOVA in a results table (Table 9.7). Since most people conducting ANOVA these days do so with automated computer software, you will likely see a results table along these lines in whatever software you use if you conduct ANOVA yourself.

Formatting Data for Computer Analysis

Most computer programs that compute ANOVAs require your data to be in a specific form. Consider the data in Table 9.11.

Table 9.11: Example data in wide form.

Group 1	Group 2	Group 3
3	2	8
4	4	5
5	6	5

Here there are three groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 9.11 is shown in Table 9.12.

Table 9.12: Reformatted data (now in long form).

Group	Y
1	3
1	4
1	5
2	2
2	4
2	6
3	8
3	5
3	5

Chapter 9 Appendix II: More about the Chi Square Distribution and its Tests

9.2.1 Chi Square Distribution[20]

A standard normal deviate is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates, and the degrees of freedom of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as $\chi^2(1)$, is simply the distribution of a single normal deviate squared. The area of the $\chi^2(1)$ distribution below 4 is the same as the area of a standard normal distribution below 2, since 4 is 2^2.

Consider the following problem: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher? Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a Chi Square (with 2 df) being six or higher is 0.050.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, but as the degrees of freedom increases, the degree of skew decreases and the Chi Square distribution approaches a normal distribution. Figure 9.6 shows density functions for three Chi Square distributions. Notice how the skew decreases as the degrees of freedom increases.

The Chi Square distribution is very important because many test statistics are approximately distributed as Chi Square. Two of the more common tests using the Chi Square distribution are

[20]This section is adapted from David M. Lane. "Chi Square Distribution." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/chi_square/distribution.html

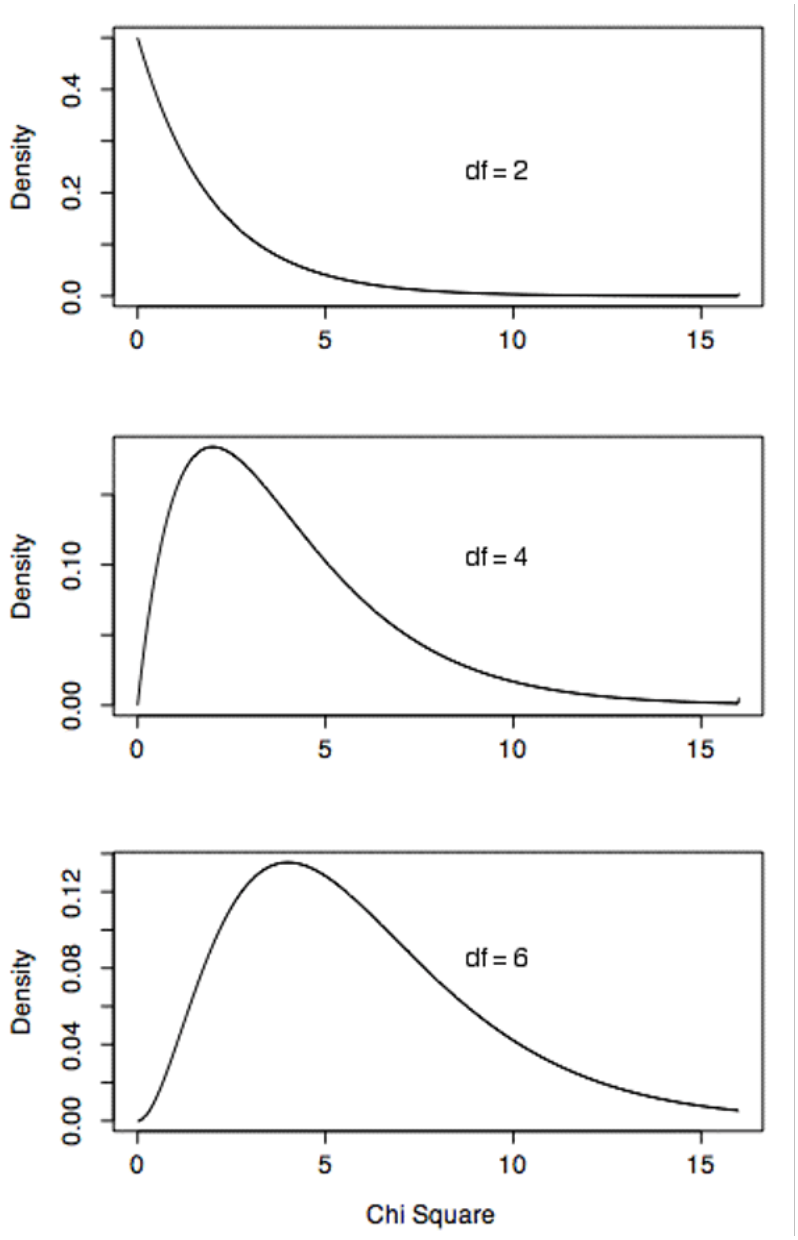

Figure 9.6: Chi Square distributions with 2, 4, and 6 degrees of freedom.

tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between qualitative variables (contingency tables). Numerous other tests beyond the scope of this work are based on the Chi Square distribution.

9.2.2 One-Way Tables[21]

The Chi Square distribution can be used to test whether observed data differ *significantly* from theoretical expectations. For example, for a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table 9.13 were obtained by rolling a six-sided die 36 times. However, as can be seen in Table 9.13, some outcomes occurred more frequently than others. For example, a "3" came up nine times, whereas a "4" came up only two times. Are these data consistent with the hypothesis that the die is a fair die? Naturally, we do not expect the sample frequencies of the six possible outcomes to be the same since chance differences will occur. So, the finding that the frequencies differ does not mean that the die is not fair. One way to test whether the die is fair is to conduct a significance (hypothesis) test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Table 9.13: Outcome Frequencies from a Six-Sided Die.

Outcome	Frequency
1	8
2	5
3	9
4	2
5	7
6	5

The first step in conducting the significance test is to compute the expected frequency for each outcome given that the null hypothesis is true. For example, the expected frequency of a "1" is 6 since the probability of a "1" coming up is 1/6 and there were a total of 36 rolls of the die.

Expected frequency = $(1/6)(36) = 6$

Note that the expected frequencies are expected only in a theoretical sense. We do not really "expect" the observed frequencies to match the "expected frequencies" exactly.

[21]This section is adapted from David M. Lane. "One-Way Tables (Testing Goodness of Fit)." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/chi_square/one-way.html

The calculation continues as follows. Letting E be the expected frequency of an outcome and O be the observed frequency of that outcome, compute

$$\frac{(E-O)^2}{E}$$

for each outcome. Table 9.14 shows these calculations.

Table 9.14: Outcome Frequencies from a Six-Sided Die.

Outcome	**E**	**O**	$\frac{(E-O)^2}{E}$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.500
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Next we add up all the values in Column 4 of Table 9.14.

$$\sum \frac{(E-O)^2}{E} = 5.333$$

This sampling distribution of

$$\sum \frac{(E-O)^2}{E}$$

is approximately distributed as Chi Square with k-1 degrees of freedom, where k is the number of categories (in our example, the six possible values we might get from a die roll). Therefore, for this problem the test statistic is

$$\chi^2_5 = 5.333$$

which means the value of Chi Square with 5 degrees of freedom is 5.333.

From a Chi Square calculator[22] it can be determined that the probability of a Chi Square of 5.333 or larger is 0.377. Therefore, the null hypothesis that the die is fair cannot be rejected.

This Chi Square test can also be used to test other deviations between expected and observed frequencies. The following example shows a test of whether the variable "University GPA"

[22] https://onlinestatbook.com/2/calculators/chi_square_prob.html

in the "SAT and College GPA" case study[23] (used previously to demonstrate regression) is normally distributed.

The first column in Table 9.15 shows the standard normal distribution divided into four ranges. The second column shows the proportions of a standard normal distribution falling in the ranges specified in the first column. For example, we see that for the range 0 to 1, the proportion is 0.341 (or 34.1%). This follows directly from what we learned in Chapter 6; 68% of the area under the standard normal distribution lies between -1 and 1, so naturally 34% (half of 68%) will be covered by [0, 1]. The expected frequencies (E) are calculated by multiplying the number of scores (105) by the proportion expected according to the standard normal distribution. The final column shows the observed number of scores (O) in each range, after standardizing the University GPA variable so that it will map onto the standard normal distribution (see Section 6.1.1). It is clear from the table that the observed frequencies vary greatly from the expected frequencies. Note that if the distribution were normal, we would expect only about 35 scores between 0 and 1, whereas 60 were observed.

Table 9.15: Expected and Observed Scores for 105 University GPA Scores.

Range	**Proportion**	**E**	**O**
Above 1	0.159	16.695	9
0 to 1	0.341	35.805	60
-1 to 0	0.341	35.805	17
Below -1	0.159	16.695	19

The test of whether the observed scores deviate significantly from the expected scores is computed using the familiar calculation.

$$\chi^2_3 = \sum \frac{(E-O)^2}{E} = 30.09$$

The subscript "3" means there are three degrees of freedom. As before, the degrees of freedom is the number of outcomes (in our example, the number of ranges listed in Table 9.15) minus 1, which is 4 - 1 = 3 in this example. A Chi Square distribution calculator shows that $p < 0.001$ for this Chi Square. Therefore, the null hypothesis that the scores are normally distributed can be rejected.

