Stat 155 Notes

Macalester Statistics Faculty (Heggeseth, Myint)

Updated: 2019-09-07

Contents

Preface

This book contains notes for STAT 155 at Macalester College. It contains definitions, data examples, and R code explanations that provide a foundation for the activities we will do in class.

This is a living document, which will get updated throughout the semester.

When you find typos or have clarifying questions, please email bheggese@macalester.edu as soon as possible.

Figure 1: Creative Commons License

0.1 Acknowledgments

These notes would not be possible without the **bookdown package** and the many open-course R packages used throughout the text. I thank the developers of these packages for making this work possible.

This structure of this books is strongly influenced by the bookdown text written by Yihui Xie.

Chapter 1

Data Collection and Quality

We live in a world where data touch nearly every aspect of our lives: health care, online shopping, transportation, entertainment. From search engines to satellite images, from cell phones to credit cards, current technology can produce data faster than we can analyze them.

This course is the beginning of your journey into the field of Statistics, a discipline whose main goal is to extract information and meaning from data. We do this is by visually exploring the data and building models to try to explain observed variability. First, we will take some time to think about the collection of data and what factors might make data more or less reliable.

1.1 What is Data?

Data is *anything* that contains information. We typically think of data being stored in spreadsheets, but it can come in many other formats such as images or collections of text (whether 280 character tweets or fictional novels).

For example, we can take the pixels of digital images or text from one of the State of the Union addresses and transform them into a tidy, rectangular format. Below, see the top 10 words (that are not 'the', 'and', etc.) from the most recent State of the Union address given in the United States.

```
## Joining, by = "word"
```

```
## Selecting by n
```

```
## # A tibble: 11 x 2
##    word          n
##    <chr>     <int>
##  1 american     31
##  2 people       26
```

```
##  3 americans    24
##  4 tonight      23
##  5 america      18
##  6 country      15
##  7 tax          14
##  8 congress     13
##  9 home         12
## 10 family       10
## 11 world        10
```

Tidy data is a table in which

- Each row of a rectangular table corresponds to an **observation** or **case** (e.g. person, classroom, country, image, word)
- Each column correspond to a characteristic or feature or **variable** for those cases (e.g. age, average classroom grade, average county income, intensity of red pixels, number of times the word "together" is used)

Variables can be either **categorical** or **quantitative** variables.

- **Categorical variable:** *A characteristic with values that are names of categories; the names of categories could be numbers such as with zipcodes. If the categories have a natural ordering, it may be called an ordinal variable, but we won't be distinguishing between different types of categorical variables in this class.*
- **Quantitative variable:** *A characteristic with measured numerical values with units.*

Note: Any quantitative variable can be converted into a categorical variable by creating categories defined by intervals or bins of values.

The following graphic from the book R for Data Science, by Garrett Grolemund and Hadley Wickham illustrates the features of tidy data.

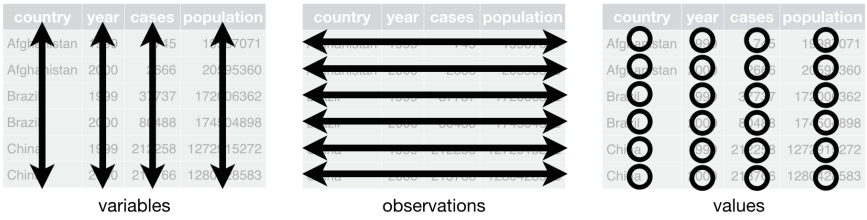


Figure 1.1: The components of a tidy dataset. Chapter 12 of R for Data Science

The transformation process from raw data to a tidy data format is often called **feature extraction** and is not a short or easy task. In this introductory course,

we will typically work with data that are already in a tidy format.

Cases are often referred to as the **units of analysis**. As data analysts, it is important for us to consider what to use as the unit of analysis when we have information, say, on both individual presidents and their speeches Do we want to understand matters at the individual or the speech level? Answers to these questions will depend on the context and the research questions.

1.2 Data Context

For any data set, you should always ask yourself a few questions to provide vital **context** about that data set.

- **Who is in the data set?** What is the observational unit or *case*? How did they end up in the data set? Were they selected randomly or were they in a particular location a particular time?
- **What is being measured or recorded on each case?** What are the characteristics, features, or *variables* that were collected?
- **Where were they collected?** In one location? Multiple locations?
- **When was the data collected?** One point in time? Over time? If data quality degrades over time (e.g. lab specimens), we should be concerned.
- **How were they collected?** What instruments and methods used for measurement? What questions were asked and how? Online survey? By phone? In person?
- **Why were they collected?** For profit? For academic research? Are there conflicts of interest?
- **Who collected this data?** An agency, a consortium of researchers, an individual researcher?

Thinking about this data context informs us how we analyze the data, what conclusions we can draw, and whether we can generalize our conclusions to a larger population.

Many of these data context questions also hint at general considerations for threats to data quality. Threats to data quality generally arise through **sampling**, **information bias**, and **study design.**

1.3 Sampling

When we study a phenomenon, we generally care about making a conclusion that applies to some **target population of interest** (e.g. all likely voters in the U.S., all eligible voters in the U.S., college students in Minnesota, etc.). However, we cannot feasibly collect data on that entire population (this is called a **census** and is very expensive to complete) due to financial and time constraints, so we collect a **sample** of individuals. We want our sample to be **representative** of the target population in that we want our sample to resemble the target

population in the characteristics we are studying.

> **Reflect:** How is representativeness affected by our research question? Can a sample be representative for one goal but not another?

When our method of selecting a sample is flawed, **sampling bias** can result, and our sample is unrepresentative of the target population. We need to be aware of how this tends to happen, and how can we avoid it.

It is first helpful to define the term **sampling frame**. A sampling frame is the complete list of individuals/units in the target population. For example, it could be a spreadsheet listing every college student that studies in Minnesota.

1.3.1 Sampling Bias

The following are common ways that sampling bias can arise, and they all share the feature that a sampling frame is *NOT* used:

- **Convenience Sampling:** *Individuals that make up a convenience sample are easy to contact or to reach (e.g. you stand on a street corner and ask passerbys to answer a few questions). The people sampled will likely be systematically different than the target population.*
- **Self-Selection and Volunteer Sampling:** *Individuals that make up a sample self-select or volunteer to be in a sample (e.g. product reviews on Amazon, individuals that call in to radio shows, blood donors, etc.). They are likely to be systematically different than the target population.*

One result of using these sampling techniques is that we can get **undercoverage** in the sample. This happens when some groups of the population are inadequately represented in the sample due to the sampling procedure. A famous example in United States history is the 1936 Literary Digest poll that completely mispredicted the presidential election. The magazine predicted a strong victory for Alfred Landon, but Franklin Delano Roosevelt ended up winning the election by a substantial margin. The survey relied on a convenience sample, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent and leaned to the right politically (they favored Landon).

If we do not have a complete sampling frame, then we have no control over what units enter the sample because we do not even have a complete list of the units that *could* be sampled. Imagine that our target population is like a pot of soup, these forms of sampling are similar to scooping only the bits of soup that float to the top of the pot without stirring.

1.3.2 Random Sampling

With a sampling frame, we can do better and hopefully avoid sampling bias by using randomization. In our soup metaphor, this amounts to mixing the soup thoroughly and dipping our spoon in random locations.

These strategies are called **probability sampling** strategies or, more colloquially, **random sampling** strategies. In probability sampling, each unit in the sampling frame has a known, nonzero probability of being selected, and the sampling is performed with some chance device (e.g. coin flipping, random number generation).

Some probability sampling techniques include:

Simple Random Sampling: *Each unit in the sampling frame has the same chance of being chosen and individuals are selected without replacement (once they have been chosen, they cannot be chosen again). With this strategy, every sample of a given size is equally likely to arise.*

Stratified Sampling: *The units in the sampling frame are first divided into categories/strata (e.g. age categories). Simple random sampling is then performed within each category/stratum. Why do this? Just by chance, simple random sampling might oversample young individuals. Stratifying by age first, then performing simple random sampling in these strata ensures a desired age distribution in the sample. With this strategy, you may be able to increase the precision of the estimates.*

Cluster Sampling: *Sometimes a sampling frame is more readily available for clusters of units rather than the units themselves. For example, a sampling frame of all hospitals in Minnesota might be more readily available than a sampling frame of all Minnesota hospital patients in a given time frame. In cluster sampling, the initial clusters are sampled with a probability sampling method (like simple random sampling or stratified sampling). All units in the sampled clusters may be chosen, or if sampling frames can be obtained for the sampled clusters, probability sampling is performed within the cluster. This strategy should only be used when a full sampling frame is unavailable or it is economically justified as this procedure generally provides less precision than the other two strategies.*

1.3.3 Nonresponse bias

Even with a random sampling method, our sample can still be unrepresentative if units in our sample do not choose to participate after they are selected. For example, if the communication method is via e-mail, individuals who do not read our e-mail may be nonresponders. If those individuals who don't participate are systematically different than those that do, this type of nonresponse bias is called **unit nonresponse bias**.

Let's say that an individual opens up our e-mail survey. They may answer the first few questions but grow weary and skip the last questions. If those

individuals who answer some but don't answer other questions are systematically different than those that do in their responses, this type of nonresponse bias is called **item nonresponse bias**.

1.4 Information bias

Lastly, separate from sampling, bias can arise in how we record or measure observations. Some of these biases are types of cognitive biases that and others are statistical biases. Here are a few types of information biases that can occur:

- **Response bias/Self-report bias/Social desirability bias:** *Bias occurs when the recorded response does not accurately represent the true value for the individual due to wording of the question, ordering of the questions, format of response, or to increase social desirability. Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes or characteristics (e.g. weight, income, alcohol consumption, mental illness) or illegal activities in a survey, particularly if the survey is completed in an interview setting and/or the results are not confidential. Additionally, this could arise if the possible responses do not allow for accurate reporting (e.g. gender identity, race).*
- **Recall bias:** *People often unintentionally make mistakes in remembering details about the past. If the study design is retrospective in that it requires units to rely on their memory, we may get bias in the information collected.*
- **Measurement error:** *Technologies that measure variables of interest may not always be accurate and human calibration of those instruments may be off as well.*

There are many other forms of bias, many of which are studied by cognitive and social pscyhologists. When designing questionnaires and surveys, it is important to keep these in mind. A researcher must first decide what specific information they want to collect and then figure out the best way to accurately collect that.

1.5 Study Design

Once we have a sample selected, data can be collected in one of two general study designs:

1. **Observational Study:** *Data is collected in such a way such that the researcher **does not** manipulate or intervene in the characteristics of the individuals. Researchers observe or record characteristics of the sample through direct measurement or through a questionnaire or survey.*

2. **Experiment:** *Data is collected in such a way such that the researcher **does** manipulate or intervene in characteristics of the individuals by randomly assigning individuals to treatment or control groups. Researchers then record characteristics of the individuals in the sample within the treatment*

and control groups.

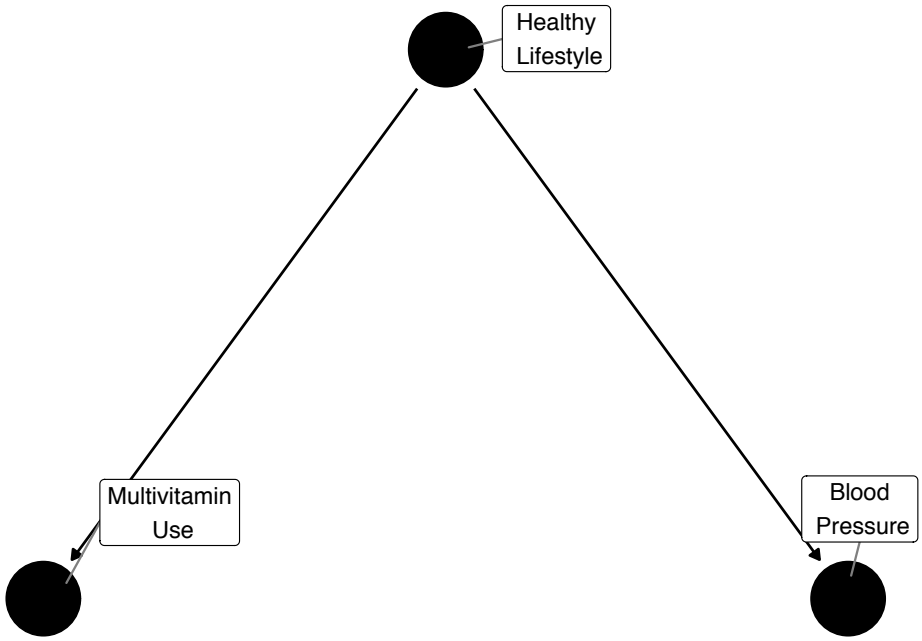
The main reason for doing an experiment is to estimate a relationship between a treatment and a response.

For example, imagine that we want to know if taking a daily multivitamin reduces systolic blood pressure. If we performed an observational study, we would select a sample (hopefully a representative one) from a population of interest and then ask whether an individual takes a daily multivitamin and then measure their blood pressure. *Would this data provide enough evidence to conclude that vitamin use causes a reduction in blood pressure?*

Individuals that take daily multivitamins may be more health-conscious, eating more fruits and vegetables and exercising more, which may be related to blood pressure. The diet and exercise would be acting as **confounding variables**, making it impossible to say for certain if vitamin use has a direct impact on blood pressure.

Confounding Variables: *A third variable that is a common cause for the "treatment" (e.g. multivitamin) and the "response" (e.g. blood pressure). In this case, a healthy lifestyle could be a variable that confounds the relationship between multivitamin use and blood pressure.*

```
confounder_triangle(y= "Blood\n Pressure",
                    z = "Healthy\n Lifestyle",
                    x= "Multivitamin\n Use") %>%
ggdag( text = FALSE, use_labels = "label") +
  theme_dag_blank()
```



Reflect: For example, say we note that on days with higher ice cream sales, there are typically a higher number of pool drownings. What variable could be confounding the relationship between ice cream sales and pool drownings?

In an experiment, we "manipulate" the characteristics for an individual by randomly assigning them to a treatment group. This random assignment is intended to break the relationship between any common causes and the treatment so as to try to reduce the impact of confounding. It is impossible to entirely remove the possibility of confounding, but the random assignment to a treatment helps. (Note: things can get complicated if individuals don't comply with the treatment such as take the multivitamin every single day. . .)

1.6 Causal Inference

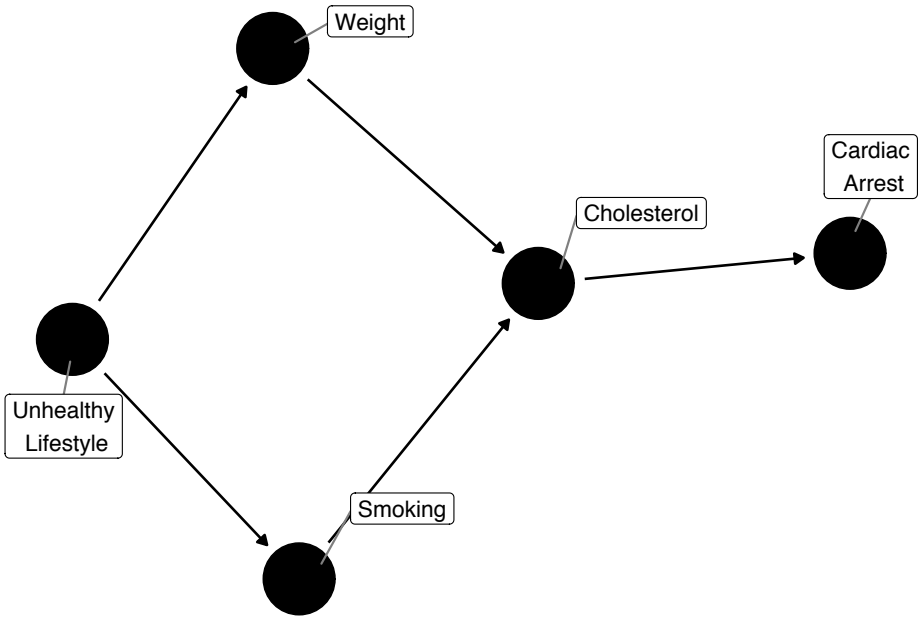
Causal inference is the process of making a conclusion about direct cause and effect between a "treatment" and a "response". It is difficult to make causal inferences/statements based on data from an observational study due to the possible presence of confounding variables.

There is a whole area of statistics dedicated to methods that attempt to overcome the potential confounding. A full overview of causal inference is beyond the scope of this course, but a first good step in this direction is to consider a **causal model**, which is a representation of the causal mechanism by which data were

generated. We typically visualize these models with a graphical structure called a **directed acyclic graph or DAG** for short. In a DAG, we have circles or **nodes** that represented variables and arrow or **directed edges** that indicate the causal pathway.