[23] https://onlinestatbook.com/2/case_studies/sat.html

Part III

Putting Data to Work

10 Research Designs and Causality

We analyze data because we wish to learn things about the world around us, but all data has limitations. How do we establish what our data can and cannot tell us? There are no simple answers to this question; the critical thinking we have been practicing with Three Questions to Always Ask about Data is one place to start. But more generally, we can help clarify what our data is capable of revealing through a **research design**. A research design describes how your data relates to the particular questions you hope to answer. Ideally, your research design is compelling enough that someone (other scientists, maybe even yourself) could be convinced to rethink their opinion if the results the design yields don't match their expectations. In other words, when we speak of research design, we are trying to separate out the *process* from the *results*. The research design describes the process through we obtain and analyze the data. Using a rigorous process will make your results more credible.

10.1 Types of Research Designs

There are many types of research designs, so it can be helpful to organize them into different categories. For example, we often distinguish between inductive and deductive research. **Inductive research** refers to drawing general conclusions (inferring broader principles or patterns) based on observations of specific examples. We often use the term **exploratory** when describing inductive research. Inductive studies can help us to develop hypotheses. We might begin an inductive study with some research questions, or we might not even have particularly precise questions at the start of the study.

Deductive research applies broader theories or principles to specific situations or data. For example, a deductive study might test a hypothesis (or the implications of some theory) in a particular setting. We can also refer to this as **confirmatory** research. One important element of confirmatory work is that it can be clearly stated what it would look like to get results that negate the argument being tested.

In practice, the lines between inductive and deductive research are often blurry. Many studies use elements of both approaches.

We can also categorize research as descriptive versus causal. **Descriptive** research addresses questions about what is. For example, we might want to know how closely the general public follows the news, or whether countries generally moved away from democracy during the past

decade. **Causal** research allows us to get at "why" questions. Why do some people follow the news more carefully than others? Why do countries move away from democracy?

Once again, the lines between categories are often blurry in practice. For example, many studies advance causal arguments about what is occurring in the world, but the actual data analyzed may not allow for drawing any strong conclusions about causality.

Another way we can categorize research is to distinguish between experimental and observational studies. In an **experiment**, the researcher is involved in manipulating one or more variables of interest. We saw an example in Chapter 5, where researchers studying a food assistance program assigned applicants to either receive text message reminders about an interview or to be part of a control group that received no such reminders. In social science, we usually want to randomize any experimental manipulation (e.g., use a random number generator to determine whether a subject receives the treatment or control), in order to match the assumptions of statistical models used for analysis.

In **observational studies**, the researcher observes variation in variables caused by something other than the researcher's own intervention. There are practical and ethical barriers to manipulating many variables we care about in the social world, so important research questions are not always suitable for experimental study. A sub-category of observational studies is **quasi-experimental studies**, which utilitize research designs aimed at assessing causality rigorously, despite lacking true experimental manipulation. For example, a researcher might study a specific policy shock, such as a court ruling that altered a policy in certain jurisdictions but not others (creating plausibly distinct "treatment" and "control" groups). While such designs are generally beyond the scope of what is covered in this text, they are an important and growing part of the social scientific literature.[1]

The concepts of internal and external validity can help us describe the strengths and weaknesses of various research designs. **Internal validity** refers to confidence that a causal conclusion can be drawn about one or more relationships among variables. Internal validity refers specifically to learning about the causal effects that exist *among the units observed in the study.* **External validity** describes confidence that the findings of a study can be generalized to a broader set of units, beyond those directly observed in the study. For an application of these concepts, consider that many classic psychology studies consisted of lab experiments conducted with undergraduate psychology students. While well-constructed lab experiments allow for strong conclusions to be reached about the causal effects of a manipulation on the students within the lab (good internal validity), such studies have also been criticized for poor external validity since undergraduate psychology students may tend to react differently to certain stimuli than the general public. More recent innovations like the use of online survey experiments have allowed psychologists to regularly collect data from more a diverse cross section of the public, although the precision and control afforded by a lab setting is weakened in an online experiment. Thus, online experiments may be generally considered to have weaker internal validity than lab

[1] For an excellent conceptual overview of several quasi-experimental designs, see Chapter 13 of Wheelan, C. (2010.) *Introduction to Public Policy.* New York: W. W. Norton & Company.

experiments (due to less precise control of experimental manipulations), while the more diverse populations associated with online experiments may afford greater external validity. Many other considerations are important for a detailed assessment of the internal and external validity of any given experiment, but this broad (and somewhat simplistic) summary of experimental psychology illustrates how these concepts can help us identify important aspects of research designs to scrutinize.

10.2 Causality

Causality is a complex concept that is difficult to precisely define. One way to think about causality in a social science context is as the sequence in which our variables are ordered. Researchers often depict variables sequentially with directional arrows showing the presumed causal connections among variables. As already introduced in Section 3.4, we give variables different designations depending on where they appear in this sequence (although we introduced this idea by focusing on prediction rather than causation). An **independent variable** is supposed to be a cause of the **dependent variable**. If we have a sequence that extends beyond two variables, we can call an in-between variable a **mediator** or **mediating variable** (e.g., if A causes B and B causes C, we consider B to be a mediator in the relationship between A and C).

10.2.1 A framework for assessing causality

How can we evaluate whether an independent variable X causes a dependent variable Y? There are many different tools for assessing causality, but for now we will introduce a simple framework that can help us informally evaluate evidence.

To begin with, we start from the assumption that an association (e.g., a correlation or non-zero regression slope) between X and Y has been found. If X does cause Y, then there should be some sort of association between the two variables,[2] even though an association is not sufficient to conclude causation. If no association is found, then our data indicate no evidence in support of a causal relationship.

Under this framework, there are five possibilities for why X is associated with Y:

1. The association is a coincidence
2. Z causes X and Y
3. Y causes X
4. Research design problems create an artificial association
5. X causes Y

[2]This association can take the form of a partial correlation (and a bivariate correlation may be altogether absent) if there is a confounding effect that masks the bivariate association between the two variables that are causally linked.

Assessing causality under this framework is a bit like detective work: we can potentially use the process of elimination to establish causality. Specifically, if we rule out options 1-4, we can conclude 5 must be true. Of course, we do not normally reach purely binary conclusions (that something is certainly true or certainly false); instead, we are weighing evidence and assessing the relatively plausibility of these 5 possibilities. The more confident we are that 1-4 are untrue, the more sure we are that 5 is true.

Let's briefly discuss some considerations under each of the five possibilities.

1. The association is a coincidence

The social world is fully of complexity and variation, so we can never hope to create a perfectly sterile environment where everything is held constant except a single variable. In other words, we always have an error term to contend with, as described in Section 6.2. There is always a risk that by pure coincidence, the random noisiness of the world will yield an apparent association in our particular sample, even if there is no systematic linkage in reality.[3] Fortunately, confidence intervals and hypothesis tests explicitly allow us to account for such random noise. Thus, the standard way studies address this first possibility is by testing whether an association is strong enough to achieve statistical significance. When we achieve statistical significance, we are essentially concluding that the relationship between variables is unlikely to be coincidental.

In many ways, this first possibility is the easiest to assess, given that the fundamental tools of statistical inference are designed to address it. Yet for some research questions, it impossible to collect large samples, making statistical significance very difficult to achieve. For example, studies of presidential elections within a single country typically suffer from small sample sizes, since current institutional practices and data availability usually extend back at most for several decades (and presidential elections typically occur once every several years).

Another common difficulty is that failing to meet model assumptions can distort the results of hypothesis tests, as when standard errors are not accurately estimated.

Cherry picking of results (or data) is another common concern, since false positives will sometimes occur due to coincidence (at a rate consistent with the chosen alpha level, at least in theory). If insignificant results are discarded and only significant results are presented, the rate of false positives among the remaining results could be dramatically inflated. Recent attention to issues of p-hacking and publication bias directly address such concerns, and efforts to adapt research designs to incorporate practices like preregistration may help to mitigate such problems in the social scientific literature.

[3] In fact, a sample correlation between two variables will almost never be exactly 0, even if two variables are unrelated to one another.

2. Z causes X and Y

This possibility is perhaps the most vexing cause for concern in observational studies. If a third variable Z causes both X and Y, then X and Y will generally exhibit an association even if there is no direct causal link between X and Y. Such a third variable may be called a **confounder** (or confounding variable). For example, suppose an observational study finds that participants in a microloan program experience substantial improvements in economic wellbeing compared to peers who did not participate in the program. If the program had an opt-in element, we should be worried about self-selection distorting accurate estimation of program effects. People with higher levels of ambition (a third variable Z) will probably be more likely to participate in the program (the X variable), but this ambition will likely also serve to boost future economic wellbeing (the Y variable). Thus, even if the program itself has no effect on future economic wellbeing, we can still expect to find a positive association between X and Y due to Z affecting both variables.