If we were interested in the causal effect of smoking on cardiac arrest, we might draw the following causal model (using previous research to inform our graph).

```
smoking_ca_dag <- dagify(cardiacarrest ~ cholesterol,
       cholesterol ~ smoking + weight,
       smoking ~ unhealthy,
       weight ~ unhealthy,
       labels = c("cardiacarrest" = "Cardiac\n Arrest",
                  "smoking" = "Smoking",
                  "cholesterol" = "Cholesterol",
                  "unhealthy" = "Unhealthy\n Lifestyle",
                  "weight" = "Weight"),
       latent = "unhealthy",
       exposure = "smoking",
       outcome = "cardiacarrest")

ggdag(smoking_ca_dag, text = FALSE, use_labels = "label") +
  theme_dag_blank()
```



Below we discuss the three basic DAG structures.

- **Chain**: *A chain includes three nodes and two edges, with one directed into*

and one edge directed out of the middle variable. In the above example, we see a few chains. There is a chain that suggests an unhealthy lifestyle leads to smoking which directly impacts cholesterol levels. Another chain suggests an unhealthy lifestyle direclty impacts human weight which impacts cholesterol levels. By our DAG, there are two causal pathways directly impacts cholesterol levels.

- **Fork**: *A fork includes three nodes with two arrows emanating from the middle variable.* The only fork in the example above involves unhealthy lifestyle (middle variable), weight, and smoking. In this case, we are assuming that an unhealthy lifestyle is a **common cause** of both weight and smoking. Thus, the lifestyle could confound the relationship between smoking and weight in that smoking and weight may look to be associated with each other because they have a common cause.
- **Colliders**: *A collider occurs when one node receives edges from two other nodes.* In the example above, smoking and weight have a **common effect**, cholesterol.

These structures are important to understand because they can help guide our analysis and modeling. Assuming the causal model is correct (which is might not be...), we have a better sense of how smoking might be related to cardiac arrest.

- There is a **direct path** from smoking through cholesterol to cardiac arrest. Smokers may have a higher cholesterol level and thus a higher risk of cardiac arrest.
- There is an **indirect (backdoor) path** from smoking to cardiac arrest through weight. Smokers also have an unhealty lifestyle which means they are likely to have a higher weight leading to a higher cholesterol level and thus a higher risk of cardiac arrest.

If we only want to estimate the direct relationship between smoking and cardiac arrest, we need to block the impact the indirect path. There are many ways to do this that we'll talk about later (stratification, including variable in model, etc.)

If you want more information about causal inference, take the Causal Inference course or check out the "Causal Inference in Statistics: A Primer" by Judea Pearl, Madelyn Glymour, Nicholas P. Jewell (Prof. Heggeseth has a copy in her office).

1.7 Ethical Considerations

Ethics play a very important role in study design and data collection procedures, especially when humans and animals are the observational units. In the United States, the Belmont Report is the main federal document that provides the "Ethical Principles and Guidelines for the Protection of Human Subjects of

Research". The three fundamental ethical principles for using any human subjects for research are:

1. **Respect for persons**: This principal is about protecting the autonomy of all people and treating them with courtesy and respect and allowing for informed consent. Researchers must be truthful and conduct no deception;
2. **Beneficence**: This principal is the philosophy of "do no harm" while maximizing benefits for the research project and minimizing risks to the research subjects; and
3. **Justice**: This principal is about ensuring reasonable, non-exploitative, and well-considered procedures are administered fairly — the fair distribution of costs and benefits to potential research participants — and equally.

For a brief, limited history of ethical regulation in human research, see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3593469/. A few key moments in international history are

- Nuremberg Code (1948) in response to medical experiments in Germany performed without consent
- Declaration of Helskinki (originated in 1964 and frequently revised) established by World Medical Association.
- The Belmont Report written in response to the Tuskegee Syphilis Study (1932 - 1972).
- Common Rule (1981) is regulatory policy which all U.S. government-funded research and nearly all U.S. academic institutions must abide.

Macalester College has an Institutional Review Board (IRB) that oversees research at Macalester that includes human participants (see: https://www.macalester.edu/irb/).

Throughout this class, we are going to stop and think about the ethical considerations of the many parts of statistical practice, ranging from data collection to model prediction.

Ethics are the norms or standards for conduct that distinguish between right and wrong. In particular, we are going to consider the ethics of

- How the data are collected
- Random assignment to treatments
- Data storage
- Data privacy
- Data use
- Choice of sample data used for predictive modeling
- Use of predictive modeling

We are going to pay extra attention to negative consequences of the above that may disproportionately impact marginalized groups of people.

Throughout the semester, you will be asked to think about answers to the

question: "What are the ethical considerations for this data set/analysis?" Like in other disciplines, the choices we make will be biased by our life experiences. Throughout this class, let us be mindful in increasing our awareness of the real consequences caused by choices we make in Statistics.

1.8 Major Takeaways

1. Any observed data is a sample of a larger population or phenomenon. But you need to consider which population. Is it the population of interest to you? If not, then why?
2. Sampling strategies impact what type of generalizations we can make about a population. Bias occurs when there are systematic differences between observed sample data and the true population of interest due to the sampling process.
3. The data we collect many not accurately reflect the truth due to information biases caused by the data collection mechanism (instrument or survey).
4. Study design impacts what type of conclusions we can make. Confounding variables prevent us from easily making cause and effect conclusions.
5. To make a causal conclusion, you must think carefully about all potential causal pathways between a treatment and response and deal with them appropriately.
6. We need to be aware of the real and ethical consequences of our choices when working with data and building statistical models.

Chapter 2

Visualizing Data

The first step in any data analysis is to visually explore your data.

There is a saying that "a picture is worth a 1000 words." In making visualizations, our goal is to quickly and easily get a better understanding of the variability, structure, and relationships that exist in the data.

Here we will cover the standard appropriate graphics for univariate variation and bivariate relationships. We will also cover techniques for multivariate relationships (3 or more variables). The choice of the graphic depends on the type(s) of variable(s): quantitative or categorical. So the first step is to think about the variables you are interested in visualizing and determining whether they are quantitative or categorical.

For each type of variable, we use a real data set to illustrate the visualizations.

2.1 Good Visualization Principles

Before we discuss the standard graphics, let's lay out the basic design principles for good data visualizations.

1. **Show the Data** This may be self-explanatory, but make sure that the data is the focus and driver of the visualization.
2. **Avoid Distorting the Data** Avoid 3D charts as the added dimension distorts the comparison. The area in a graph should equal the magnitude of the data it is representing.
3. **Simplify** In 1983, Edward Tufte said that "A large share of ink on a graphic should present data-information, the ink changing as the data change. Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented." Remove

any unnecessary "ink" that does not assist in the presentation of the data. Remove distractions.

4. **Facilitate Comparisons** In order to explain variation, we want the graphics to facilitate comparisons between groups. The design should make it easier to compare between groups rather than harder.
5. **Use Contrast** Humans have developed to seek out visual contrast. When choosing colors and annotation, strive for more contrast in luminance (white to dark) to make it easier for everyone to visually perceive.
6. **Use Color Appropriately** Think about your audience. A small proportion of the population is color-blind; try printing it in grayscale to see if the color palette is still effective. Also, every culture has different associations with colors; ask others for feedback on color choices. Neuroscience research has shown that humans are more sensitive to red and yellow, so those are good colors to use for highlighting key points.
7. **Annotate Appropriately** Informative text is crucial for providing data context. Make sure to use informative axis labels and titles. It may be worth adding text to explain extreme outliers.

For examples of good data visualizations in the news and discussion around them, check out the New York Times column "What's Going on in This Graph?".

2.2 Brief Intro to R

Throughout this class, we use R and RStudio to visualize, analyze, and model real data. To straighten out which is which: R is the name of the language itself (syntax, words, etc.) and RStudio is a convenient software interface that you'll interact with on the computer.

While you'll be learning about and using R throughout the course, this is not a course on R. Our focus will be on data and statistical modelling. We will be using R and RStudio as tools to help us get information from data.

2.2.1 Basic Syntax

For this class, we will have data that we want to pass to a function that performs a particular operation (does something cool) on our data. Thus, we'll pass **inputs** as arguments to a **function**:

```
FunctionName(argument1 = a1, argument2 = a2,..., argumentk = ak)
```

Note the `FunctionName` and the use of parantheses. Inside the parantheses, the argument name (`argument1`) goes first and the value you are passing as an input is after = (`a1`).

We may want to save the **output** of the function by assigning it a name using the assignment operator, `<-`:

```
OutputName <- FunctionName(argument1 = a1, argument2 = a2,..., argumentk = ak)
```

R allows us to be lazy and not include the argument name as long as we provide the input in the correct order:

```
OutputName <- FunctionName(a1, a2,..., ak)
```

We can also **nest** functions by first performing one operation and then passing that as an input into another function. In the code below, `Function1()` would first run with the input `data` and create some output that is then passed as the first input in `Function2()`. So R evaluates functions from the inside-out.

```
Function2(Function1(data))
```

As we go through real examples below, notice the names of the functions that we use. The name comes right before `(` and the inputs we pass in right after `(`.

Additionally, we are going to use a shortcut that makes our code more readable. It is called a **pipe** and looks like `%>%`. What this does is pass the output on its left as the first argument to the function on the right. The following two sections of code do exactly the same thing but the second is easier to read. For this code, we take data and summarize the variable height and then take the mean of the heights.

```
summarize(data, mean(height))
```

```
data %>%
  summarize(mean(height))
```

There is so much more we could say about functions in R, but we will stop here for now.

With this in mind, we'll point to external references if you'd like to go deeper in your understanding of R as a programming language throughout this class.

To get a broad sense of R, you can work through R primers (https://rstudio.cloud/learn/primers) in RStudio Cloud in addition to any coursework and use the R cheatsheets available (https://rstudio.cloud/learn/cheat-sheets).

2.3 Anatomy of a ggplot command

To learn more about visualizing data with the ggplot2 R package, see Hadley Wickham's textbook.

In this course, we'll largely construct visualizations using the `ggplot()` function from the `ggplot2` R package. NOTE: `gg` is short for "grammar of graphics". Plots constructed from the `ggplot()` function are constructed in layers, and the syntax used to create plots is meant to reflect this layered construction. As you read through the rest of this chapter, pay attention to how the syntax generally follows this structure:

```
data %>%
    ggplot(aes(x = X_AXIS_VARIABLE, y = Y_AXIS_VARIABLE)) +
    VISUAL_LAYER1 +
    VISUAL_LAYER2 +
    VISUAL_LAYER3 + ...
```

We pass the aesthetic mapping from the data set to the plot with `aes()`. The visual layers are features such as points, lines, and panels. We'll introduce these soon. The `+`'s allow us to add layers to build up a plot (note this is not the pipe!).

> **Reflect:** What are the function names in the example above? There are only two as it is written right now.

2.4 One Categorical Variable

First, we consider survey data of the electoral registrar in Whickham in the UK (Source: Appleton et al 1996). A survey was conducted in 1972-1974 to study heart disease and thyroid disease and a few baseline characteristics were collected: age and smoking status. 20 years later, a follow-up was done to check on mortality status (alive/dead).

Let's first consider the age distribution of this sample. Age, depending on how it is measured, could act as a quantitative variable or categorical variable. In this case, age is recorded as a quantitative variable because it is recorded to the nearest year. But, for illustrative purposes, let's create a categorical variable by separating age into intervals.

Distribution: *the way something is spread out (the way in which values vary).*

```
#Note: anything to the left of a hashtag is a comment and is not evaluated as R code

library(dplyr) #load the dplyr package
library(ggplot2) #load the ggplot2 package
data(Whickham) #load the data set from Whickham R package

Whickham <- Whickham %>%
    mutate(ageCat = cut(age, 4)) #Create a new categorical variable with 4 categories

head(Whickham)
```

```
##   outcome smoker age      ageCat
## 1   Alive    Yes  23 (17.9,34.5]
## 2   Alive    Yes  18 (17.9,34.5]
## 3    Dead    Yes  71 (67.5,84.1]
```

```
## 4   Alive     No  67   (51,67.5]
## 5   Alive     No  64   (51,67.5]
## 6   Alive    Yes  38   (34.5,51]
```

Reflect: What do you lose when you convert a quantitative variable to a categorical variable? What do you gain?

2.4.1 Bar Plot

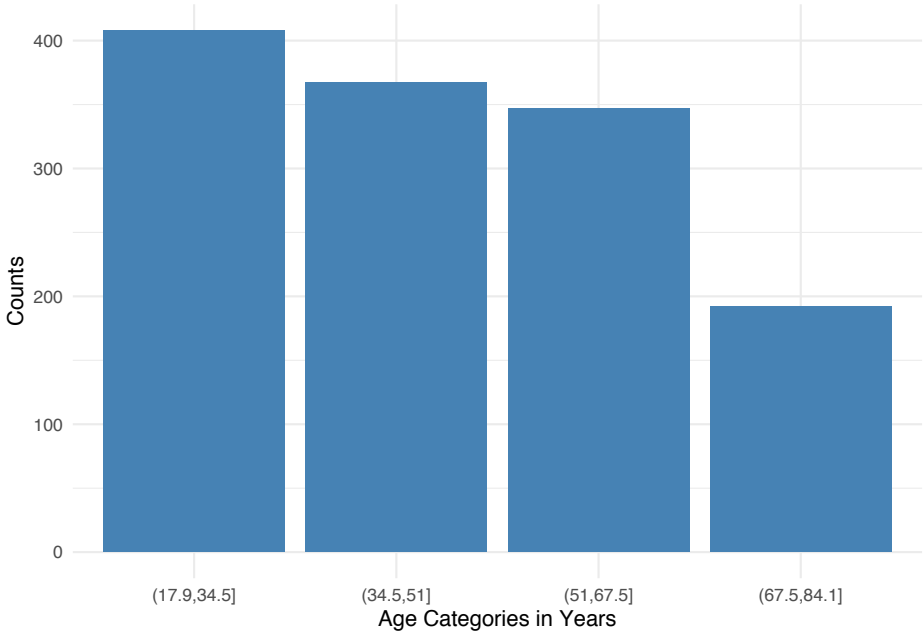
One of the best ways to show the distribution of one categorical variable is with a bar plot. For a bar plot,

- The **height of the bars** is the only part that encodes the data (width is meaningless).
- The height can either represent the **frequency** (count of cases) or the **relative frequency** (proportion of cases).

```
## Numerical summary (frequency and relative frequency)
Whickham %>%
    count(ageCat) %>%
    mutate(relfreq = n / sum(n))
```

```
## # A tibble: 4 x 3
##   ageCat          n relfreq
##   <fct>       <int>   <dbl>
## 1 (17.9,34.5]   408   0.311
## 2 (34.5,51]     367   0.279
## 3 (51,67.5]     347   0.264
## 4 (67.5,84.1]   192   0.146
```

```
## Graphical summary (bar plot)
Whickham %>%
    ggplot(aes(x = ageCat)) +
    geom_bar(fill="steelblue") +
    xlab('Age Categories in Years') +
    ylab('Counts') +
    theme_minimal()
```



Reflect: What do you notice? What do you wonder?

2.4.2 Pie Chart

Pie charts are only useful if you have 2 to 3 possible categories and you want to show relative group sizes.

This is the best use for a pie chart:

We are intentionally not showing you how to make a pie chart because a bar chart is a better choice.

Here is a good summary of why many people strongly dislike pie charts: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6. Keep in mind Visualization Principle #4: Facilitate Comparisons. We are much better at comparing heights of bars than areas of slices of a pie chart.

2.5 Two Categorical Variables

Now, let's consider two other variables in the same Whickham data set. What is the relationship between the 20-year mortality outcome and smoking status at the beginning of the study?