If we can successfully identify any (and all) confounders and are able to perfectly measure them, we can control for them (include them as additional independent variables) in a regression, which will generally address this concern. Specifically, if Z is the only confounder of concern, we can run a multiple regression that includes both X and Z as independent variables (and Y as the dependent variable). If X exhibits a (significant) association with Y in this multiple regression, we can generally be satisfied that Z was not the cause of the association between X and Y since multiple regression will estimate an association for X independent of Z.

However, as a practical matter, it is very difficult to be confident we have identified and precisely measured all potential confounders. Going back to the microloan example, a practical difficulty is that ambition is difficult to precisely measure, challenging our ability to fully remove any confounding effect of ambition by controlling for it in a regression. Given such difficulties, the most persuasive tests of causality generally rely on examining variation in X that is believed to be random (as in an experiment) or near-random (e.g., varying substantially and sharply in response to a clear cause, such that third factors are unlikely to be varying in a similarly arbitrary pattern). If the value of X was randomly assigned (e.g., determined by the result of a random number generator), then we have no reason to worry that some confounder Z caused both X and Y (since the result of the random number generator should have no direct effect on Y). This is why experiments utilizing random assignment are considered the gold standard for building evidence of causality.

3. Y causes X

Sometimes, we can be fairly confident that this is not a concern. For example, if X clearly precedes Y in time and there is no concern about anticipatory effects (i.e., it is implausible that Y could be predicted or that people adjusted X in anticipation of Y), we might logically conclude that Y causing X is unlikely. Or we might simply deem it rather implausible that Y would affect X based on our existing understanding of social behavior. For example, we might

assumed that voting intention does not affect economic conditions, since it is hard to imagine a mechanism by which macroeconomic conditions would notable shift in response to how people planned to vote (at least assuming a reasonably close election for which the results were in doubt ahead of time).

Sometimes, collecting data over time (e.g., panel data) will help us better evaluate this possibility. Random or near-random variation again provides some of the best means of address this concern (where such variation can plausibly be identified), since a random assignment of values to X implies that Y was not causing the values of X.

4. Research design problems create an artificial association

This fourth possibility is quite open ended, since artificial findings of an association may arise due to a variety of issues associated with a study's design. We cannot possibly provide an exhaustive list here, so some examples will have to suffice. A study might suffer attrition (people dropping out of a study) in particular patterns that distort the picture of how variables are associated with one another. More generally, non-random patterns of missing data may bias estimates of associations. Measurement error can also bias results, especially if misreporting is correlated with another variable of interest. Another common concern in social science is that people may distort their behavior due to awareness that they are being studied (a Hawthorne effect) or treated (placebo effects); good research designs will make efforts to mitigate such effects by, for example, creating a carefully constructed control condition for an experiment.

5. X causes Y

Beyond considering whether there are any good rival explanations (possibilities 1-4), it is important to assess the plausibility of this relationship itself. Is there a theory or a plausible mechanism that explains how X could affect Y? If we are examining the effects of a policy change on future electoral outcomes, is the public broadly aware of the policy or its effects? If not, it is probably difficult to imagine how the policy change could have a large effect on a subsequent election.[4]

[4]Of course in some settings, it might be plausible that elites (who have greater awareness of the policy change) can affect public sentiment through endorsements or campaign contributions. The greater point is that the plausibility of such mechanisms should be assessed on their own terms; establishing a plausible mechanism for how X could affect Y makes this fifth possibility itself more plausible when weighing it against the other four possibilities in this framework.

10.2.2 Establishing Causation in Experiments[5]

Consider a simple experiment in which subjects are sampled randomly from a population and then assigned randomly to either the experimental group or the control group. Assume the condition means on the dependent variable differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in "minutes slept" could have been due to a chance difference between the control group and the experimental group rather than due to the drug's effect.

This problem seems intractable since, by definition, it is impossible to measure an "unmeasured variable" just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods (e.g., the independent-groups t test from Section 9.1) determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the difference obtained in the experiment. If that probability is low, then it is inferred (that's why they call it inferential statistics) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

[5]This subsection and the next are adapted from David M. Lane. "Causation." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/research_design/causation.html

10.2.3 Causation in Non-Experimental Designs

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the third variable problem and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975)[6] to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better, though admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010)[7] found a strong correlation between public debt and GDP growth. Although some have argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.[8]

Chapter 10 Appendix: Classic Experimental Designs from Psychology

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one independent variable or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

[6] Li, C. (1975) *Path analysis: A primer.* Boxwood Press, Pacific Grove, CA.

[7] Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, https://www.nber.org/papers/w15639

[8] For a video on causality featuring evidence that smoking causes cancer, see https://www.learner.org/series/against-all-odds-inside-statistics/the-question-of-causation/

10.2.1 Between-Subjects Designs

In a **between-subjects** design, the various experimental treatments are given to different groups of subjects. For example, in the "Teacher Ratings"[9] case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor's lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the independent variable is "Condition" and has two levels (charismatic teacher and punitive teacher). It is a between-subjects variable because different subjects were used for the two levels of the independent variable: subjects were in either the "charismatic teacher" or the "punitive teacher" condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using inferential statistics, it can be calculated that the probability of finding a difference as large or larger than the one obtained if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the "Smiles and Leniency" case study the independent variable is "type of smile" and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

10.2.2 Multi-Factor Between-Subject Designs

In the "Bias Against Associates of the Obese"[10] experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as

[9] https://onlinestatbook.com/2/case_studies/ratings.html
[10] https://onlinestatbook.com/2/case_studies/obesity_relation.html

an Associate's Weight (2) x Associate's Relationship (2) factorial design. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an *interaction* (also known as moderation) between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

10.2.3 Within-Subjects Designs

A **within-subjects** design differs from a between-subjects design in that the same subjects perform at all levels of the independent variable. For example consider the "ADHD Treatment"[11] case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under *each* of the four levels of the independent variable "dose," the design is a within-subjects design and dose is a within-subjects variable. Within-subjects designs are sometimes called repeated-measures designs.

10.2.4 Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more power (ability to find precise estimates) than between-subjects designs. That is, this makes within-subjects

[11] https://onlinestatbook.com/2/case_studies/adhd.html

designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called "repeated-measures" designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

10.2.5 Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the "Weapons and Aggression"[12] case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

[12] https://onlinestatbook.com/2/case_studies/guns.html

11 Measurement

Measuring what we care about in the social world is often difficult. Attitudes, behaviors, and cultures do not easily lend themselves to being recorded succinctly as a column in a spreadsheet. Thus, a central concern with data in social science is measurement.

To distinguish what it is we truly care about from the things we are able to measure, we use the term **construct** to describe the concept or property we wish to study. By contrast, the data that ends up in our files is a variable—a term we've been using already throughout this book. For example, we can create a personality questionnaire to measure someone's extroversion, but there will always be a gap (measurement error) between the values that end up in our spreadsheets—the variable—and the "true" value of the construct extraversion—a complex personality trait that is difficult to precisely quantify. For complex constructs that defy easy measurement, an **operational definition** describes a particular approach to practically measuring the construct. The distinction between construct and variable is particularly pronounced in psychology (where many variables of interest are difficult to measure precisely), so literature drawing on that discipline is where you are most likely to encounter this terminology. By contrast, suppose we are interested in something relatively easy to measure, like someone's age. It is difficult to articulate a difference between the concept of age and the measured values of age, so the distinction between construct and variable is not particularly useful in this instance.

11.1 Validity and reliability

How do we evaluate whether a particular measurement approach is effective? We want measures that are **valid**, meaning that they (on average) reflect the underlying construct (a property known as **construct validity**). We also want measures to be **reliable**, meaning that they are precise and we get consistent results from the measurement approach.

There are many ways to evaluate validity, often identified as different types of validity. A full accounting is beyond the scope of this text, but two broadly-applicable examples are worth discussing. First, **face validity** refers to a qualitative judgement of whether the measurement approach appears reasonable. You can always ask yourself whether a measure makes sense, based on what you know about the topic being studied. Second, **criterion validity** refers to a measure exhibiting associations with other variables in expected ways. When we see that a

variable tracks with other variables that should be interrelated, that builds some confidence that we have not gone horribly astray in our attempts to measure a construct.

Reliability is usually evaluated by repeating measurement in some manner and then comparing how similar the results are across the different measurements. If a measure is highly reliable, the various measurements should give us similar results (unless there's reason to believe the true value of the construct has changed between measurement attempts). Various types of reliability scores can be calculated. While the details differ, they usually have a range of either 0 to 1 or -1 to 1, with 0 or -1 indicating no reliability and 1 indicating perfect reliability (equivalent scores from the different measurement attempts). Three common methods for estimating reliability are test-retest reliability, Cronbach's alpha, and inter-rater reliability.

Test-retest reliability involves administering a measure once and then repeating the measure, usually at a later date. In order for the test-retest approach to make sense, we generally need to be measuring a highly stable construct (at least during the period separating the two measurements). For example, personality refers to a stable set of characteristics (at least in theory), so test-retest reliability is often used to assess measures of personality. By contrast, emotional states are generally more transient, so finding that someone indicates a different emotional state at two different points in time does not indicate that the measurement approach is unreliable; the subject may simply be experiencing a different emotional state than last time they were measured.