2.5.1 Side by Side Bar Plot

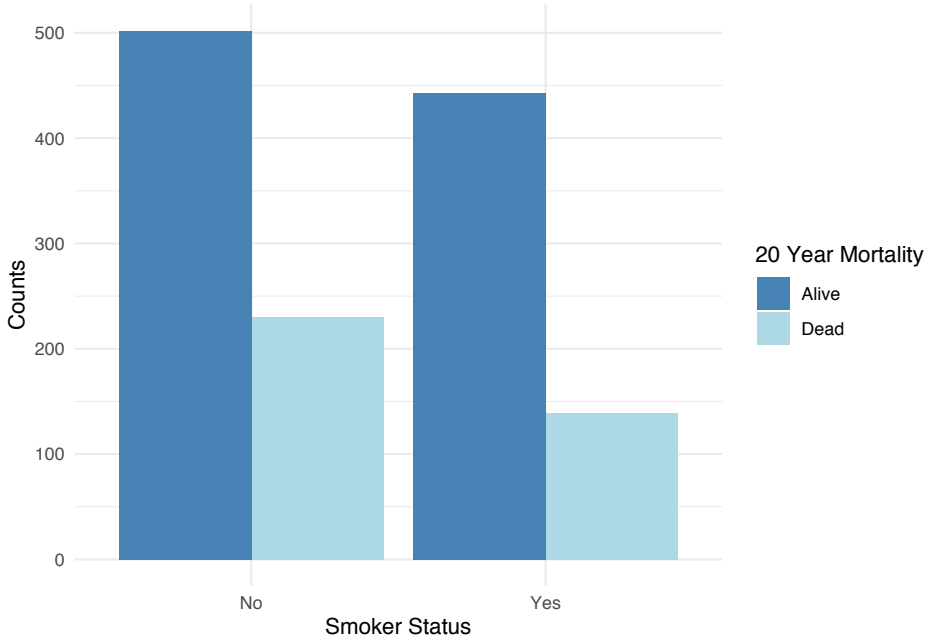
There are a few options for visualizing the relationship between two categorical variables. One option is to use a bar plot and add bars for different categories next to each other, called a **side-by-side bar plot**. For these plots,

- The **height of the bars** shows the frequency of the categories within subsets.

```
## Numerical summary (frequency and overall relative frequency)
Whickham %>%
  count(outcome, smoker) %>%
  mutate(relfreq = n / sum(n))
```

```
## # A tibble: 4 x 4
##   outcome smoker     n relfreq
##   <fct>   <fct>  <int>   <dbl>
## 1 Alive   No       502   0.382
## 2 Alive   Yes      443   0.337
## 3 Dead    No       230   0.175
## 4 Dead    Yes      139   0.106
```

```
## Graphical summary (side-by-side bar plot)
Whickham %>%
  ggplot(aes(x = smoker, fill = outcome)) +
  geom_bar(position = position_dodge()) +
  xlab('Smoker Status') +
  ylab('Counts') +
  scale_fill_manual('20 Year Mortality', values = c("steelblue", "lightblue")) +
  theme_minimal()
```



Reflect: What additional information do you gain by considering smoking status?

2.5.2 Stacked Bar Plot

Another way to show the same data is by stacking the bars on top of each other with a category. For a **stacked bar plot**,

- The **height** of the entire bar shows the **marginal distribution** (frequency of the X variable, ignoring the other variable).
- The **relative heights** show **conditional distributions** (frequencies within subsets), but it is hard to compare distributions between bars because the overall heights differ.
- The **widths** of the bars have no meaning.

```
## Numerical summary (conditional distribution - conditioning on outcome)
Whickham %>%
    count(outcome, smoker) %>%
    group_by(outcome) %>%
    mutate(relfreq = n / sum(n))
```

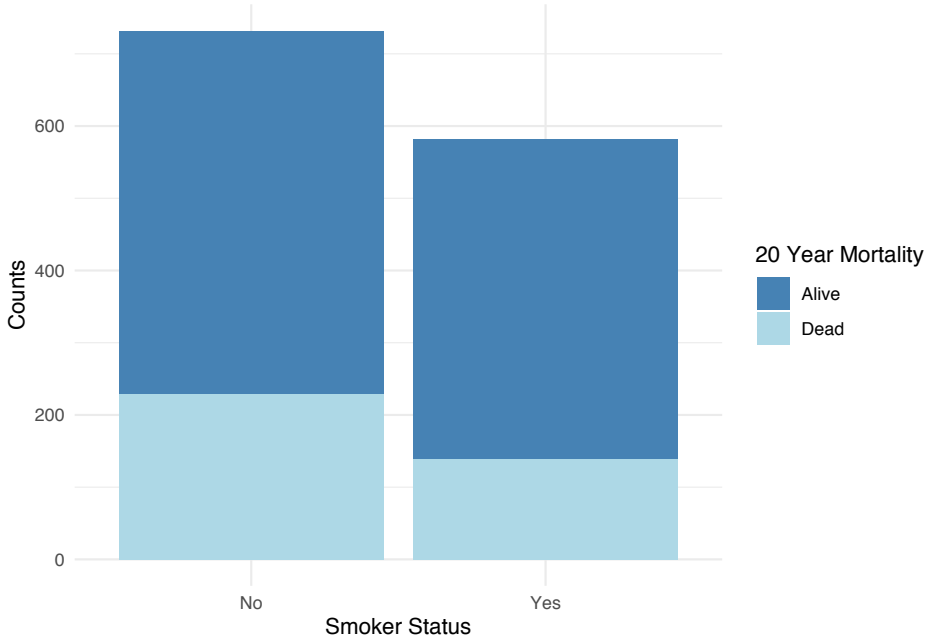
```
## # A tibble: 4 x 4
## # Groups:   outcome [2]
##   outcome smoker     n relfreq
##   <fct>   <fct>  <int>   <dbl>
```

```
## 1 Alive   No       502   0.531
## 2 Alive   Yes      443   0.469
## 3 Dead    No       230   0.623
## 4 Dead    Yes      139   0.377
```

```
## Numerical summary (conditional distribution - conditioning on smoker)
Whickham %>%
    count(outcome, smoker) %>%
    group_by(smoker) %>%
    mutate(relfreq = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   smoker [2]
##   outcome smoker     n relfreq
##   <fct>   <fct>  <int>   <dbl>
## 1 Alive   No       502   0.686
## 2 Alive   Yes      443   0.761
## 3 Dead    No       230   0.314
## 4 Dead    Yes      139   0.239
```

```
## Graphical summary (stacked bar plot)
Whickham %>%
    ggplot(aes(x = smoker, fill = outcome)) +
    geom_bar() +
    xlab('Smoker Status') +
    ylab('Counts') +
    scale_fill_manual('20 Year Mortality',values=c("steelblue", "lightblue")) +
    theme_minimal()
```



Reflect: What information is highlighted when you stack the bars as compared to having them side-by-side?

2.5.3 Stacked Bar Plot (Relative Frequencies)

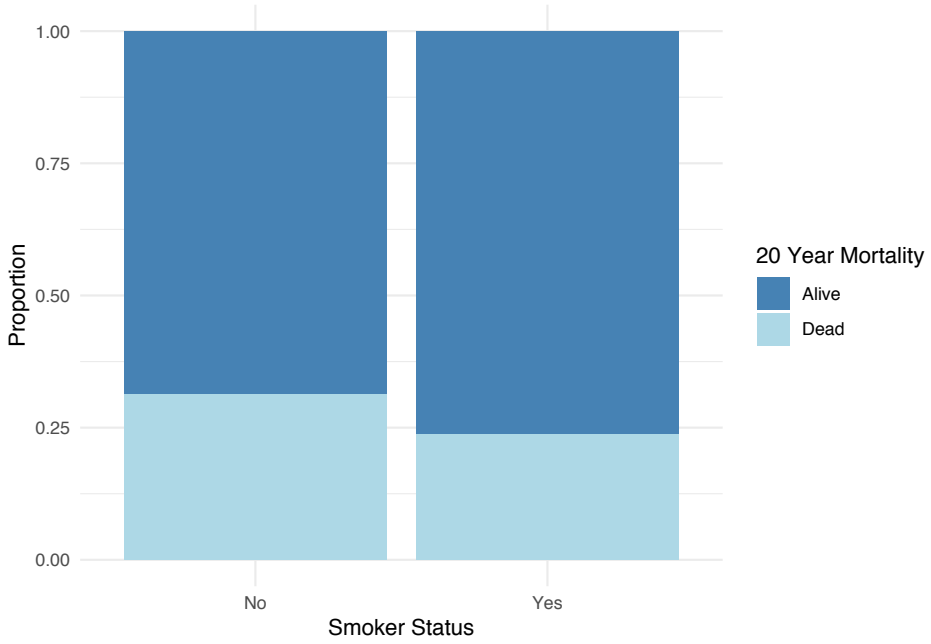
We can adjust the stacked bar plot to make the heights the same, so that you can compare conditional distributions. For a **stacked bar plot based on proportions** (also called a **proportional bar plot**),

- The **relative heights** show **conditional distributions** (relative frequencies within subsets).
- The **widths** of the bars have no meaning.

The code below computes the conditional distributions manually first (fractions of outcomes within the two smoking groups). Then these calculations are plotted directly.

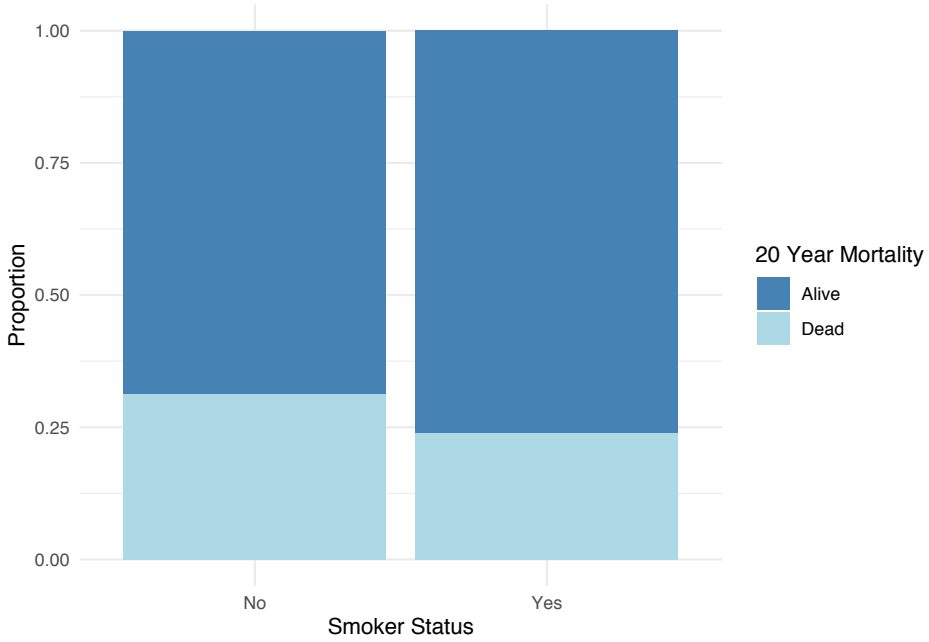
```
Whickham %>%
    count(outcome, smoker) %>%
    group_by(smoker) %>%
    mutate(relfreq = n / sum(n)) %>%
    ggplot(aes(x = smoker, y = relfreq, fill = outcome)) +
    geom_bar(stat = 'identity') +
    xlab('Smoker Status') +
    ylab('Proportion') +
```

```
    scale_fill_manual('20 Year Mortality', values = c("steelblue", "lightblue")) +
    theme_minimal()
```



Another way to make a proportional bar plot is to use the `position = "fill"`

```
Whickham %>%
    ggplot(aes(x = smoker, fill = outcome)) +
    geom_bar(position = "fill") +
    xlab('Smoker Status') +
    ylab('Proportion') +
    scale_fill_manual('20 Year Mortality', values = c("steelblue", "lightblue")) +
    theme_minimal()
```

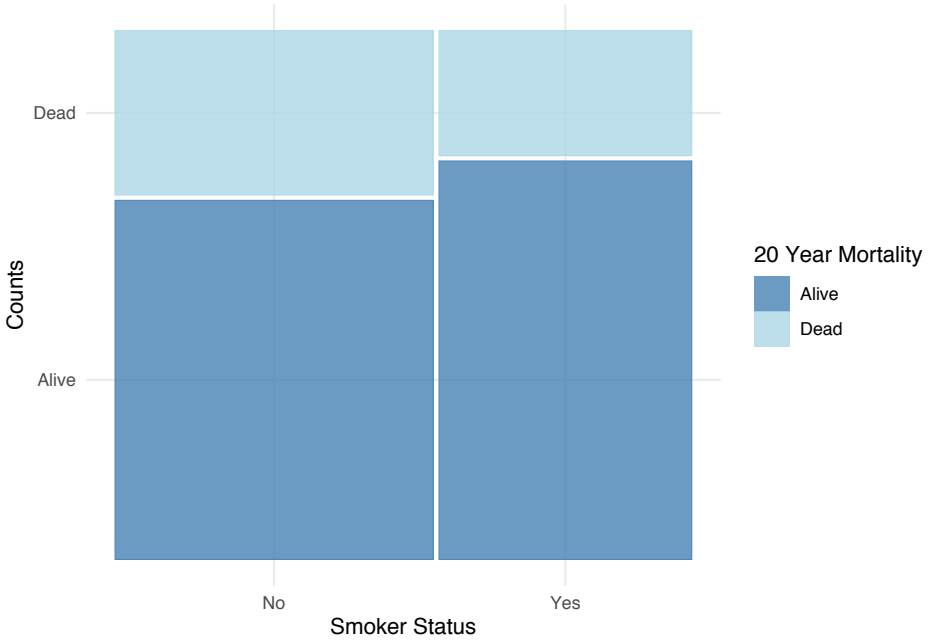


2.5.4 Mosaic Plot

The best (Prof. Heggeseth's opinion) graphic for two categorical variables is a variation on the stacked bar plot called a **mosaic plot**. The total heights of the bars are the same so we can compare the conditional distributions. For a **mosaic plot**,

- The **relative height** of the bars shows the **conditional distribution** (relative frequency within subsets).
- The **width** of the bars shows the **marginal distribution** (relative frequency of the X variable, ignoring the other variable).
- Making mosaic plots in R requires another package: `ggmosaic`

```
library(ggmosaic)
Whickham %>%
    ggplot() +
    geom_mosaic(aes(x = product(outcome, smoker), fill = outcome)) +
    xlab('Smoker Status') +
    ylab('Counts') +
    scale_fill_manual('20 Year Mortality', values = c("steelblue", "lightblue")) +
    theme_minimal()
```



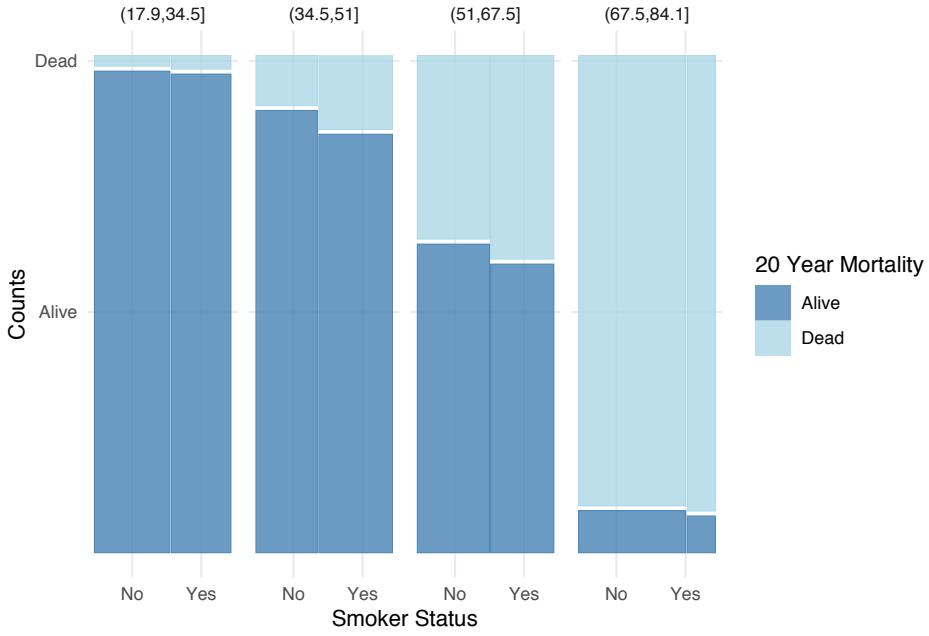
Reflect: What information is highlighted when you focus on relative frequency in the mosaic plots as compared to other bar plots?

With this type of plot, you can see that there are more non-smokers than smokers. Also, you see that there is a higher mortality rate for non-smokers.

Reflect: Does our data suggest that smoking *is associated* with a lower mortality rate? Does our data suggest that smoking *reduces* mortality? Note the difference in these two questions - the second implies a cause and effect relationship.

Let's consider a third variable here, age distribution. We can create the same plot, separately for each age group.

```
Whickham %>%
    ggplot() +
    geom_mosaic(aes(x = product(outcome, smoker), fill = outcome)) +
    facet_grid( . ~ ageCat) +
    xlab('Smoker Status') +
    ylab('Counts') +
    scale_fill_manual('20 Year Mortality', values = c("steelblue", "lightblue")) +
    theme_minimal()
```



Reflect: What do you gain by creating plots within subgroups?

Reflect: How is it that our conclusions are exactly the opposite if we consider the relationship between smoking and mortality within age subsets? What might be going on?

This is called **Simpson's Paradox,** which is a situation in which you come to two different conclusions if you look at results overall versus within subsets (e.g. age groups).