Cronbach's alpha can be computed when multiple indicators are combined into an index that measures the construct of interest. The classic example is a survey with several items related to one construct (as in the measure of extraversion we have repeatedly referenced) or an exam with multiple problems. Cronbach's alpha reflects the *internal consistency* of the indicators used to form the index. In other words, it tells us how similar our various indicators are to one another. Conbach's alpha also increases—all else equal—as the number of indicators increases. So a 10-item index will have a higher Cronbach's alpha than a 3-item index, assuming the two indices have items that are equally internally consistent. The reason for this is that as the number of indicators increases, the idiosyncrasies associated with individual items matter less to the overall index (just as larger sample sizes result in less noisy estimates). Whether this property of indices implies that we should generally use long multi-item scales to measure complex psychological or behavioral constructs is a topic of debate among survey researchers.

Finally, **inter-rater reliability** can be computed when multiple sources are rating (or coding) the same material. For example, a study might rely on multiple research assistants to rate the level of charisma exhibited by a speaker, using a rubric that details specific tactics of charismatic speech that are to be counted. One can use a measure of inter-rater reliability to determine how similar the ratings are from the different research assistants. This requires that there is a sample (could be a subsample) of speeches that have each been rated by more than one person, so that direct comparisons of the scores can be made. If all raters give the same score to every speech, there will be perfect inter-rater reliability. If raters give highly inconsistent scorings of the same speech, inter-rater reliability will be low.

For all types of reliability, researchers often rely on "rules of thumb" about what threshold (e.g., 0.8) a reliability score must reach to constitute "good" or "acceptable" levels of reliability. Trying to identify meaningful thresholds for the entirety of the social sciences is perhaps a hopeless tasks, since different constructs and types of measures allow for different realistic levels of reliability to be achieved. Within a given field, there will probably be established norms regarding acceptable levels of reliability.

Validity and reliability are both important. However, because reliability is often easier to evaluate quantitatively, you may find that more space is devoted to discussions of reliability than validity in many social science journals. Some scholars even argue that the scientific norms associated with scrutinizing reliability have led survey researchers to unjustifiably sacrifice validity in their scale development in order to achieve reliability levels that are deemed sufficient.[1]

11.2 Scaling

Scaling refers to combining multiple indicates of a construct into a single variable called an index. The simplest scaling method involves taking the average (or sum) of the indicators. We call the result a **summative index**. While taking the average and taking the sum might seem like entirely distinct ways of creating an index, they are in some sense equivalent since each is a linear transformation of the other: divide the sum by the number of indicators, and you will have the average. Just as our results should not meaningfully change if we decide to measure something in inches instead of feet (Section 2.6), using a sum versus an average to construct an index will make not difference to our results so long as we remember to interpret our units correctly.

If the indicators don't have a common scale (or even if they do), it is often a good idea to first standardize the items before combining them into an index. Some scaling approaches will automatically do this in the background, but if you are creating a summative index you may need to make this transformation first before calculating a sum/average.

Factor analysis refers to various methods for scaling that involve calculating different weights to apply to the various indicators. By contrast, with a summative index we are effectively applying an equal weight to all indicators, making it so that all indicators contribute equally to the final index. By assigning different weights, we make some indicators more important than others. This makes conceptual sense if we believe that some indicators are more precise or offer more unique information about the true value of the construct. *Confirmatory factor analysis* (as opposed to exploratory factor analysis) requires that you specify a measurement model indicating how various indicators are linked to constructs (as well as other linkages

[1] Clifton, Jeremy D. W. 2020. "Managing validity versus reliability trade-offs in scale-building decisions." *Psychological Methods* 25(3): 259.

indicators may have to one another) and yields results that can be used as tests of whether the measurement model is plausible.

Principal component analysis (PCA, also called principal component factors or PCF) is a widely used technique that is often (mis)labled a type of factor analysis and accomplishes something similar, in that it creates an index based on calculating different weights for the indicators. Unlike confirmatory factor analysis, PCA does not require the user to map out a model of measurement. The basic intuition underlying PCA is that it selects values for weights in a way that maximizes the extent to which a common (latent) factor can explain the variation in the various indicators.

The factor loadings (or weights) from factor analysis or PCA will indicate how closely alligned each indicator is to the index. There are different ways in which these values can be reported, depending on the technique and what transformations might be applied. But generally speaking, loadings closer to 0 indicate less alignment of the indicator with the index. Negative loadings mean that an indicator is negatively associated with the index (e.g., an indicator of *introversion* should have a negative loading for an index of *extraversion*).

11.3 Measurement error

Measurement error usually distorts our ability to make valid estimates. An exception is that random measurement error in a dependent variable will not necessarily violate any regression assumptions since we can consider the measurement error to be part of the error term (so long as the measurement error conforms with the particular assumptions made about the error term). Unfortunately, measurement error often extends to our independent variables as well when we are examining data about the social world. This brings a serious source of concern regarding the validity of our estimate, including the validity of our inferential statistical results (confidence intervals and significance tests).

If we are only examining a bivariate relationship (e.g., how X relates to Y, without any control variables), then we can at least say that *random measurement error* in the independent variable should lead to **attenuation bias**, meaning that we will tend to underestimate the strength of an association. For example, if the actual correlation between two constructs is 0.6, attenuation bias means that we will systematically tend to get estimates that are smaller than this (e.g., 0.5 or 0.4). By *random* measurement error, I mean that the value of the variable's measurement error is unrelated to the true value of either construct (and is also unrelated any the measurement error in the other variable).

Unfortunately, as soon as we move to the world of multiple regression (to be covered more in Chapter 12), random measurement error in the independent variables can easily lead to inflated estimates of associations (meaning the strength of an association is overstated) or even systematically wrong-signed estimates (e.g., a negative instead of a positive association).

Generally speaking, it is difficult to correctly anticipate the direction of bias that might occur from measurement error (among independent variables) in the context of multiple regression.

Correlated measurement error generates similarly disruptive problems for estimation, even when looking at bivariate relationships. *Correlated* measurement error refers to errors in measurement that are correlated with underlying constructs or with errors in the measurement of other variables. For example, *common method variance* is a frequent source of potentially correlated measurement error in survey research. Suppose that we are using a survey of employees and want to estimate the association between one's work motivation and job performance. If we rely on self-reported survey scales (a "common method") to measure both variables, our variables will likely exhibit correlated measurement error. Respondents who think particularly highly of themselves (or wish to convey a positive image of themselves on a survey) are likely to overstate both their own motivation and their performance. They will have high values for both variables. Respondents with a more humble disposition will tend to report lower values for both variables. Thus, measurement error will likely push the association in a positive direction (high values of one variable paired with high values in the other variable, and low values paired with low values). This can lead to an association even if none exists in the underlying constructs.

Two main sets of tools exist that can create corrections for measurement errors. They emerge out of distinct traditions of statistical analysis emerging from the disciplines of psychology and economics. The psychology tradition has developed rather elaborate tools that utilize structural equation modeling (SEM) to estimate associations while accounting for measurement error. From the economic tradition, there is errors-in-variables regression, which allows for estimation of regression models that account for known error in the measurements of variables. Both sets of tools can be helpful for testing the sensitivity of findings to different assumptions about measurement error, but the tools are also somewhat limited in that they generally require strict assumptions about the nature of measurement error than cannot be fully tested.

12 Regression Models

Regression is the most important tool for statistical analysis in the social sciences, and we have already seen several examples of how regression is used, starting from Section 3.4. In this chapter, we will learn more about the assumptions that typically underlie our regression models as well as how we can think about using regression to test more complex relationships among variables than we have examined so far.

12.1 Regression Assumptions

There are many different articulations of the assumptions that underlie our typical linear regression models, with some authors providing longer lists than others. Here, we will focus on the list provided by Gelman, Hill, and Vehtari (2021), which has the benefit of being arranged in decreasing order of importance. The authors caution, though, that this list of assumptions is for *predictive* inferences; drawing *causal* conclusions requires additional considerations, as implied by the discussion of causality in Chapter 10.

1. Validity

Just as our list of Three Questions to Always Ask about Data prods us to start by asking "what is being measured?", the first regression assumption highlights the importance of valid measurement of variables (see discussion of validity in Chapter 11). With any statistical analysis (whether using regression or not), the data in the sample must validly measure whatever you are trying to understand, or else the results will be of no use. In their concept of validity, Gelman and colleagues also indicate the need to "include all relevant" independent variables and to have observations that fall within the realm of the phenomena of interest (e.g., a study of employee attitudes should use a sample that consists of employees).

2. Representativeness

The data should be appropriate for generalizing to the broader phenomena of interest (external validity). As a simple example, a representative sample from a population (as would be found in expectation under random sampling) meets this assumption. It is not always necessary to have a perfectly representative sample to draw valid conclusions about associations, so long

as the associations among variables are the same in the sample as in the population. For example, a sample that overrepresents young people could potentially yield accurate estimates of the association between exercise and happiness in the general population, so long as the exercise-happiness relationship plays out similarly among older and younger people. It is also the case that we are not always studying a well defined population. Thus, we sometimes need to interpret this assumption as indicating that the observations in the sample are representative instances of whatever it is we care to learn about (even if that phenomena of interest is not easily defined as a population).