Let's look at the marginal distribution of smoking status within each age group. For groups of people that were 68 years of age or younger, it was about 50-50 in terms of smoker vs. non smoker. But, the oldest age group were primarily nonsmokers.

Now look at the mortality rates within each age category. The 20-year mortality rate among young people (35 or less) was very low, but mortality increases with increased age. So the oldest age group had the highest mortality rate, due primarily to their age, and also had the highest rate of non-smokers. So when we look at everyone together (not subsetting by age), it looks like smoking is associated with a lower mortality rate, when in fact age was just confounding the relationship between smoking status and mortality.

2.6 One Quantitative Variable

Next, we use data from one of the largest ongoing health studies in the USA, named NHANES. In particular, we will focus on data from the NHANES between 2009-2012 (Source: CDC). For more info about NHANES: https://www.cdc.gov/nchs/nhanes/index.htm.

Since sleep is vitally important to daily functioning, let's look at the number of hours of sleep respondents reported.

2.6.1 Histogram

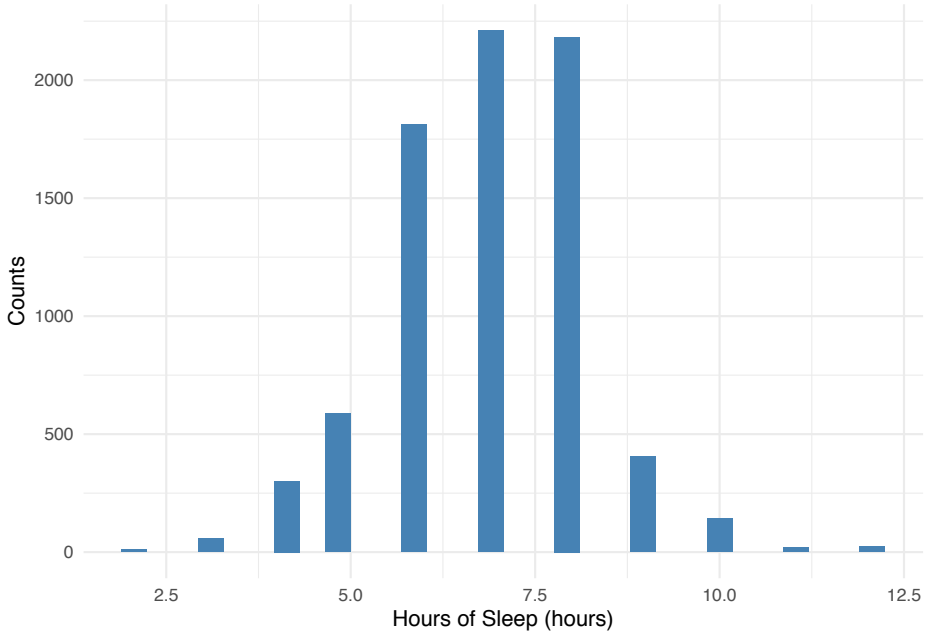
One main graphical summary we use for quantitative variables is a histogram. It resembles a bar plot, but there are a few key differences:

- The x-axis is a number line that is divided into intervals called **bins**. Bins technically do not all have to be of equal width but almost always are. When making histograms in R, R chooses a default bin width, but you have options to change the number and/or width of the bins/intervals.
- The **height** of the bars shows either the **frequency within intervals** (counts of cases that fall into that bin/interval) or the **density** (fraction of cases that fall into that bin/interval).
- Gaps between bars are meaningful. They indicate absence of values within an interval.

```
data(NHANES)
NHANES %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(fill = "steelblue") +
    xlab('Hours of Sleep (hours)') +
    ylab('Counts') +
    theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_bin).
```

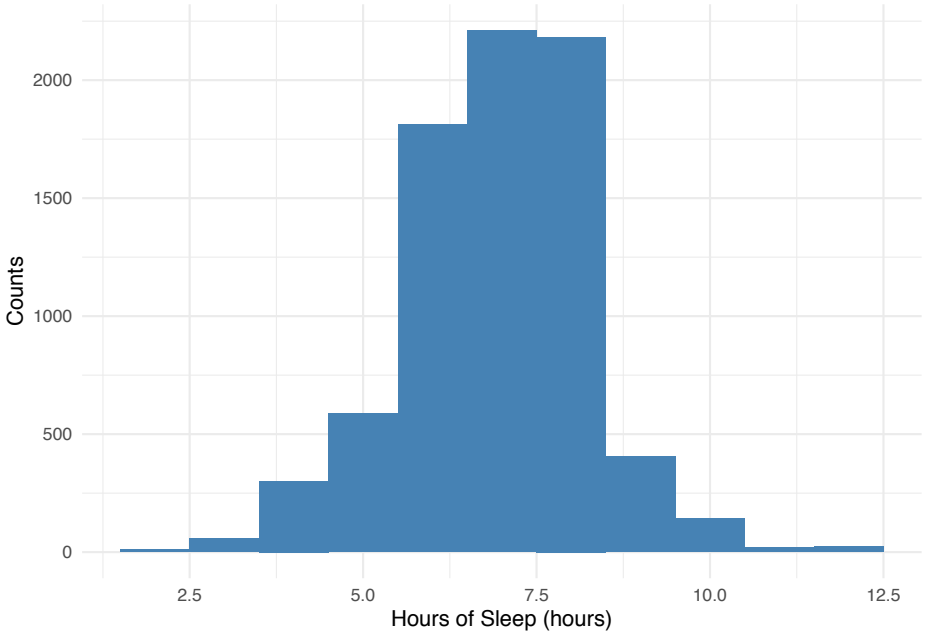


Note the warning message above: "Removed ___ rows containing non-finite values (stat_bin)." Sometimes there is missing information for a variable for some cases in the dataset. We cannot plot these because we don't know their values! This warning message is just a friendly reminder from R to let you know what it is doing.

Also note the message that R gives about bin width to remind us that we can choose this if we wish. If we want to specify the width of the intervals or bins, we can specify `binwidth = DESIRED_BIN_WIDTH` within `geom_histogram`.

```
NHANES %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(binwidth = 1, fill = "steelblue") +
    xlab('Hours of Sleep (hours)') +
    ylab('Counts') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_bin).
```

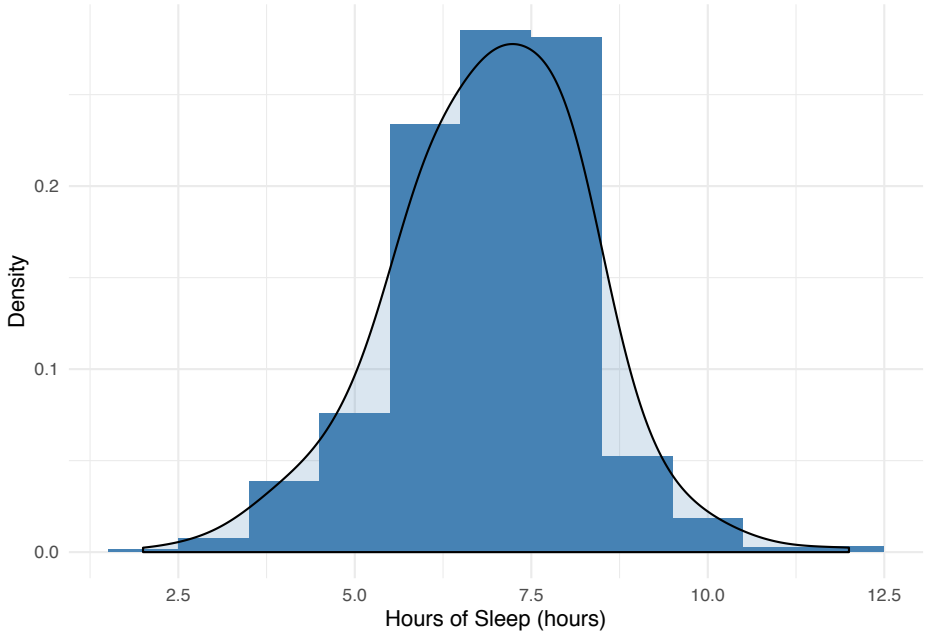


Lastly, notice that the y-axis in the previous two histograms has been the counts (or frequency) within each sleep hour interval. We can adjust this to **density**, which is relative frequency adjusted for the width of interval so that the sum of the areas of the bars (height x width) equals 1.

```
NHANES %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(aes(y = ..density..), binwidth = 1, fill = "steelblue") +
    geom_density(alpha = 0.2, fill = "steelblue", adjust = 3) +
    xlab('Hours of Sleep (hours)') +
    ylab('Density') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_density).
```



The smooth curved line on this plot is called a **density plot**. It is essentially a smoother version of the histogram. Both the area under a density plot and the total area of all the rectangles in a density histogram equal 1.

When describing a distribution, we focus on three aspects of the histogram:

- **Shape:** Is it **symmetric** (can you fold it in half and the sides match up)? or is it **skewed to the right or left**? (A distribution is **left-skewed** if there is a long left tail and **right-skewed** if it has a long right tail.) How many **modes** ("peaks"/"bumps" in the distribution) do you see?
- **Center:** Where is a typical value located?
- **Spread** (or variation): How spread out are the values? Concentrated around one or more values or spread out?
- **Unusual features:** Are there **outliers** (points far from the rest)? Are there gaps? Why?

Here is another data set for comparison. Here are the annual salaries for the highest paid CEOs in 2016 (Source: NYTimes). To get the data, we are scraping the data from a NYTimes website. For fun, you can look at the code below.

```
nyturl <- 'https://www.nytimes.com/interactive/2017/05/26/business/highest-paid-ceos.ht
dat <- read_html(nyturl)
ceo <- dat %>%
    html_nodes(".nytg-compensation , .nytg-year") %>%
    html_text() %>%
    str_replace('\\$|-','') #webscraping data
ceo <- data.frame(matrix(ceo,ncol = 2,byrow = TRUE))
```

```
names(ceo) <- c('year','salary')
ceo$salary <- as.numeric(ceo$salary)
ceo <- ceo %>%
    filter(year == '2016')
```

Let's create a density histogram of the annual salaries for the highest paid CEO's in the U.S. in 2016.

```
ceo %>%
    ggplot(aes(x = salary)) +
    geom_histogram(aes(y = ..density..), binwidth = 15, fill = "steelblue") +
    geom_density(alpha = 0.2, fill = "steelblue") +
    xlab('Salary ($ Millions)') +
    ylab('Counts') +
    theme_minimal()
```

We note that some of the highest salaries were close to 200 million U.S. dollars (in 2016), but the majority of the salaries in this sample are closer to 50 million U.D. dollars.

Reflect: Is this distribution of salaries left-skewed or right-skewed? In what populations do you think salaries might be left-skewed? Right-skewed?

2.6.2 Center

There are some choices for numerically summarizing the center of a distribution:

- **Mean**: The sum of the values divided by the number of values (sample size), $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
  - Sensitive to outliers, but it efficiently uses all the data
- **Median**: The "middle" value. The number for which half of the values are below and half are above.
  - Insensitive to outliers, but it doesn't use all the actual values
- **Trimmed means**: Drop the lowest and highest k% and take the mean of the rest.
  - A good compromise, but not widely used.

> **Math Box:** The Greek capital letter sigma, $\sum$, is used in mathematics to denote a sum. We let y_i represent the value of the ith person for a variable called y. So $\sum_{i=1}^{n} y_i$ is the sum of all the n values of a variable y, all the way from the 1st person to the nth person.

We can calculate all of these in R.

- **Hours of sleep per night from the NHANES dataset**

```
NHANES %>%
  select(SleepHrsNight) %>%
  summary()
```

```
##  SleepHrsNight
##  Min.   : 2
##  1st Qu.: 6
##  Median : 7
##  Mean   : 7
##  3rd Qu.: 8
##  Max.   :12
##  NA's   :2245
```

```
NHANES %>%
  summarize(mean(SleepHrsNight, na.rm = TRUE), median(SleepHrsNight, na.rm = TRUE), mea
```

```
## # A tibble: 1 x 3
##   `mean(SleepHrsNight, n~ `median(SleepHrsNight~ `mean(SleepHrsNight, trim~
##                     <dbl>                  <int>                      <dbl>
## 1                    6.93                      7                       6.95
```

- **CEO salary information from NYT**

```
ceo %>%
  select(salary) %>%
  summary()  # Note the differences between mean and median
```

```
##      salary
##  Min.   : 20.0
##  1st Qu.: 36.8
##  Median : 58.0
##  Mean   : 68.4
##  3rd Qu.: 96.0
##  Max.   :176.0
```

```
ceo %>%
  summarize(mean(salary), median(salary), mean(salary, trim = 0.05))
```

```
##   mean(salary) median(salary) mean(salary, trim = 0.05)
## 1         68.4             58                      66.2
```

Note that the mean, median, and trimmed mean are all fairly close for the sleep hours distribution, which looks fairly symmetric.

Note also that the mean, median, and trimmed mean are somewhat different for the salary distribution, which looks right skewed. Often with right skewed distributions, the mean tends to be higher than the median because particularly large values are being summed in the calculation. The median and trimmed mean are not as sensitive to these outliers because of the sorting that is involved in their calculation.

2.6.3 Boxplot

An alternative graphical summary is a boxplot, which is a simplification of the histogram. The plot consists of

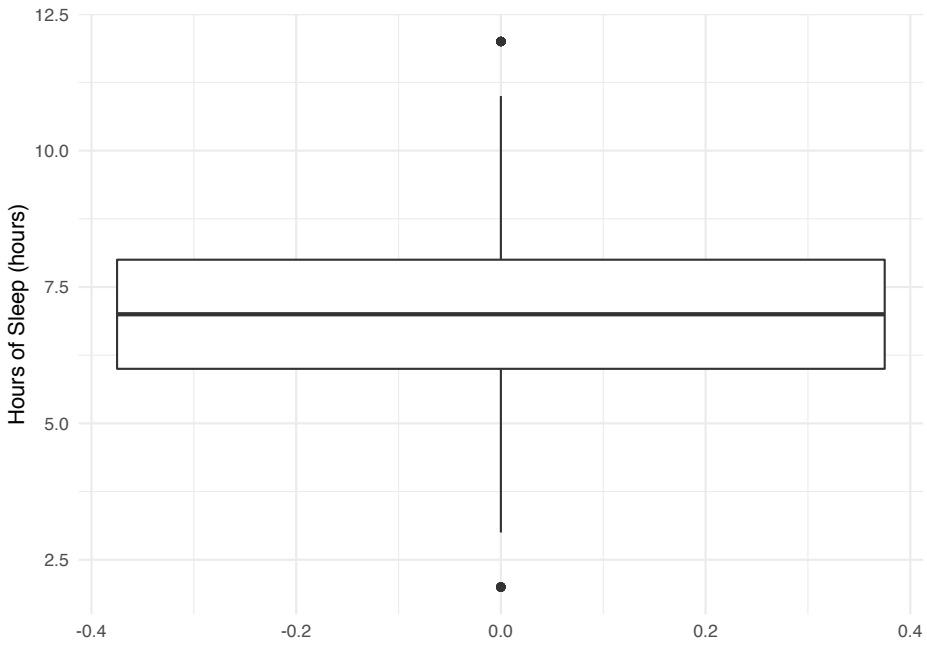
- A Box: the bottom of the box is at the 25th percentile ($Q1$) and top of the box is at the 75th percentile ($Q3$)
- Line in Box: the line in the middle of the box is at the 50th percentile, the median
- Tails/Whiskers: The lines extend out from the box to most extreme observed values within $1.5 \times (Q3 - Q1)$ from $Q1$ (bottom) or $Q3$ (top)
- Points: If any points are beyond $1.5 \times (Q3 - Q1)$ from the box edges, they are considered outliers and are plotted separately

Math Box: A percentile is a measure indicating the value below which a given percentage of observations in a group of observations fall. So the 25th percentile is the value at which 25% of the values are below. The 95th percentile is the point at which 95% of the observations are below.

Here is a boxplot of the sleep amount from NHANES.