3. Additivity and linearity

We use the term "linear regression" to refer to the standard regression model because of its linear form: the value of each independent variable is multiplied by a constant, and then these products are added up to form the predicted value of the dependent variable (together with the intercept). Predictions from a linear regression model will always follow this pattern of additivity and linearity. If we wish to create predictions that cannot be represented through a linear combination of independent variables, a linear regression is the wrong tool to use. Note, however, that sometimes relationships that are not strictly linear can still be approximated through a linear function. Linear relationships offer a simplicity that is not always apparent in other functional forms, so sometimes we may prefer the straightforward interpretability of linear regression results at the cost of the flexibility we might be able to achieve with other types of models. For example, if our primary concern is whether two variables generally exhibit a positive (versus negative) association and what the general strength of that association is, we may prefer a linear model of that association since it can provide a single number (a slope coefficient) that indicates direction and magnitude of association, even if this number oversimplifies a bit (as when there is some curvature in the true line describing their association).

Another important consideration to mention here (that we will explore in more detail later on in this chapter, and in the next) is that certain non-linear relationships can be modeled through linear regression, so long as they can be expressed by manipulating variables to create a linear function that represents these non-linearities. For example, the relationship between two variables need not follow a straight line if we transform the independent variable by squaring it, allowing for a prediction line to follow the shape of a parabola (for the original, untransformed variable).

If an independent variable is binary, the assumption of linearity is not practically restrictive. Since binary variables can only take on two different values (typically coded as 0 and 1), the coefficient associated with a binary variable will simply indicate how much to shift the prediction when going from one value to the other. The shape of the "line" connecting the two points is immaterial. Thus, linear regression can generally be considered "non-parametric" (meaning minimal assumptions are imposed) when studying only binary independent variables. In such cases, we can think of regression as simply using an equation to compare means across

groups, as previously demonstrated in Chapter 4 (and in Section 9.1.3.2, where we noted linear regression yields equivalent results to ANOVA).

4. Independence of errors

The "errors" referred to in this list of assumptions come from the error term in a regression model, as introduced in Section 6.2. This fourth assumption implies that each observation in the sample represents a truly unique datapoint compared to all the others, at least when it comes to the value of the error term. Because the social world is full of interconnections, we often see violations of this assumption. Suppose, for example, that customer attitudes are measured using a survey of 1000 customers collected at 20 different restaurants. Though there is a sample size of 1000, each of the 20 restaurants may have its own peculiaries that shape customer attitudes in particular ways. Thus, the errors of prediction for the individuals may be interrelated (rather than fully independent) within each restaurant.[1]

Fortunately, there are several techniques that can help us adjust our regression models to accommodate non-independence of errors, so long as we can accurately identify the structure(s) by which observations' errors are interrelated.[2] The simplest structure is when we can group observations into mutually exclusive "clusters," as in the case of the restaurant example above. There are multiple techniques that can adjust our regression estimates for such clustering, with the simplest being to make adjustments to our standard error estimates using one of several techniques known as "cluster robust standard errors." Most statistical software packages will easily allow you to implement estimation of such standard errors. More advanced (and flexible) approaches to dealing with clustered observations can be found using tools from multi-level modeling.[3]

Another setting where we often adjust for violations of this assumption is when we are analyzing panel (or time series) data. Since the same units (e.g., individuals or organizations) are being observed multiple times within a dataset, observations are not expected to be fully independent

[1] One way to conceptualize potential consequences of violating of this assumption is that you are effectively overstating the sample size: since observations are not truly independent, each observation adds less than a full unit (of new information) to the degrees of freedom.

[2] Some additional techniques beyond those mentioned in the main text are spatial regression models, time series and panel regression techniques, and fixed effects models.

[3] Take, for example, a study of whether employee job satisfaction is associated with changes in the size of an organization's budget. Suppose a survey is conducted with employees of several dozen organizations, yielding thousands of individual-level survey responses. This seems to provide a very large sample, but the independent variable—size of the organization's budget—is measured at the level of the organization, not at the level of the individual. And only a few dozen organizations were included in the sample. This is a classic example of multi-level data (since job satisfaction is measured at the individual level while budget size is measured at the organizational level). With multi-level data, it is difficult to define the sample size because the sample size differs depending on the variable: individual-level variables will have many more distinct observations than organizational-level variables. If we run a regression at the individual level, we risk dramatically overstating the precision of our estimates due to acting as though our sample size is much larger than it really is (for the independent variable).

of one another (e.g., an individual with above-average satisfaction in one time period will likely continue to be fairly satisfied in the following period). A variety of techniques have been developed to address concerns associated with the non-independence of errors when working with panel (or time series) data, and effectively working with such data will typically require serious study of such techniques.

5. Equal variance of errors[4]

This assumption, known as **homoskedasticity**, indicates that the variance around the regression line is the same for all values of the independent variable(s). A clear violation of this assumption is shown in Figure 12.1. Notice that the predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs are not very good. In other words, the points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students are not.

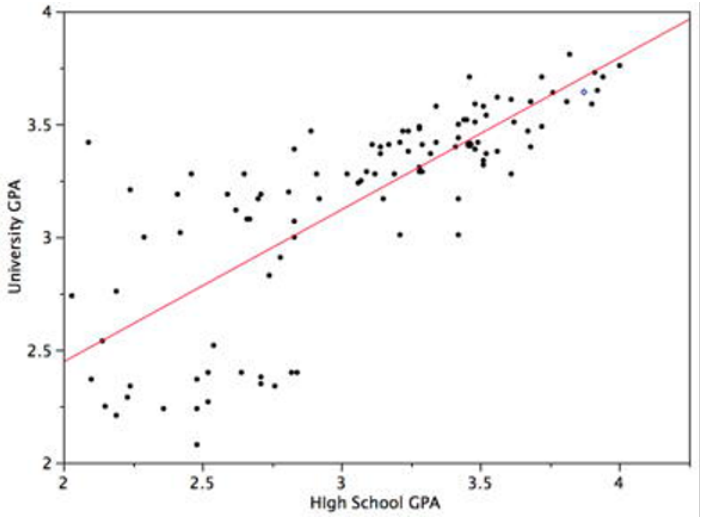

Figure 12.1: University GPA as a function of High School GPA.

When errors have unequal variance, we call this **heteroskedasticity**. A common solution (easily implemented in most statistical software) is to calculate standard errors that are robust to heteroskedasticity. Some analysts will default to always using heteroskedastic-robust standard errors, although such corrections do not always work well in small samples.[5]

[4]The first paragraph of this subsection is adapted from David M. Lane. "Inferential Statistics for b and r." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/inferent

[5]Rajh-Weber, H., Huber, S.E. & Arendasy, M. A practice-oriented guide to statistical inference in linear modeling for non-normal or heteroskedastic error distributions. *Behav Res* **57**, 338 (2025). https://doi.org/10.3758/s13428-025-02801-4

6. Normality of errors

This assumption is considered to the be least important. Still, regression results can sometimes become unreliable or imprecise when the distribution of errors has excessive outliers that are associated with certain non-normal distributions. This is particularly true when working with small samples. Thus, it can be useful to check variables for outliers, as well as examining the distribution of the residuals (estimated values of the error term in a regression model), which can easily be obtained in statistical software. Robust versions of regression are readily available in most statistical software and should be less vulnerable to problems created by violations of the normality assumption.[6]

12.2 Multiple Regression[7]

In simple linear regression, a dependent variable is predicted from one independent variable. In multiple regression, the dependent variable is predicted by two or more variables. For example, in the SAT case study we've used several times already to illustrate regression, you might want to predict a student's university grade point average on the basis of their High-School GPA (HSGPA) and their total SAT score (verbal + math). The basic idea is to find a linear combination[8] of HSGPA and SAT that best predicts University GPA (UGPA). That is, the problem is to find the values of β_1 and β_2 in the equation shown below that give the best predictions of UGPA. As in the case of simple linear regression, we define the best predictions as the predictions that minimize the squared errors of prediction (the "least squares" criterion).

$$\widehat{UGPA} = \alpha + \beta_1 HSGPA + \beta_2 SAT$$

where $\widehat{UGPA}$ is the predicted value of University GPA and α is a constant. For these data, the best prediction equation is shown below:

[6] Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy*, *98*, 19-38. https://doi.org/10.1016/j.brat.2017.05.013

Baissa, D. K., & Rainey, C. (2020). When BLUE is not best: non-normal errors and the linear model. *Political Science Research and Methods*, *8*(1), 136-148. https://doi.org/10.1017/psrm.2018.34

[7] The initial material in this section (up until Section 12.2.1) is adapted from Rudy Guerra and David M. Lane. "Introduction to Multiple Regression." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/multiple_regression.html

[8] A linear combination of variables is a way of creating a new variable by combining other variables. A linear combination is one in which each variable is multiplied by a coefficient and the products are summed. For example, if

$$Y = 3X_1 + 2X_2 + .5X_3$$

then Y is a linear combination of the variables X_1, X_2, and X_3.

$$\widehat{UGPA} = 0.540 + 0.541 \times HSGPA + 0.008 \times SAT \tag{12.1}$$

In other words, to compute the prediction of a student's University GPA, you add up (a) 0.540, (b) their High-School GPA multiplied by 0.541, and (c) their SAT multiplied by 0.008.

For comparison purposes, here is the regression equation from the simple regression discussed in Section 3.4.

$$\widehat{UGPA} = 1.097 + 0.675 \times HSGPA \tag{12.2}$$

Table 12.1 shows the data and predictions for the first five students in the dataset based on the multiple regression (Equation 12.1).

Table 12.1: Data and Predictions

$HSGPA$	SAT	$\widehat{UGPA}$
3.45	1232	3.38
2.78	1070	2.89
2.52	1086	2.76
3.67	1287	3.55
3.24	1130	3.19

The values of β (β_1 and β_2) are called "regression slope coefficients."

The multiple correlation (R) is equal to the correlation between the predicted scores and the actual scores. In this example, it is the correlation between $\widehat{UGPA}$ and $UGPA$, which turns out to be 0.79. That is, R = 0.79. Note that R will never be negative since if there are negative correlations between the predictor variables and the criterion, the regression coefficients will be negative so that the correlation between the predicted and actual scores will be positive.

Interpretation of Regression Coefficients

A regression coefficient in multiple regression is the slope of the linear relationship between the criterion variable and the part of a predictor variable that is independent of all other predictor variables. There are multiple ways to explain this computation, with additional descriptions provided in the appendix. As one approach, the regression coefficient for HSGPA can be computed by first predicting HSGPA from SAT and saving the errors of prediction (the differences between $HSGPA$ and $\widehat{HSGPA}$). These errors of prediction are called "residuals" since they are what is left over in HSGPA after the predictions from SAT are subtracted, and represent the part of HSGPA that is independent of SAT. These residuals are referred to as HSGPA.SAT, which means they are the residuals in HSGPA after having been predicted by SAT. The correlation between HSGPA.SAT and SAT is necessarily 0.

The final step in computing the regression coefficient is to find the slope of the relationship between these residuals and UGPA. This slope is the regression coefficient for HSGPA. The following equation is used to predict HSGPA from SAT:

$$\widehat{HSGPA} = -1.314 + 0.0036 \times SAT$$

The residuals are then computed as:

$$HSGPA.SAT = HSGPA - \widehat{HSGPA}$$

The linear regression equation for the prediction of UGPA by the residuals is

$$\widehat{UGPA} = 3.173 + 0.541 \times HSGPA.SAT$$

Notice that the slope (0.541) is the same value given previously for the estimate of β_1 in the multiple regression equation.

This means that the regression coefficient for HSGPA is the slope of the relationship between the dependent variable and the part of HSGPA that is independent of (uncorrelated with) the other independent variables. It represents the change in the prediction for the dependent variable associated with a change of one in the independent variable when all other independent variables are held constant. Since the regression coefficient for HSGPA is 0.54, this means that, holding SAT constant, a change of one in HSGPA is associated with a change of 0.54 in $\widehat{UGPA}$. If two students had the same SAT and differed in HSGPA by 2, then you would predict they would differ in UGPA by $(2)(0.54) = 1.08$. Similarly, if they differed by 0.5, then you would predict they would differ by $(0.50)(0.54) = 0.27$.

The slope of the relationship between the dependent variable and the part of an independent variable that is unique from (independent of) other independent variables is its partial slope. Thus, the regression coefficient of 0.541 for HSGPA and the regression coefficient of 0.008 for

SAT are partial slopes. Each partial slope represents the relationship between the independent variable and the dependent variable holding constant all of the other independent variables.

It is difficult to compare the coefficients for different variables directly because they are measured on different scales. A difference of 1 in HSGPA is a fairly large difference, whereas a difference of 1 on the SAT is negligible. Therefore, it can be advantageous to transform the variables so that they are on the same scale. The most straightforward approach is to standardize the variables (see Section 2.6.1) so that they each have a standard deviation of 1. A regression coefficient for standardized variables is called a "standardized coefficient" or "beta coefficient." For these data, the standardized coefficients are 0.625 and 0.198. These values represent the change in the prediction for the dependent variable (in standard deviations) associated with a change of one standard deviation on an independent variable (holding constant the value(s) on the other independent variable(s)). Clearly, a change of one standard deviation on HSGPA is associated with a larger difference than a change of one standard deviation of SAT. In practical terms, this means that if you know a student's HSGPA, knowing the student's SAT does not aid the prediction of UGPA much. However, if you do not know the student's HSGPA, his or her SAT can aid in the prediction since the standardized coefficient in the simple regression predicting UGPA from SAT is 0.68. For comparison purposes, the standardized coefficient in the simple regression predicting UGPA from HSGPA is 0.78. As is typically the case, the partial slopes are smaller than the slopes in simple regression.

12.2.1 Deciding Which Independent Variables to Include[9]

It is a bit hard to generalize regarding the criteria for deciding which varaibles to include as independent variables, because it depends on the research question posed and whether the goal is to describe general patterns of association, to identify a predictive relationship, or to identify a causal or possibly causal relationship.

To help guide our discussion of variable selection, we will distinguish between *key independent variables of interest* (those that the analyst is particularly interested in learning about) and **control variables**. We will represent the former using X and the latter as Z. When we add an independent variable Z to a regression not because we are particularly interested in estimating the association of Z with the dependent variable Y but instead because we think including Z in the regression will yield better estimates for how X is associated with Y, we often refer to Z as a control variable. Control variables are not different from independent variables, as far as the statistical estimation is concerned. They are merely different labels that signal a difference in the researcher's substantive interest in the slope coefficients for these variables (e.g., control variables are probably not the subjects of any hypotheses).

Obviously, any variable X of substantive interest should be included in a regression, although it is sometimes beneficial to include various X variables one at a time if one is not interested in describing how each one relates to to Y independent of the other X variables.

[9]The remainder of the chapter was written by Nathan Favero.

Where the interest lies (at least to some extent) in considering causal relationships among variables, you should generally include potential confounders to the X-Y relationship as control variables (additional independent variables) in a regression.

You might also consider including as a control variable any factor that you believe will be a strong cause of the dependent variable, so long as this factor is not itself caused by X. Adding such variables will generally improve the precision of our estimates (making standard errors smaller).

When there is an interest in causal relationships, it is generally best to avoid adding as control variables anything that may be caused by X. One reason is that when X causes Z and Z causes Y, Z is a mediator, and therefore Z is one route (or mechanism) through which X may affect Y. Thus, by controlling for (or pulling out) one route through which X may affect Y, we are distorting our ability to observe the total effect of X on Y.

For estimations of causal effects, a more complete assessment of which variables should and should not be included in a regression can be faciliated through the use of Directed Acyclic Graphs (DAGs), an increasingly popular tool for applied researchers that is beyond the scope of what can be covered in this text.

12.2.1.1 More on Mediating Relationships

When a variable is a mediator, it is one pathway by which an original independent variable affects the dependent variable. Mediators provide an interesting case that can highlight how there are sometimes multiple valid ways to select a list of independent variables, with different selections yielding different insights.

Let us consider the case of gender and wages. In studies of the gender wage gap, one might use a sample of people in the workforce to test for differences in wages between men and women. The question of whether to control for various employee characteristics—such as professional background, expertise, and industry—turns out to be a highly controversial one. One the one hand, gender was determined at birth (at least for most of the population), and employee characteristics tend to result from either processes that occur after birth (and may be affected by gender) or things like family background that shouldn't be correlated with gender (since we can assume in most contexts that sex is randomly determined). Thus, employee characteristics that are associated with gender can be viewed as potential mediators of the gender-pay relationship. If we are trying to understand the net-total effect of gender on pay, we should probably estimate the gender-pay association without controlling for employee characteristics. But the net-total effect is not necessarily the only thing we care about. For example, if we are trying to learn something about potential gender discrimination by employers, we might want to control for factors that we believe were determined prior to an employee's interaction with the firm. In other words, we are now interested in a particular subset of pathways by which gender may be associated with pay, while ignoring other pathways that are not related to firm discrimination. This example illustrates how direct associations and partial associations can both add value

to our understanding. In fact, a common way to empirically study relationships in which mediation is believed to exist is to run more than one regressions—one that does not control for the mediator (to estimate a total effect of the independent variable) and one that does include the mediator as a control (in part to estimate the "direct" effect of the independent variable, independent of its indirect effect through the mediator). To better describe the mediated path, one can also run a regression with the mediator as the dependent variable, in order to discern the link between the independent variable and the mediator.

While we often find phenomena in the world that we believe can be described through a mediating relationship, it is very difficult to comprehensively test the causal claims implied by a mediating relationship. As such, most mediating relationships cannot be firmly established empirically with a single study.[10] Thus, while mediation is important to consider when mapping out possible relationships, we should be somewhat modest in terms of our expectations for being able to easily test mediating relationships. If we are willing to *assume* that a mediating relationship exists, there are regression we can run (such as those described in the prior paragraph) to describe the nature of the linkages in this relationship, but it is much more difficult to evaluate quantitatively whether we have correctly identified the causal ordering in a mediating relationship.