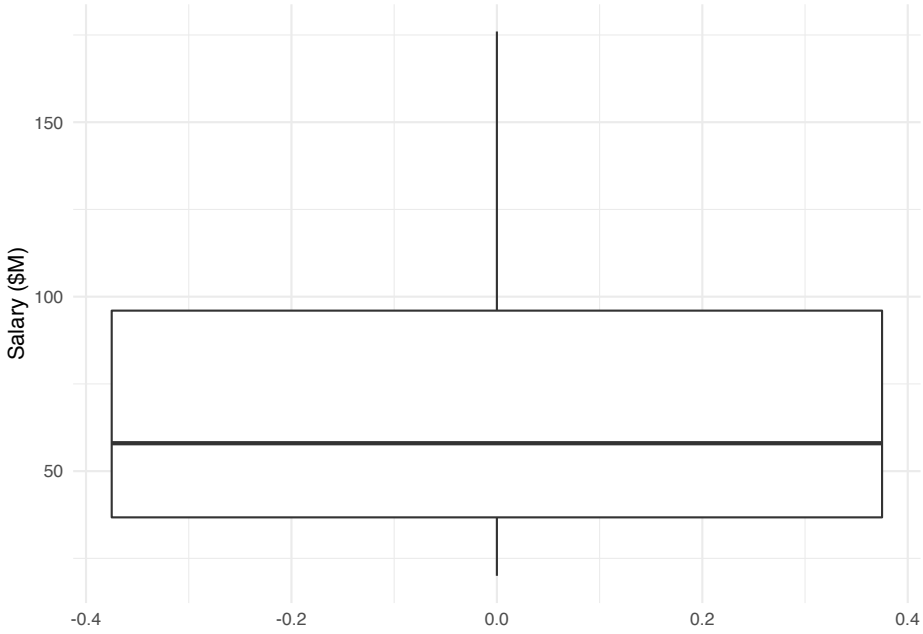
```
NHANES %>%
    ggplot(aes(y = SleepHrsNight)) +
    geom_boxplot() +
    ylab('Hours of Sleep (hours)') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_boxplot).
```



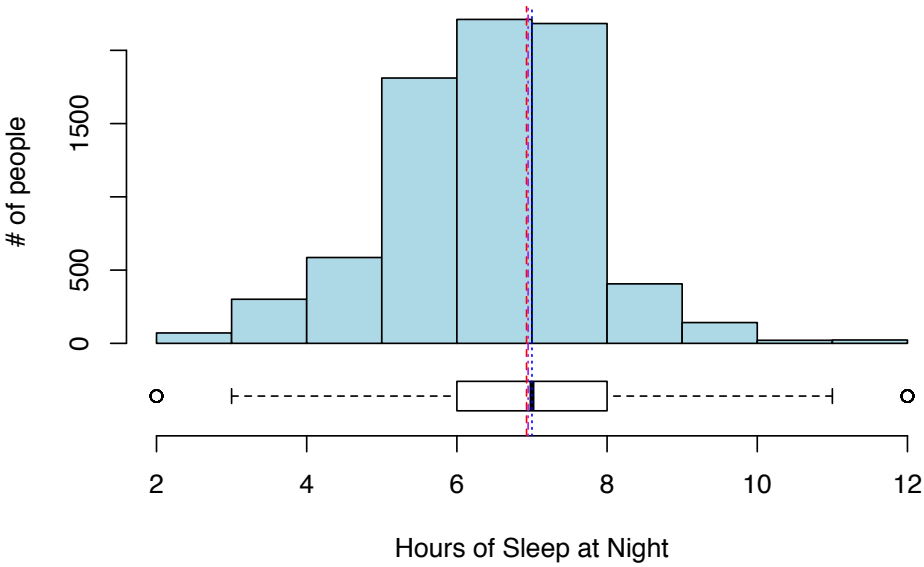
Compare that to the boxplot of the CEO salaries.

```
ceo %>%
    ggplot(aes(y = salary)) +
    geom_boxplot() +
    ylab('Salary ($M)') +
    theme_minimal()
```



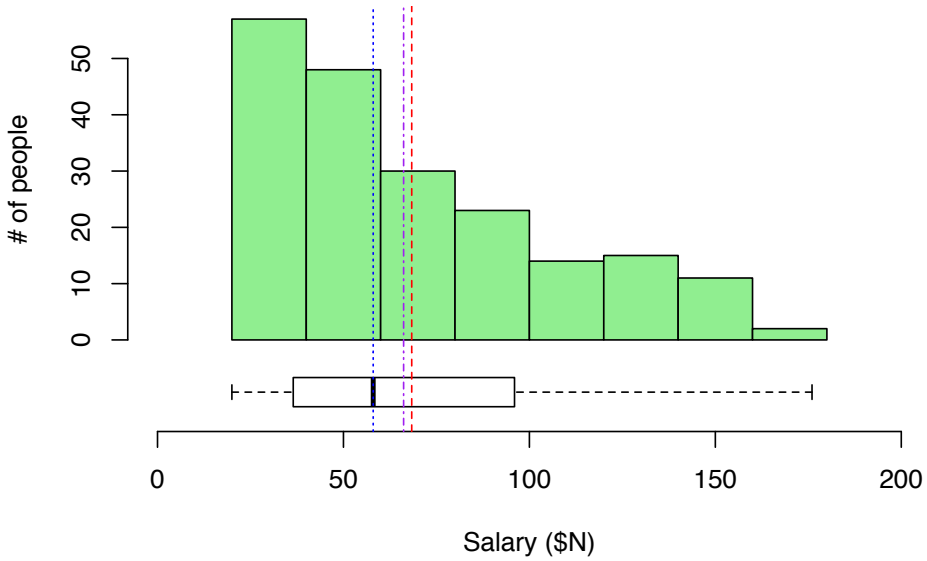
Note: In these 2 plots above, the x-axis has number labels, but they don't mean anything.

Let's put the boxplots next to the histograms so we can better compare the two types of visualizations. Also, let's add the mean (red dashed), median (blue dotted), and 5% trimmed mean (purple dash-dot) as annotations.



For the hours of sleep, the mean, median, and 5% trimmed mean are all pretty

much the same. Note also that the distribution looks pretty symmetric based on the histogram.



For CEO salaries, the mean and 5% trimmed mean are a bit higher than the median. **The mean is always pulled toward the long tail.**

> **Reflect:** What would the boxplot look like if all of the values were exactly the same? Sometimes when making multiple boxplots for each of multiple groups, a group may only have one value or a small number of values that all happen to be identical. What will this look like?

2.6.4 Spread

There are some choices for numerically summarizing the spread of a distribution:

- **Range**: the maximum value - the minimum value
  - Sensitive to the outliers since it's the difference of the extremes
  - Units (e.g. inches, pounds) are the same as the actual data
- **IQR**: the interquartile range : $Q3 - Q1$ (75th percentile - 25th percentile).
  - Length of the box in a boxplot
  - Spread of middle 50% of data
  - Like the median. Less sensitive because it doesn't use all of the data
  - Units are the same as the actual data
- **Standard deviation** (SD): Root mean squared deviations from mean, $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$

  - Roughly the average size of deviation from the mean ($n-1$ instead of n)
  - Uses all the data but very sensitive to outliers and skewed data (large values are first squared).
  - Units are the same as the actual data
- **Variance**: Square of the standard deviation
  - Units are the squared version of the actual data's units (e.g. squared inches, pounds)
  - Standard deviation is preferred for interpretability of units
  - Variance will come up when we discuss models in the next chapter

We can calculate all of these in R.

- **Hours of sleep per night from the NHANES dataset**

```
NHANES %>% summarize(diff(range(SleepHrsNight, na.rm = TRUE)), IQR(SleepHrsNight,
    na.rm = TRUE), sd(SleepHrsNight, na.rm = TRUE), var(SleepHrsNight, na.rm = TRUE))
```

```
## # A tibble: 1 x 4
##   `diff(range(Sleep~ `IQR(SleepHrsNigh~ `sd(SleepHrsNigh~ `var(SleepHrsNig~
##                <int>              <dbl>             <dbl>             <dbl>
## 1                 10                  2              1.35              1.81
```

```
# range gives max and min; take difference betwee max and min IQR = Q3-Q1 sd
# = standard deviation var = variance
```

- **CEO salary information from NYT**

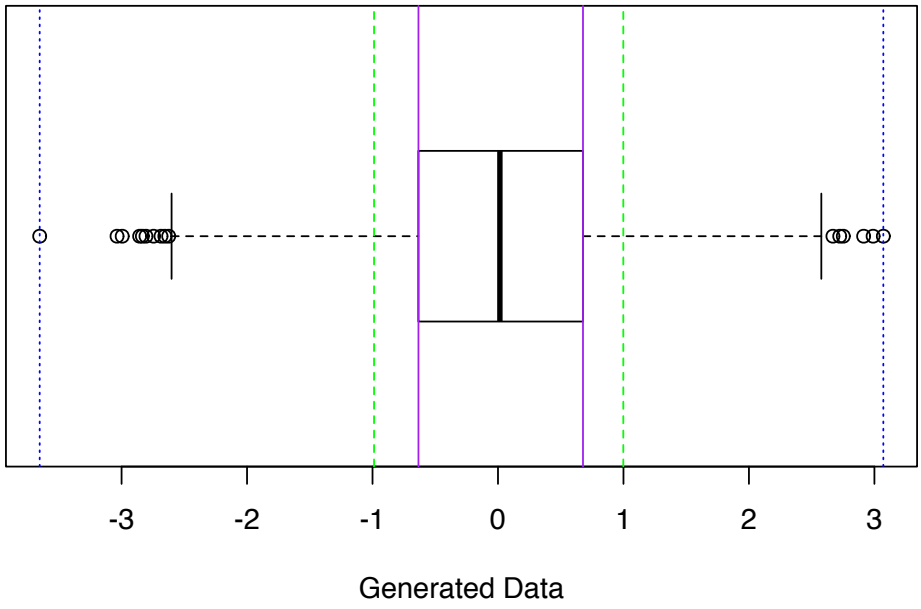
```
ceo %>% summarize(diff(range(salary)), IQR(salary), sd(salary), var(salary))
```

```
##   diff(range(salary)) IQR(salary) sd(salary) var(salary)
## 1                 156        59.2       39.1        1526
```

2.6.5 Some data accounting

We've looked at different measures of the spread of a distribution. Do some measures of spread encompass a lot of the data? Just a little? Can we be more precise about how much of the data is encompassed by intervals created from different spread measures?

```
x <- rnorm(1500)
boxplot(x, horizontal = TRUE, xlab = "Generated Data")
abline(v = range(x), col = "blue", lty = 3)
abline(v = quantile(x, c(0.25, 0.75)), col = "purple", lty = 1)
abline(v = c(mean(x) - sd(x), mean(x) + sd(x)), col = "green", lty = 2)
```



Reflect: What percentage of the data is between the blue dotted lines (length of interval is range)?
What percentage of the data is between the purple solid lines (length of interval is IQR)?
What percentage of the data is between the green dashed lines (length of interval is 2*SD)?

The code below computes the fraction of data points, `x`, that fall between the lower bound of 1 SD below the mean and the upper bound of 1 SD above the mean.

```
sum(x > mean(x) - sd(x) & x < mean(x) + sd(x))/length(x)
```

```
## [1] 0.678
```

So with this data set, about 68% of the data values fall within 1 SD of the mean.

Reflect: If we had a different data set, do you know that answer to the following questions? *You should know the answer to 2 of them at this point...*

- What percentage of the data would be between the minimum and maximum (blue dotted lines above)?
- What percentage of the data would be between bottom and top of the box (purple solid lines above)?
- What percentage of the data would be between 1 SD below the mean and 1 SD above the mean (green dashed lines)?

2.6.6 Z-scores

How do you decide when an outlier is really unusual (think: athletic victory being very impressive or a data point that may be a typing error such as a human weight of 3000 lbs)?

If the observation is far from the rest of the measurements in the data, we tend to say that the value is unusual. We want to quantify this idea of "unusual".

To do this, we often calculate a **z-score**, a standardized data value which we denote with the letter, z. To calculate a z-score,

- Calculate how far an observation, y, is below (or above) the mean of the sample, denoted as $\bar{y}$.
- Then divide the difference by the standard deviation (measure of spread), denoted as s_y.

$$z = \frac{y - \bar{y}}{s_y}$$

The z-score tells you how many standard deviations the observation is above or below the mean.

Reflect: Say that you got a z-score of 1 on an exam with mean = 80 and SD = 5. That means that you got an 85 on the exam because your exam is one SD above the mean ($mean + z \times SD = 80 + 1 \times 5$).
If you got a z = -2 on an exam with mean = 80 and SD = 5, that means you got a 70 on the exam because your exam is two SD below the mean ($mean + z \times SD = 80 + -2 \times 5$).

In general, it is quite common to have z-scores between -3 and 3, but fairly unusual to have them greater than 3 or less than -3.

Often, if you have data with a **unimodal, symmetric distribution**,

- about 68% of the z-scores are between -1 and 1,
- about 95% of the z-scores are between -2 and 2,
- about 99.7% of the z-scores are between -3 and 3.

This is not true for every histogram, but it will be true for a particularly special distribution that we will see later when we cover models. (This distribution is called the **normal distribution** or **Gaussian distribution**.)

However, we do know that z-scores of 5 or larger in magnitude (ignoring negative sign) are very unusual, no matter the shape of the histogram/distribution. For those inclined, see the mathematical theorem below that tells us this.

Math Box: (Optional) Chebyshev's inequality gives bounds for the percentages no matter the shape of the distribution. It states that for any real number $k > 0$, the chance of getting a z-score greater in magnitude (ignoring the negative sign) than k is less than or equal to $1/k^2$,

$$P(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = \frac{|X-\mu|}{\sigma}$ is a z-score, μ is the mean, and σ is the standard deviation.
If we plug in values for k, we see that the chance of getting a z-score

- at least 3 in magnitude (> 3 or < -3) is less than $(1/3^2) = 0.11 = 11\%$.
- at least 4 in magnitude (> 4 or < -4) is less than $(1/4^2) = 0.06 = 6\%$.
- at least 5 in magnitude (> 5 or < -5) is less than $(1/5^2) = 0.04 = 4\%$.

This is true for any shaped distribution (skewed, bimodal, etc.). See proof here based on probability theory.

In summary, for a quantitative variable,

- Use a histogram to display the distribution of one variable and describe the shape and any unusual features.
- For "well-behaved" distributions (symmetric, unimodal, no outliers), use the mean and standard deviation to describe the center and spread. Then z-scores will roughly follow the 68-95-99.7 rule stated above.
- For other distributions (skewed or bimodal), use the IQR and median. You can report both mean and median, but it's usually a good idea to state why.

2.7 One Quant. and One Cat. Variable

Let's return to the NHANES data. We noticed variation in the amount people sleep. Why do some people sleep more than others? Are there any other characteristics that may be able to *explain that variation*?

Let's look at the distribution of hours of sleep at night within subsets or groups of the NHANES data.

2.7.1 Multiple Histograms

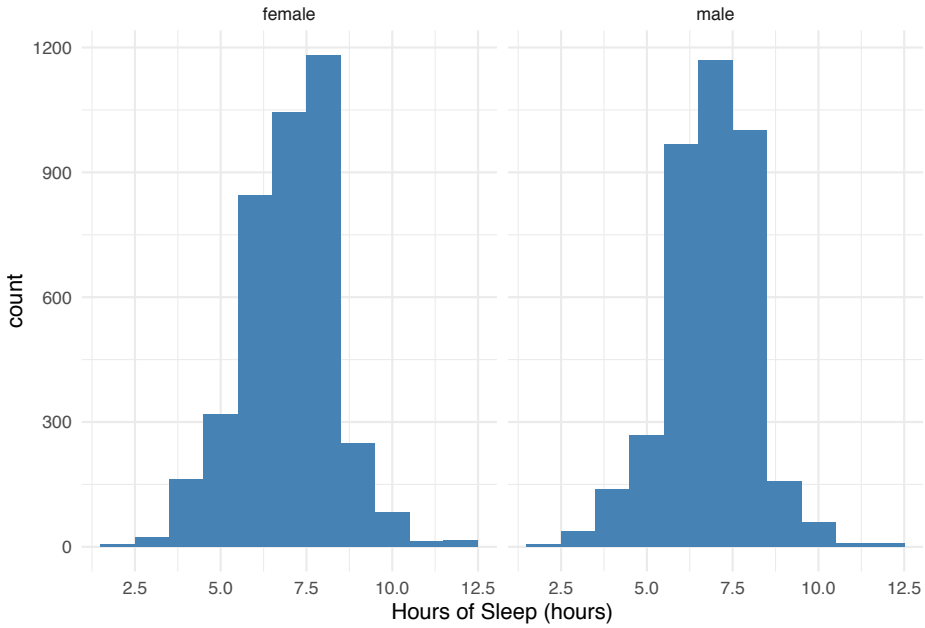
Does the recorded binary gender explain the variability in the hours of sleep?

> **Reflect:** What are the *ethical implications* of collecting gender identity as a binary variable (male/female) if some individuals do not identify with these categories?
> What might be the *causal mechanism* between gender identity and sleep? Might you be more interested in hormone levels, which might not necessarily correspond to gender identity? How might you change the data collection procedure so that the data can address the underlying research question?