12.3 Modeling Non-linear Relationships

One way to model a non-linear relationship is to square the values of an independent variable, and then include both the original (non-squared) version of the variable as well as the squared variable in the regression. Doing so with the SAT variable, we find our estimates generate the following equation (with university GPA as the dependent variable):

$$\widehat{UGPA} = -11.331 + 0.021 \times SAT - 0.0000074 \times SAT^2$$

This results in the curved line shown in Figure 12.2, which better tracks the data in the scatterplot than a straight line would. Adding a squared line allows for the shape of the line to follow the shape of a parabola. If we want to allow for a second bend in the line (of a certain sort), we could add a cubed term. Higher-level polynomials can also be added.

Now, let's consider a **moderating** relationship, meaning that two independent variable interact such that they each alter the association of the other with the dependent variable. Especially when looking at data that is well-suited for drawing causal conclusions, we might also describe tools for modeling moderation as checking for **heterogeneous effects**. One of the simplest ways to look for potential moderation is to divide a sample into subsamples according to the value of

[10] Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *The Annals of the American Academy of Political and Social Science*, *628*(1), 200-208.

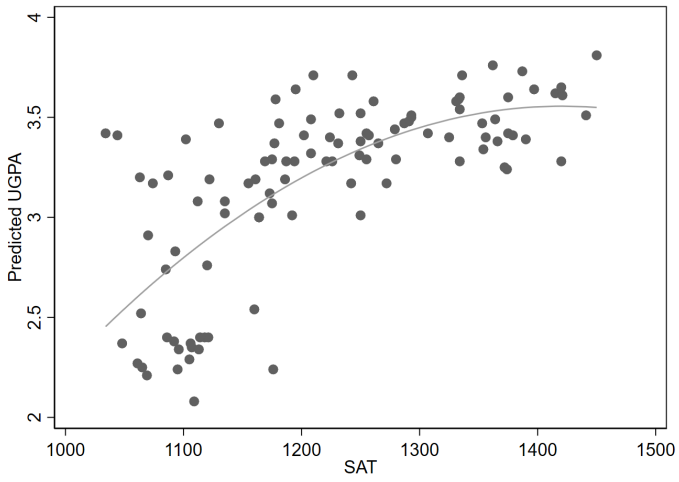

Figure 12.2: Adding a squared term to a regression allows the regression line to curve

a moderating variable; we can then estimate the association of the independent and dependent variable in each subsample and see whether results differ notably across subsamples.

Let's check whether the SAT-UGPA relationship appears to differ depending on the value of high school GPA. We split the sample into two subsamples, with one subsample (the "low HSGPA" subsample) containing all students with a high school GPA at or below the full sample's median. The other subsample ("high HSGPA") contains students with a value of high school GPA above the median. Note that there are many different ways we might split the sample, such as using the mean instead of the median, or creating three subsamples (low, medium, high).

Estimating a regression line among our low HSGPA subsample yields a slope of 0.0029:

$$\widehat{UGPA} = -0.448 + 0.0029 \times SAT$$

Estimated among students in the high HSGPA subsample, the slope shrinks to just 0.00062:

$$\widehat{UGPA} = 2.663 + 0.00062 \times SAT$$

Figure 12.3 shows these two distinct lines, along with the underlying subsamples from which they are estimated.

It appears that SAT scores have a stronger association with university GPA (a steeper slope) when high school GPA is low.

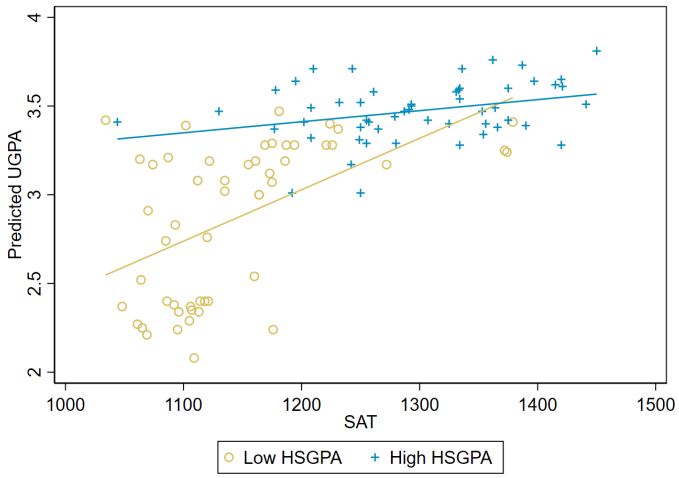

Figure 12.3: Estimating two regressions separately by subsample

We can also model moderation with a single regression for the whole sample using an **interaction term**. We create an interaction term by multiplying two independent variables together. In this example, we should multiply SAT by high school GPA. We can then run a regression where we have three independent variables: the original SAT variable, the original HSGPA variable, and the interaction term. The regression line is estimated to be:

$$\widehat{UGPA} = -3.516 + 1.780 \times HSGPA + 0.0043 \times SAT - 0.0011 \times HSGPA \times SAT$$

When we have nonlinear relationships, graphical depictions are typically helpful for demonstrating our results. Using statistical software, we can create a graph that shows us how the predicted value of UGPA changes depending on SAT and certain HSGPA values that we pick out for illustrative purposes. In this case, I chose values of 3.7 (approximately the 90th percentile), 2.3 (approximately the 10th percentile, and 3.0 (half way between the 10th and 90th percentiles).

From Figure 12.4, we again see evidence that for higher levels of HSGPA, the relationship of SAT with university GPA gets weaker. However, an important caveat to this finding is that the coefficient for the interaction term is not statistically significant at the traditional .05 alpha level (p=0.069), so there is not strong statistical evidence that the slope does in fact vary depending on HSGPA. In other words, the changes we see in the slope in Figure 12.4 could easily be a coincidence, given the imprecision of our estimate. A larger sample could help us better assess whether the moderating relationship that we seem to observe is more than a statistical anomaly.

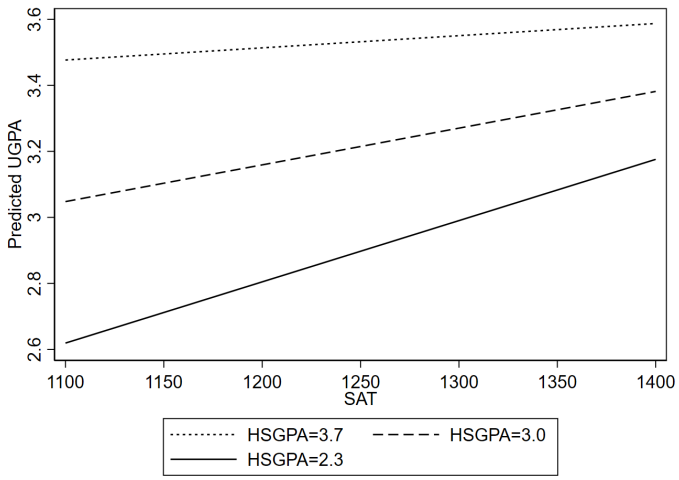

Figure 12.4: Visualization of a moderating relationship, estimated with an interaction term

More generally, it typically requires a large sample size to generate precise estimates of non-linear relationships.[11]

While the approaches described in this section are widely employed in the social sciences, caution is warranted whenever modeling non-linearities with quantitative independent variables (interaction effects involving one or more binary independent variables are more straightforward to work with). Adding interaction or polynomial terms to linear regression equations is a somewhat inflexible way of dealing with potential non-linearities; we can easily go wrong if there are relatively small deviations from the functional form we assume in the regression equation we choose to use for our model. Put differently, it is hard to be confidence we are not violating regression assumptions 1 and 3 in a consequential way when relying on these simplistic methods to describe the details of a non-linear relationship. Fortunately, it is not too difficult to find more flexible approaches that can help validate findings of non-linearities in which we are interested, although doing so may require learning tools (e.g., nonparametric regression) beyond the scope of what is covered in this text.[12]

[11] It is difficult to say precisesly how large, but see the following resources for more concrete guidance:

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect [blog post]. https://statmodeling.stat.columbia.edu/2018/03/15/need16/

Sommet, N., et al. (2023). How many participants do I need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction. Advances in Methods and Practices in Psychological Science, 6(3), 25152459231178728.

Baranger, D. A., et al. (2023). Tutorial: Power analyses for interaction effects in cross-sectional regressions. Advances in Methods and Practices in Psychological Science, 6(3), 25152459231187531.