Let's make a histogram for each gender category by adding `facet_grid(. ~ Gender)` which separates the data into groups defined by the variable, `Gender`, and creates two plots along the x-axis.

```
NHANES %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(binwidth = 1, fill = "steelblue") +
    xlab('Hours of Sleep (hours)') +
    facet_grid(. ~ Gender) +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_bin).
```



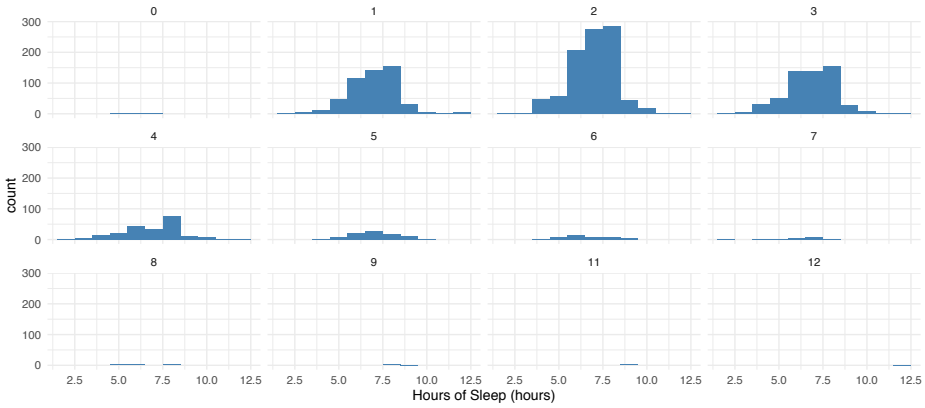
> **Reflect:** Do you notice any differences in sleep hour distributions between males and females?
> What is easy to compare and what is hard to compare between the two histograms?

Does the number of children a woman has explain the variability in the hours of sleep?

> **Reflect:** Who have we excluded from our analysis by asking this question?

```
NHANES %>%
    filter(!is.na(nBabies)) %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(binwidth = 1, fill = "steelblue") +
    xlab('Hours of Sleep (hours)') +
    facet_wrap(. ~ factor(nBabies), ncol = 4) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    theme_minimal()
```

```
## Warning: Removed 8 rows containing non-finite values (stat_bin).
```



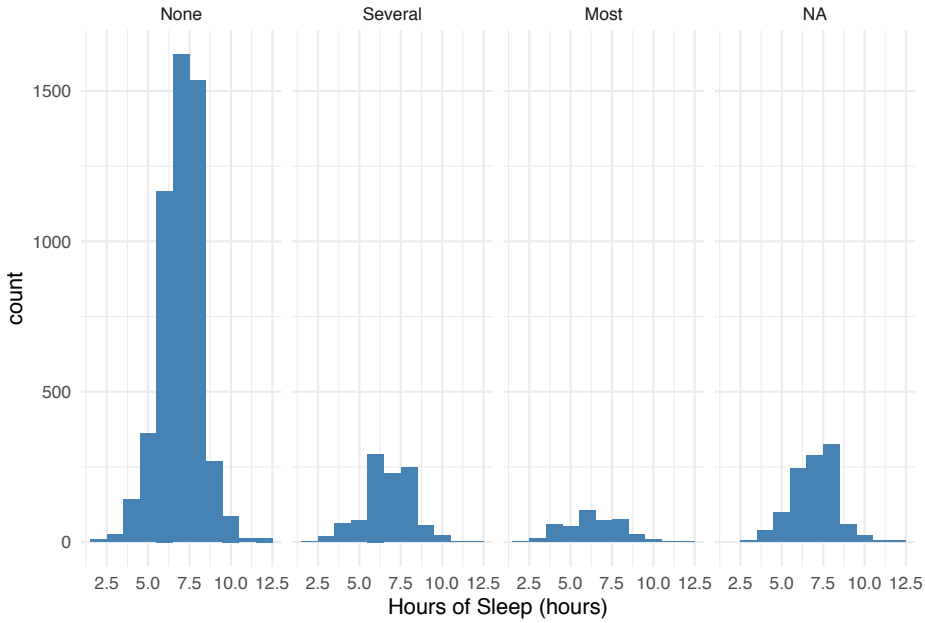
The 0 to 12 labels at the top of each of these panels correspond to the number of babies a woman had.

Reflect: Do you notice any differences in sleep hour distributions between these groups?
Note the x and y axes are the same for all of the groups to facilitate comparison. What is easy to compare and what is hard to compare between the histograms?

Does the number of days someone has felt depressed explain the variability in the hours of sleep?

```
NHANES %>%
    ggplot(aes(x = SleepHrsNight)) +
    geom_histogram(binwidth = 1, fill = "steelblue") +
    xlab('Hours of Sleep (hours)') +
    facet_grid(. ~ Depressed) +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_bin).
```



What's the rightmost "NA" category? Some individuals in this study did not answer questions about days that they might have felt depressed, but they did report their hours of sleep per night.

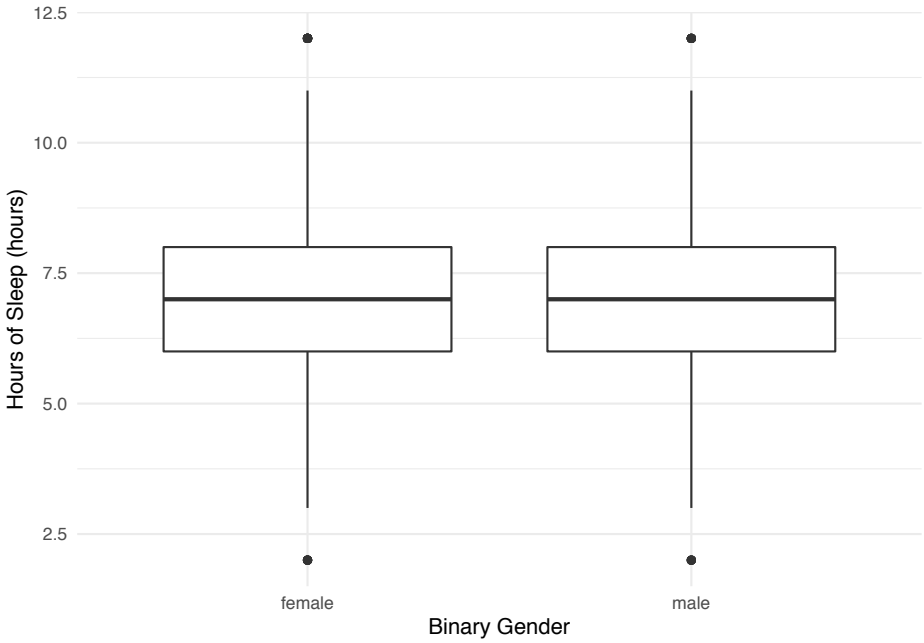
Reflect: What type of biases might be at play here?

2.7.2 Multiple Boxplots

Let's visualize the same information but with boxplots instead of histograms and see if we can glean any other information.

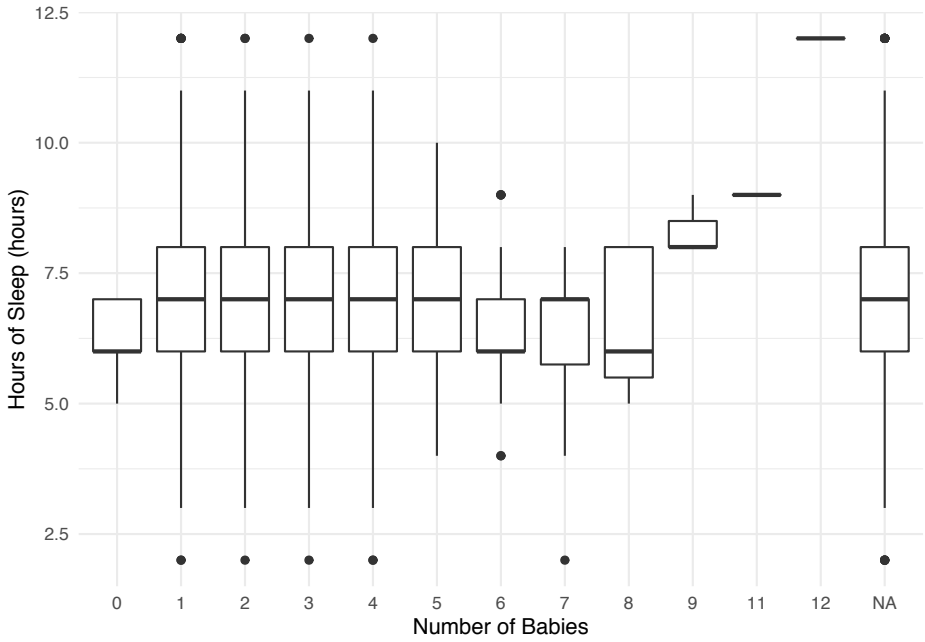
```
NHANES %>%
    ggplot(aes(x = Gender, y = SleepHrsNight)) +
    geom_boxplot() +
    ylab('Hours of Sleep (hours)') +
    xlab('Binary Gender') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_boxplot).
```



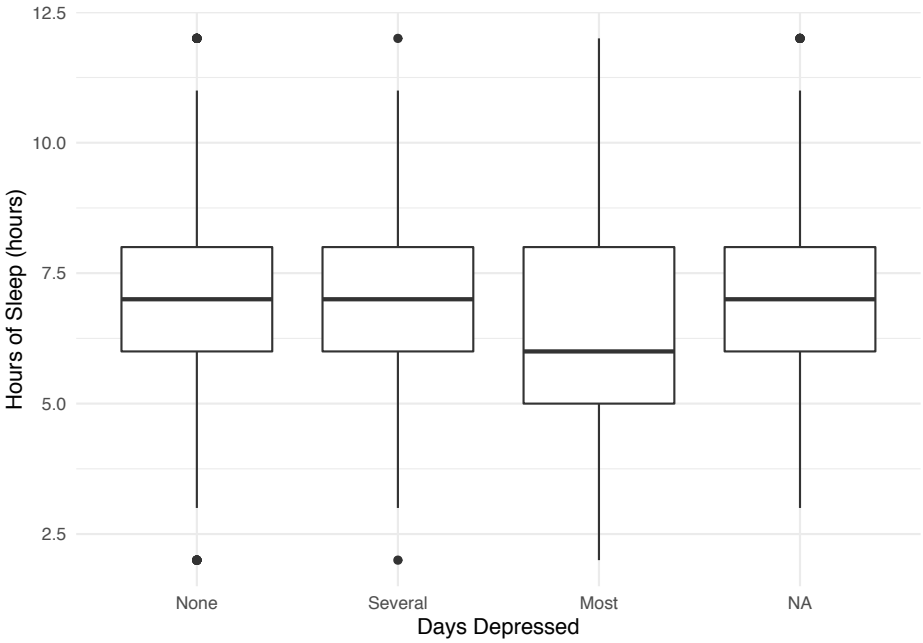

```
NHANES %>%
    ggplot(aes(x = factor(nBabies), y = SleepHrsNight)) +
    geom_boxplot() +
    ylab('Hours of Sleep (hours)') +
    xlab('Number of Babies') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_boxplot).
```




```
NHANES %>%
    ggplot(aes(x = factor(Depressed), y = SleepHrsNight)) +
    geom_boxplot() +
    ylab('Hours of Sleep (hours)') +
    xlab('Days Depressed') +
    theme_minimal()
```

```
## Warning: Removed 2245 rows containing non-finite values (stat_boxplot).
```



Reflect: What is easy to compare and what is hard to compare between the boxplots?
Why might you use multiple boxplots instead of multiple histograms?

2.7.3 Is this a Real Difference?

If we notice differences in the the sleep distributions for groups based on self-reported Depression, is it a "REAL" difference? That is, is there a difference in the general U.S. population? Remember, we only have a random *sample* of the population. *NHANES is supposed to be a representative sample of the U.S. population collected using a probability sampling procedure.*

What if there were no "REAL" difference? Then the Depressed group labels wouldn't be related to the hours of sleep.

Investigation Plan:

1. Take all of the observed data on the hours of sleep and randomly shuffle them into new groups (of the same sizes as before). This breaks any associations between the Depressed group labels and the reported hours of sleep.
2. Calculate the difference in mean hours of sleep between the groups. Record it.
3. Repeat steps 1 and 2 many times (say 1000 times).

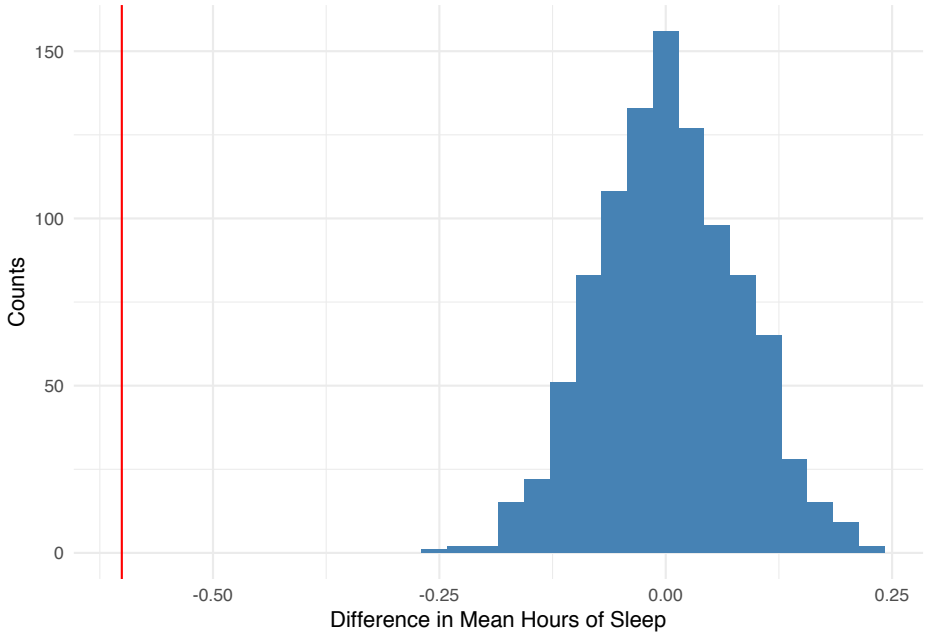
4. Look at the differences based on random shuffles & compare to the observed difference.

```
library(mosaic)
NHANES <- NHANES %>%
  mutate(DepressedMost = (Depressed == 'Most')) #TRUE or FALSE (converted Depressed to

obsdiff <- data.frame(d = diff(mean(SleepHrsNight ~ DepressedMost, data = NHANES, na.r

sim <- do(1000)*diff(mean(SleepHrsNight ~ shuffle(DepressedMost), data = NHANES, na.rm
```

Below, we have a histogram of 1000 values calculated by randomly shuffling individuals in the sample into two groups (assuming no relationship between depression and sleep) and then finding the difference in the mean amount of sleep. The red vertical line showed the observed difference in mean amount of sleep in the data.

```
sim %>%
  ggplot(aes(x = TRUE.)) +
  geom_histogram(fill = 'steelblue') +
  geom_vline(aes(xintercept = d), obsdiff, color = 'red') +
  xlab('Difference in Mean Hours of Sleep') +
  ylab('Counts') +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Reflect: The observed difference in mean hours of sleep (red line) is quite far from the distribution of differences that results when we break the association between depression status and sleep hours (through randomized shuffling of group labels). Thus, it is unlikely to get a difference that large if there were no relationhip.
What do you think this indicates? How might you use this as evidence for or against a "real" population difference?

2.8 Two Quantitative Variables

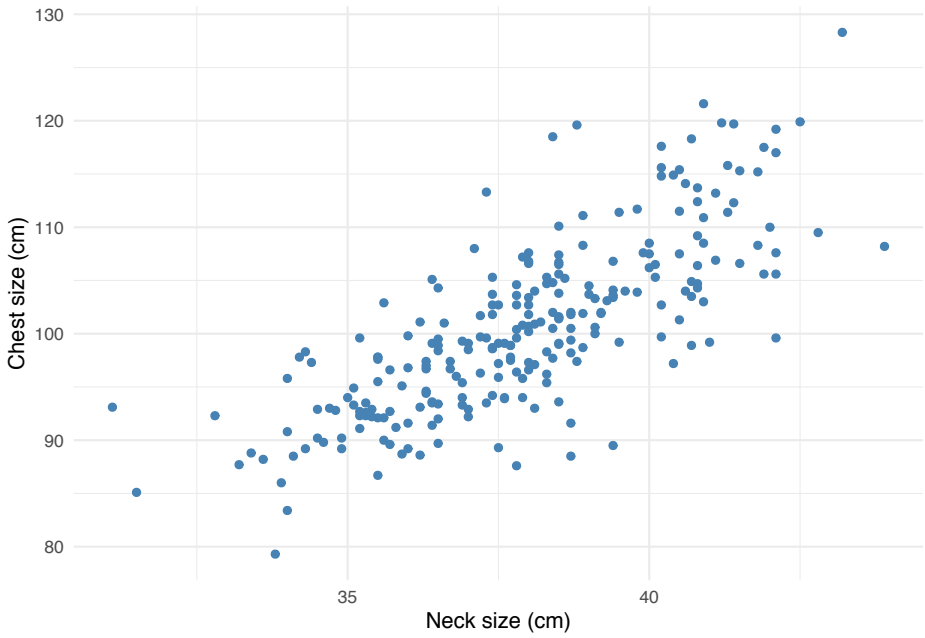
To discuss two quantitative variables, let's switch to new data set and consider a thought experiment.

Imagine that you are an entrepreneur selling button-down dress shirts. Clothing sizing are quite variable across clothing brands, so we are going to use our own data to come up with appropriate sizes for our customers. Two of the key measurements that we will use are the neck size in centimeters and chest size in centimeters of a customer. There are other variables in the data set, but let's focus on these two for the moment.

2.8.1 Scatterplot

When you have two quantitative variables, a **scatterplot** is the main appropriate graphical display of the relationship. Each point represents the neck and chest size of one customer.