[12] Simonsohn U. Interacting With Curves: How to Validly Test and Probe Interactions in the Real (Nonlinear) World. *Advances in Methods and Practices in Psychological Science.* 2024;7(1). doi:10.1177/25152459231207787

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative

Chapter 12 Appendix: More Explanation of Partial Slopes/Associations

The simple linear regression equation can be written as:

$$\hat{y_i} = \alpha + \beta x_i \tag{12.3}$$

We estimate the value of the slope coefficient (using least squares) as:

$$\hat{\beta} = \frac{Cov(x, y)}{Var(x)} \tag{12.4}$$

When we have two independent variables (x and z), the corresponding equations are:

$$\hat{y_i} = \alpha + \beta_1 x_i + \beta_2 z_i \tag{12.5}$$

$$\hat{\beta_1} = \frac{Cov(x, y)Var(z) - Cov(z, y)Cov(x, z)}{Var(x)Var(z) - Cov(x, z)^2} \tag{12.6}$$

Notice that in the numerator of Equation 12.6, we begin with the covariance between x and y (which we also scale by multiplying by the variance of z). Then, to avoid a spurious relationship between x and y that might stem from z affecting both x and y, we subtract out the covariance of z and y times the covariance of z and x. Conceptually, we throw out the joint variation among all three variables—the portion of the overlap between x and y that also reflects overlap between z and y and between z and x.

One way to think about the partial associations obtained through multiple regression is to use a Venn diagram to illustrate how covariance among three (or more) variables is handled.[13] This is not a perfectly precise representation of how multiple regression works, but it can serve as a helpful tool for understanding the basic intuition. Think of overlap in circles as representing common variation or covariation.

---

interaction models? Simple tools to improve empirical practice. *Political Analysis*, *27*(2), 163-192. https://doi.org/10.1017/pan.2018.46

Simonsohn U. Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions. *Advances in Methods and Practices in Psychological Science*. 2018;1(4):538-555. doi:10.1177/2515245918805755

[13] Kennedy (2002) and Cohen and Cohen (1975) have been instrumental in developing this Ballentine diagram approach to explaining multiple regression.

Kennedy, P. (2008). *A guide to econometrics*. Malden, MA: Blackwell Publishing.

Cohen, J., & Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

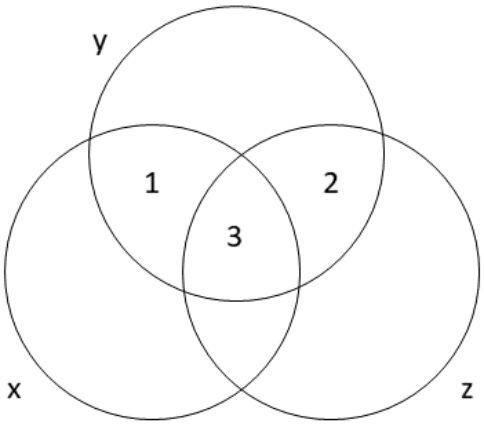

Figure 12.5

When estimating partial slopes (multiple regression), the coefficient estimate for x is based on area 1. The coefficient estimate for z is based on area 2. Shared variation among all variables can't be easily attributed to x or z, so area 3 isn't be used to estimate either coefficient in multiple regression. If we estimate the coefficient of x using simple linear regression (without including z in the regression model), we might get a misleading estimate of the relationship between x and y since we'll be using areas 1 and 3 to estimate the coefficient for x. If z is a confounding variable, then including area 3 when estimating a slope for x may be undesirable since this portion of the covariation between x and y is due to a common cause (z) rather than a direct link between x and y.

But what about when the two independent variables x and z are uncorrelated? If the covariance between them is zero, then their variation is already independent (at least linearly independent). And therefore, finding the independent association of each one with y is equivalent to just finding the association with y. Using the Ballantine visual, there is no overlapping part 3 to subtract out, as seen in Figure 12.6.

Looking to Equation 12.6, we can try subbing in the value 0 for $Cov(x, z)$:

$$\begin{aligned}\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)(0)}{Var(x)Var(z) - (0)^2} \\ &= \frac{Cov(x,y)Var(z)}{Var(x)Var(z)} = \frac{Cov(x,y)}{Var(x)}\end{aligned}$$

This yields the same result as Equation 12.4, indicating that we obtain the same slope coefficient estimate for x regardless of whether we use simple regression or control for z in the case where x and z are perfectly uncorrelated. Note that this result only applies to the slope's point estimate; its standard error and confidence interval will likely differ.

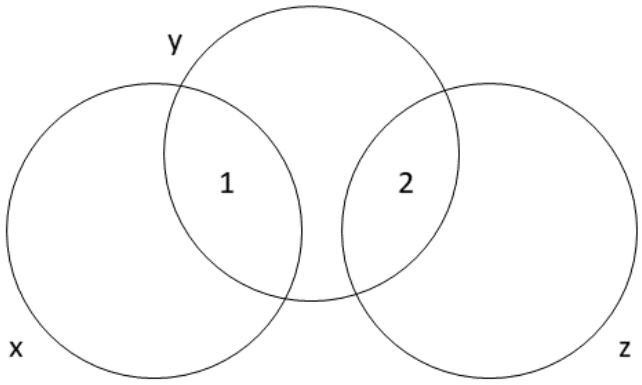

Figure 12.6

We can also use the language of conditional expected values to demonstrate the difference between the bivariate (zero-order) associations described by simple regression and the partial associations found in multiple regression. Regression can be conceptualized as a model of a conditional expected value (conditional mean), where the predicted value (for the dependent variable) is an expected value and we are conditioning on the independent variables in the regression.

Returning to the example from Section 12.2, we can indicate the difference in expected university GPA between two students where the only thing we know about them is that one (student A) has a high school GPA of 3.5 and the other (student B) has a high school GPA of 2.5. If high school GPA is the only piece of information we have access to, we should use the simple linear regression Equation 12.2 to determine this expected difference:

$$\begin{aligned}&\mathbb{E}[UGPA|HSGPA = 3.5] - \mathbb{E}[UGPA|HSGPA = 2.5] \\ &= [1.097 + 0.675 \times (3.5)] - [1.097 + 0.675 \times (2.5)] = 0.675\end{aligned}$$

Now, suppose that we also know the two students' SAT scores. If they are both 1100, that changes the predictive effect of a 1-point difference in high school GPA because would would have guessed that student A (with the 3.5 high school GPA) had a higher SAT score than student B (with a 2.5 high school GPA) since these two variables are positively correlated. But with partial slopes, we are considering how changing the value of one variable alters our prediction *while holding all other variables constant*. The partial slope for $HSGPA$ in Equation 12.1 indicates the difference in the expected value of university GPA when the $HSGPA$ differ by one but the SAT scores are equal:

$$\mathbb{E}[UGPA|HSGPA = 3.5, SAT = 1100] - \mathbb{E}[UGPA|HSGPA = 2.5, SAT = 1100]$$

$$= [0.540 + 0.541 \times (3.5) + 0.008 \times (1100)] - [0.540 + 0.541 \times (2.5) + 0.008 \times (1100)]$$
$$= 0.541$$

The key point to emphasize here is that the difference in the expected value of university GPA associated with a 1-point difference in high school GPA changes depending on whether we are conditioning on solely high school GPA (yielding a difference of 0.675) or if we are also conditioning on SAT (yielding a difference of 0.541). That is because holding SAT constant is not what we would normally expect when observing a difference in high school GPA, since high school GPA and SAT are (positively) correlated.

13 Practical points

[content to be added]

14 Teaching Resources

If you're using this text, I'd love to know. You can fill out this brief form (https://forms.gle/qBUFdb4vEuDUkzBu6), where you can also sign up to receive emails when I post updated versions or related materials.

14.1 Stata/R Labs

I've created a number of Stata and R labs that I use when I teach. There are also some handouts, including a couple covering Excel. Such resources are available here: https://github.com/favero-nate/minus-the-math/tree/main/labs

14.2 Lecture Slides/Videos

While they do not directly correspond to this version of the text, there are some (Stata-based) lecture slides and videos I created to use alongside this book when I teach. They are currently available here: https://github.com/favero-nate/minus-the-math/tree/main/lecture_slides

14.3 A Few More Details about What's Unique in this Text

1. There is a bit of Stata code in one appendix (Chapter 4), but otherwise all examples are presented apart from any statistical software package.
2. The treatment of probability theory skips much of the typical material in favor of discussing probabilistic modeling, which I believe is far more relevant to quantitative social science.
3. In addition to traditional statistical inference ideas that describe estimating parameters of population from a sample, I emphasize that we often use inference to draw conclusions about counterfactuals. This approach is informed by Kass (2011): https://doi.org/10.1214%2F10-STS337

15 Change Log

PDFs of past versions are currently available at https://github.com/favero-nate/minus-the-math/tree/main/past_versions

- Version 1.3 updates: New material on multiple regression (end of Chapter 3, Section 5.2.1, Section 7.3.3, and Section 8.7). Expanded discussion of confidence intervals (**?@sec-estimation**), including new Section 5.2.2 on interpreting confidence intervals. Expanded discussion of ANOVA (end of Section 9.1) and of contingency tables (Section 9.2). Notation updated in line with conventions: regression parameters are redone, and $\bar{X}$ is now used for the sample mean and n for sample size. Slight extension of section on the standard normal distribution (**?@sec-the-standard-normal-distribution**). Section on degrees of freedom moved to an appendix (end of Chapter 7). Various formatting updates (book was recreated using Quarto) and minor (mostly non-substantive) edits throughout.
- Version 1.2 updates: The discussion of transforming variables now appears in Chapter 2 (rather than Chapter 3).