```
body <- read.delim("Data/bodyfat.txt")

body %>%
    ggplot(aes(x = Neck, y = Chest)) +
    geom_point(color = 'steelblue') +
    xlab('Neck size (cm)') +
    ylab('Chest size (cm)') +
    theme_minimal()
```



What do you notice about:

1. **Direction** of relationship (positive, negative, or neutral)
2. **Form** of relationship (linear, curved, none, or other)
3. **Strength** of relationship (compactness around the average relationship)
4. **Unusual** features (outliers, differences in variability in y variable across different values of x variable)

> **Reflect:** How might you use this information to determine shirt sizes for your new business venture? Come up with a few ways you could define sizes such as small, medium, large, extra large, etc.

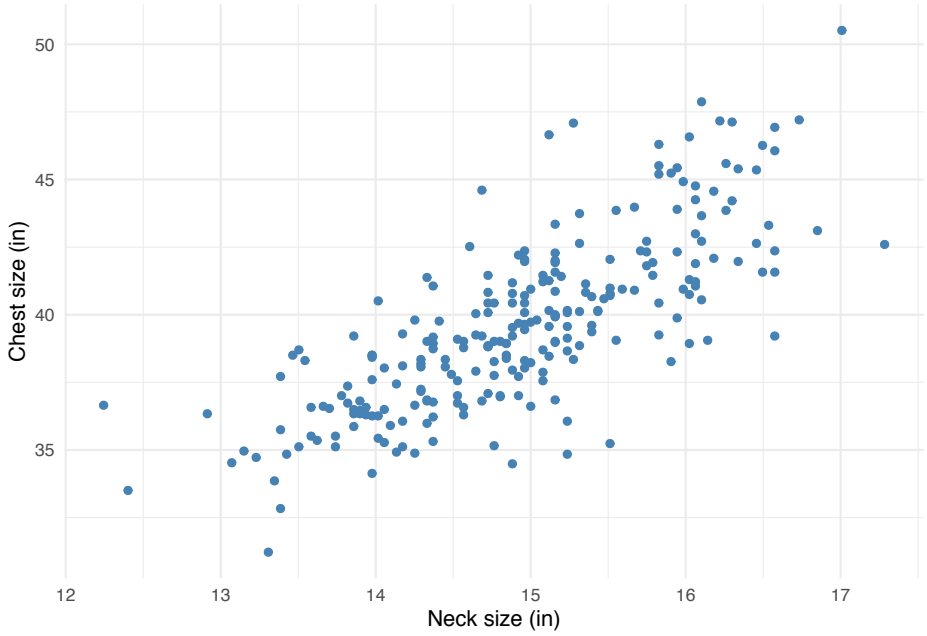
Suppose instead of *Chest* in inches and *Neck size* in cm, we plotted *Chest* in inches and *Neck size* in inches.

Does the strength of the relationship change after transformation?

Look at the plot in inches below. Does this plot look the same as the centimeters plot?

```
body %>%
  ggplot(aes(x = Neck/2.54, y = Chest/2.54)) +
  geom_point(color = 'steelblue') +
  xlab('Neck size (in)') +
  ylab('Chest size (in)') +
```

```
  theme_minimal()
```

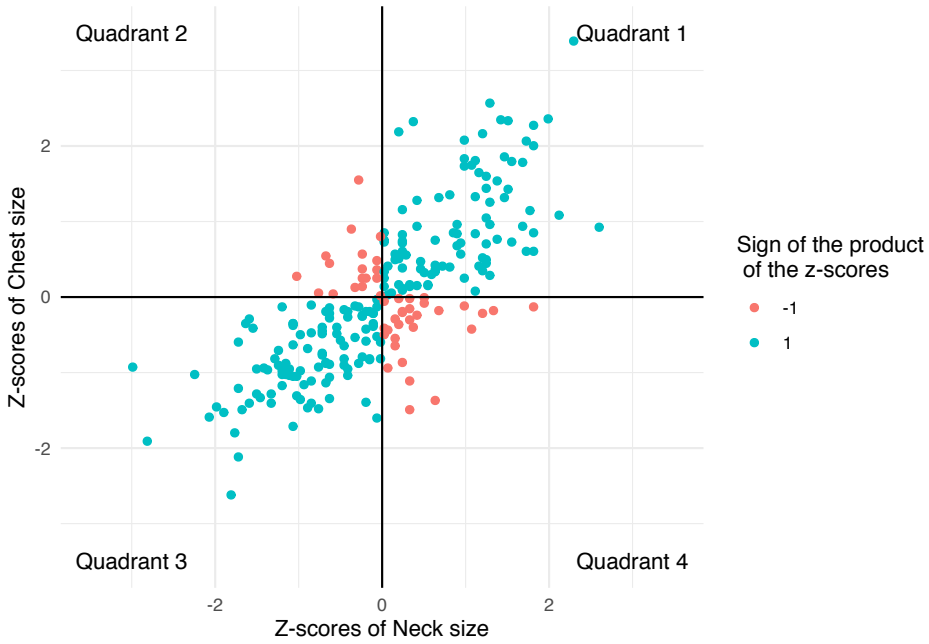


You should see that the x-axes changed but the overall shape of the plot stayed the same. Thus, the strength of the relationship was not affected by tranforming neck size from centimeters to inches (by dividing by 2.54).

2.8.2 Correlation Coefficient

Since **shifting** (adding or subtracting) and **scaling** (multiplying or dividing) make no difference in the strength of the relationship, let's standardize both variables into z-scores (recall z-scores from Section 2.6.5).

Below we plot Neck and Chest sizes after changing them to z-scores with the function `scale()` and we add some color:



The blue points in the upper right (Quadrant 1) and lower left (Quadrant 3) quadrants are either both positive or both negative in their z-score values. This means that those individuals are above average in both Neck Size and Chest Size (upper right), or they are below average in both Neck Size and Chest Size (lower left). If we multiply the z-scores of the Neck and Chest values for the blue points, we will get a positive value.

The red points in the upper left (Quadrant 2) and lower right (Quadrant 4) quadrants are positive in one and negative in the other. This means that those individuals are either above average in Neck Size but below average in Chest Size (lower right) or they are below average in Neck Size and above average in Chest Size (upper left). If we multiply the z-scores of the Neck and Chest values for the red points, we will get a negative value.

> **Reflect:** If we were to have a weaker positive relationship, how would this plot change?
> If we were to have a stronger positive relationship, how would this plot change?
> If we were to have a negative relationship, how would this plot change?

We want one number to represent **strength** and **direction** of a linear relationship.

- Points in Quadrants 1 and 3 (blue) have the z-scores of the **same sign**.

- Points in Quadrants 2 and 4 (red) have z-scores of the **opposite sign**.

What if we took the product of the z-scores for x and y variables?

Situation 1: An individual far above the means in both the x and y variables or far below the means in both the x and y variables has a very large, positive product of z-scores.

Situation 2: An individual far above the mean in x and far below the mean in y has a very large, negative product of z-scores. (The same goes for low x and high y.)

The (almost) *average* of products of the z-scores is the **correlation coefficient**,

$$r_{x,y} = \frac{\sum z_x z_y}{n-1}$$

We notate the correlation coefficient between variables x and y as $r_{x,y}$.

Some observations:

- If most of our data points follow Situation 1, the correlation coefficient is an average of mostly large positive values. Thus the correlation coefficient will be large and positive.
- If most of our data points follow Situation 2, the correlation coefficient is an average of mostly large negative values. Thus the correlation coefficient will be large and negative.
- If about an equal number of data points follow Situation 1 and Situation 2, we will be balancing positive and negative numbers, which results in a value close to zero. Thus, the correlation coefficient will be close to zero.

Which points contribute the most to this average?

Let's look at the correlation for the entire sample first. Then let's calculate the correlation for individuals around the mean Neck size.

```
body %>%
  summarize(cor(Neck, Chest)) # All data points used in calculation
```

```
##   cor(Neck, Chest)
## 1            0.769
```

```
body %>%
    filter(Neck > 35 & Neck < 40) %>% # Keep individuals with Neck size between 35cm and 40cm
      summarize(cor(Neck, Chest)) # Only middle subset of data points used in calculation
```

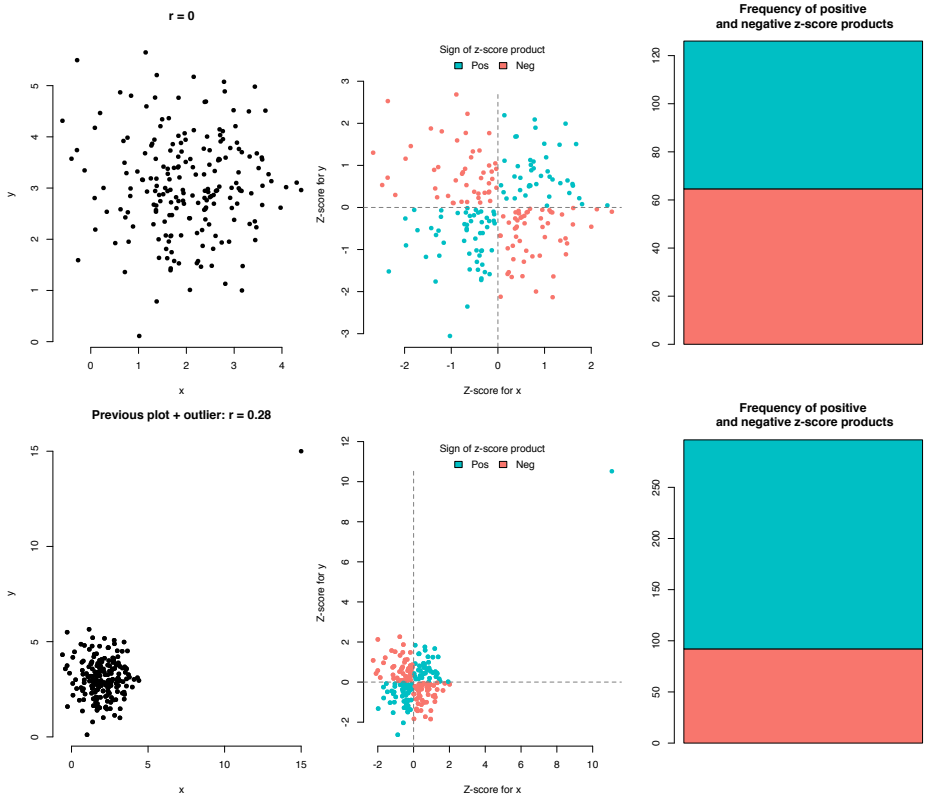
```
##   cor(Neck, Chest)
## 1            0.566
```

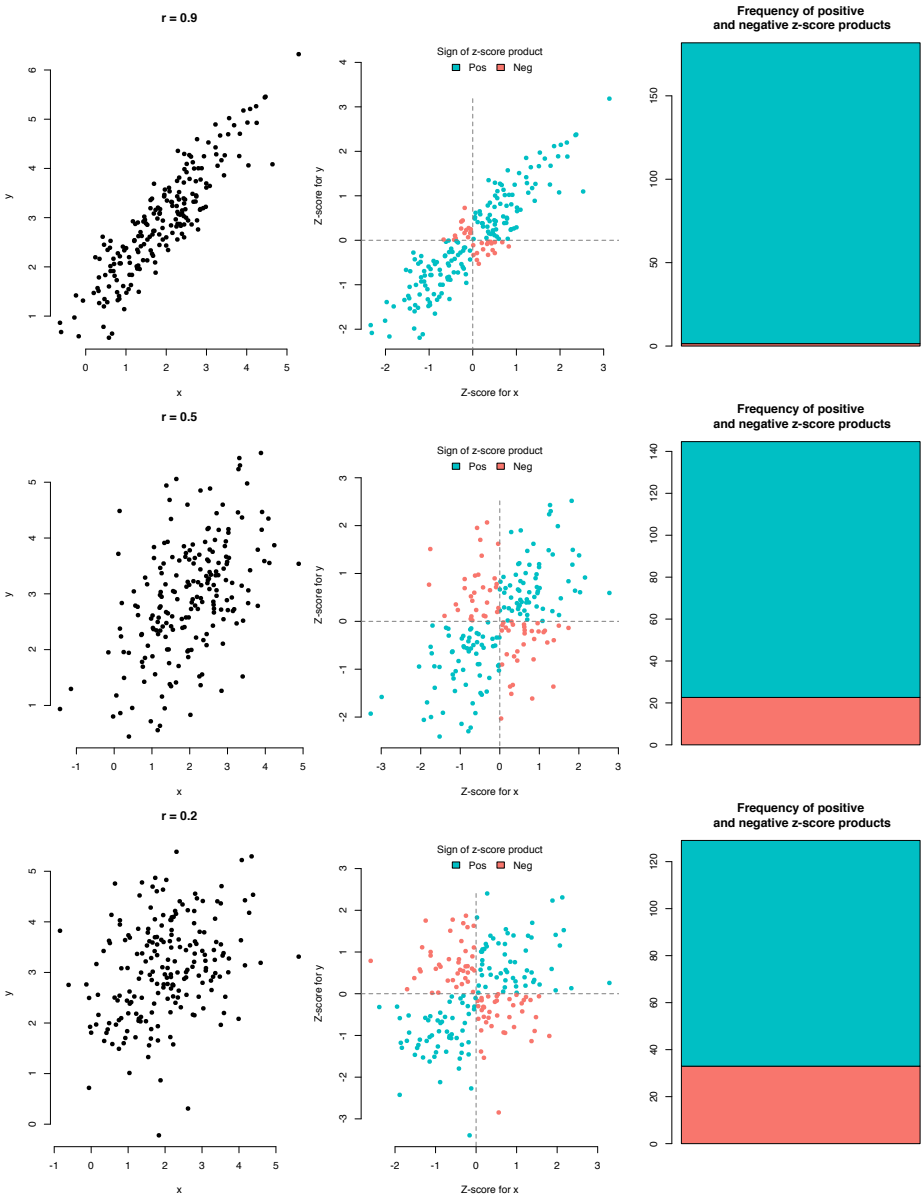
The value is much larger and more positive when all data points are used. The points that are far from the means of x and y have a larger product of z-scores and thus increase the correlation coefficient value.

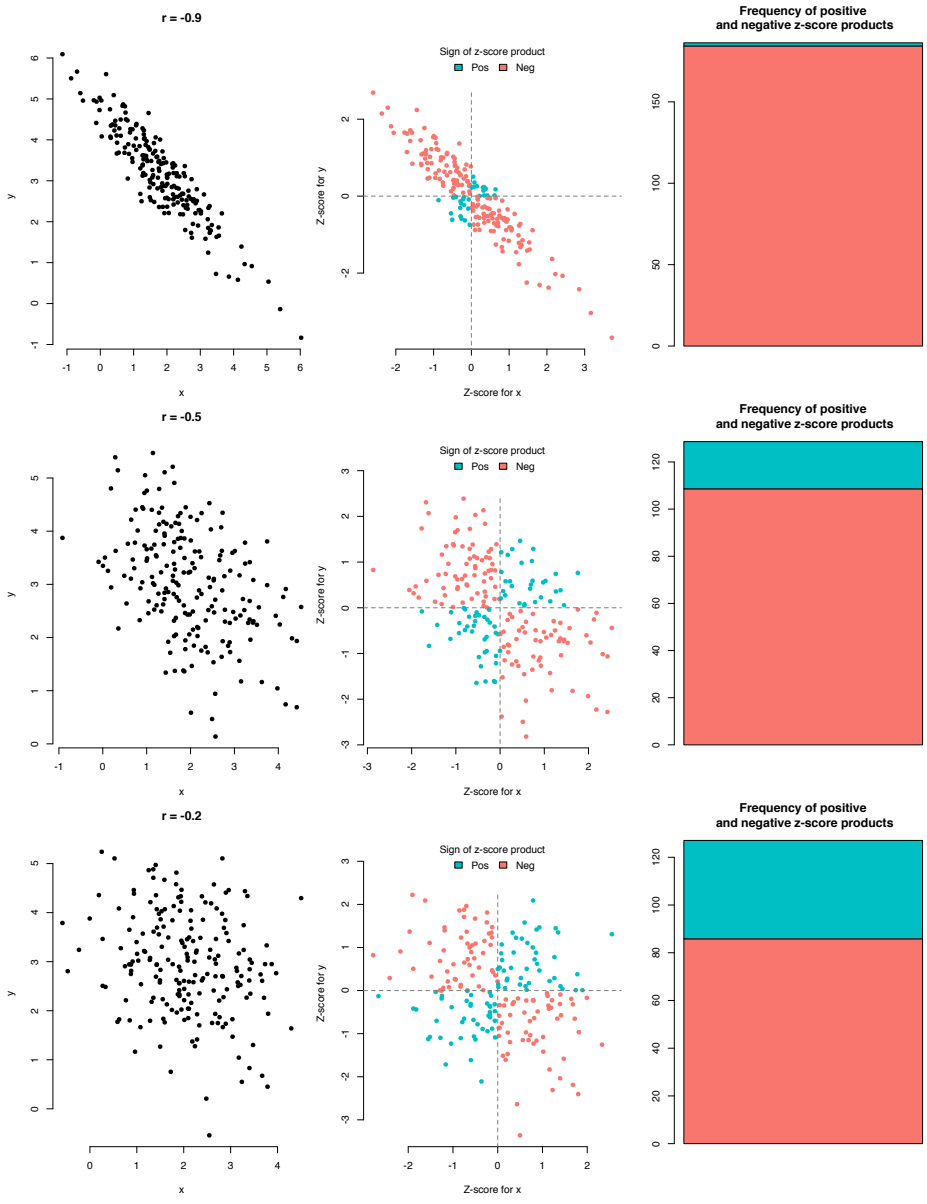
2.8.3 Properties

- $-1 \leq r \leq 1$ (due to the Cauchy-Schwarz Inequality for those are inclined)
- The sign of r goes with the direction of the relationship.
- $r_{x,y} = r_{y,x}$, it doesn't matter which variable is x and which is y.
- $r_{ax+b,cy+d} = r_{x,y}$, linear change of scale doesn't affect r. Why?
- r measures strength of *linear* relationship (not a curved relationship).
- One outlier can completely change r.

Let's look at a few scatterplot examples and the corresponding correlation.




r = 0.9
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products
r = 0.5
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products
r = 0.2
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products


r = -0.9
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products
r = -0.5
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products
r = -0.2
x
y
Sign of z-score product
Pos
Neg
Z-score for x
Z-score for y
Frequency of positive
and negative z-score products

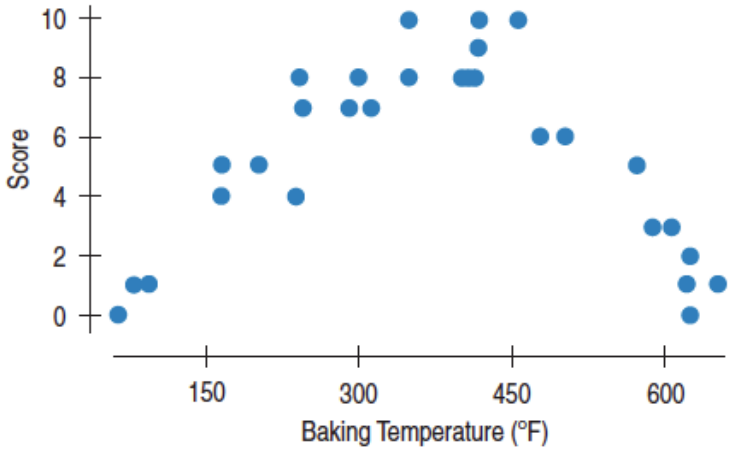
Math Box: (Optional) Here are other equivalent expressions for r for the mathematically intrigued:

$$
\begin{aligned}
r &= \frac{\sum z_x z_y}{n-1} \\
&= \frac{\sum \frac{(x_i-\bar{x})}{s_x} \times \frac{(y_i-\bar{y})}{s_y}}{n-1} \\
&= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} \\
&= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{(n-1)\sqrt{\sum \frac{(x_i-\bar{x})^2}{n-1}}\sqrt{\sum \frac{(y_i-\bar{y})^2}{n-1}}} \\
&= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}} \\
&= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}
\end{aligned}
$$

2.8.4 Is correlation always the right way to judge strength?

The plot below shows the relationship between brownie quality and oven temperature at which the brownie is baked.

The correlation coefficient is near 0, but it doesn't mean that there's no relationship. We can clearly see a quadratic relationship, but there's not a **linear** relationship.



The correlation coefficient, r, is more formally called the Pearson correlation coefficient, named after Karl Pearson who published this work in 1895. Read more about this measure of linear relationship here.

2.9 Three or more variables

In complex data sets that contain many variables, it is necessary to get a fuller understanding of the relationships in the data than we can see with plots that look at only one or two variables.

The following visual features can help us turn bivariate plots into multivariate plots:

- Color of points and lines
- Shape of points
- Size of points
- Panels (facets)

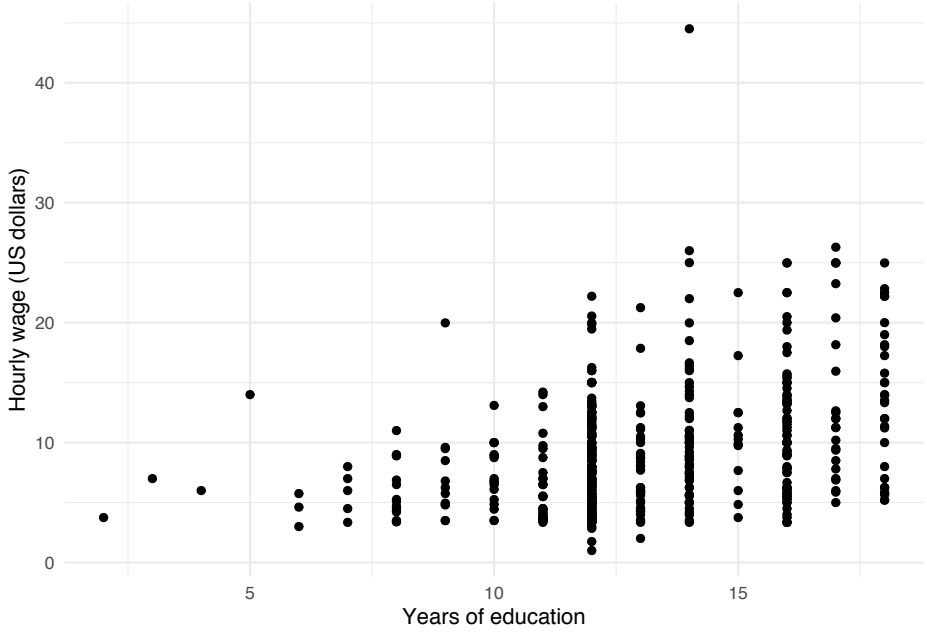
Let's look at another data set, the 1985 Current Population Survey. This is a smaller scale survey administered by the United States government in the intervening years of the decennial census.

2.9.1 A bivariate scatterplot

Our primary interest is the `wage` variable which gives the hourly wage for each individual in the data set in US dollars. What is the relationship between years of education and hourly wage?

```
data(CPS85) # Load the data from the R package CPS85

CPS85 %>%
```

```
    ggplot(aes(x = educ, y = wage)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    theme_minimal()
```

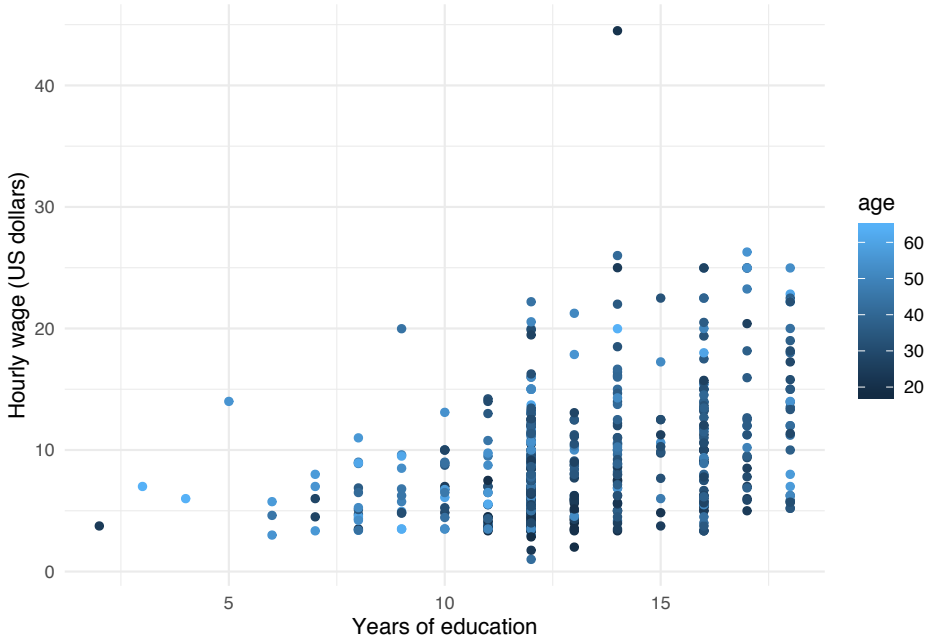


We can see that years of education and hourly wage are positively correlated. What about the impact of other variables?

2.9.2 Enriching with color

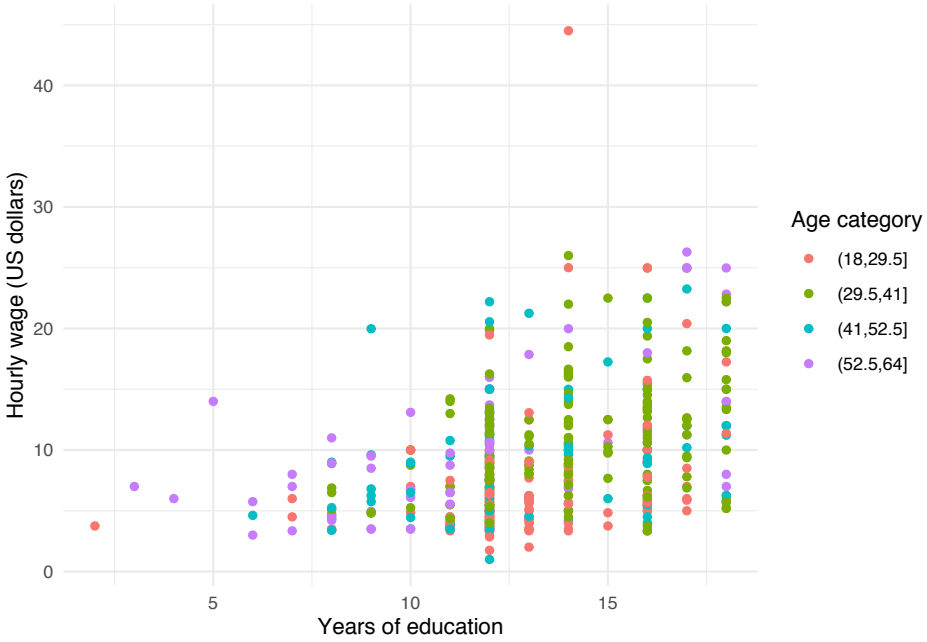
We can enrich this bivariate scatterplot by showing additional information via color.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage, color = age)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    theme_minimal()
```



Adding color for a quantitative variable, age, does not reveal any obvious patterns; that is, we don't see obvious clustering by color. Perhaps this is because there are too many colors (remember Visualization Principle #6: Use Color Appropriately). Are any patterns revealed if we use 4 age categories instead?

```
CPS85 %>%
    mutate(age_cat = cut(age, 4)) %>%
    ggplot(aes(x = educ, y = wage, color = age_cat)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    guides(color = guide_legend(title = "Age category")) + # Legend title
    theme_minimal()
```

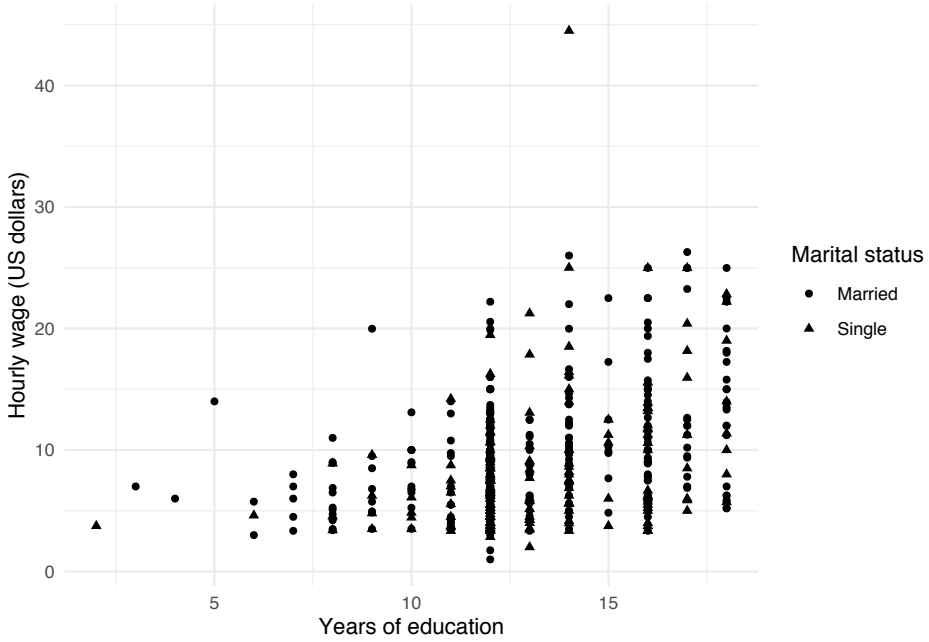


With 4 age categories, no age patterns are evident, but this does help us see that the least educated people in this data set are mostly in the youngest and oldest age categories.

2.9.3 Enriching with shape

We can also encode information via point shape. Here we let shape encode marital status.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage, shape = married)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    guides(shape = guide_legend(title = "Marital status")) + # Legend title
    theme_minimal()
```

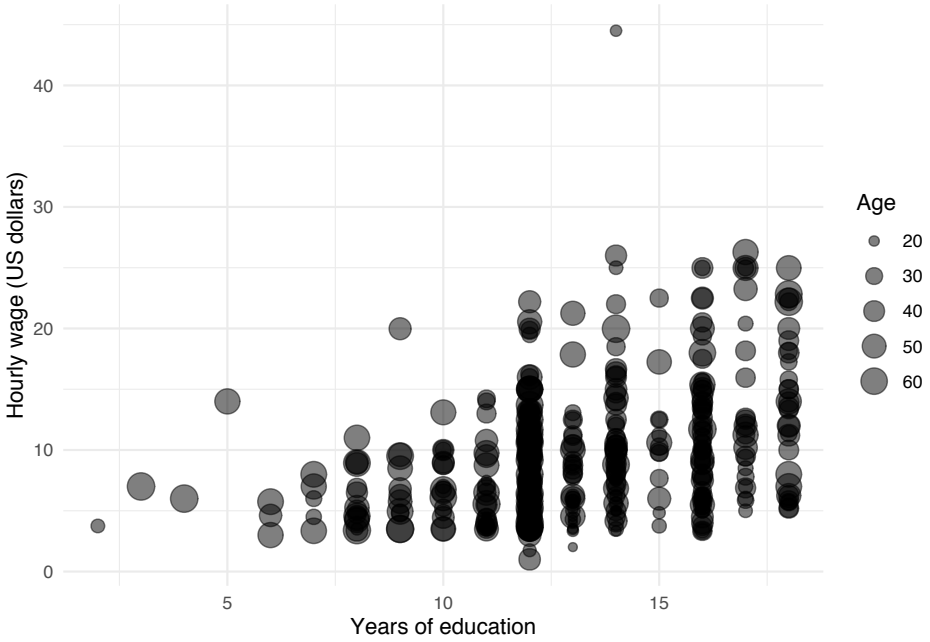


Often encoding information with color is preferable to encoding it with shapes because differences in shapes are not as easily discernible. Remember that statistical visualizations are meant to help you better understand your data. If you are having trouble easily picking out patterns when using a certain visual feature (e.g. shape, color), try another feature to see if the clarity of the plot increases for you.

2.9.4 Enriching with size

The size of a point is useful for conveying the magnitude of a quantitative variable. For example, we may wish to see non-categorized age information with point size.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage, size = age)) +
    geom_point(alpha = 0.5) + #alpha specifies the level of transparency of the points
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    guides(size = guide_legend(title = "Age")) + # Legend title
    theme_minimal()
```

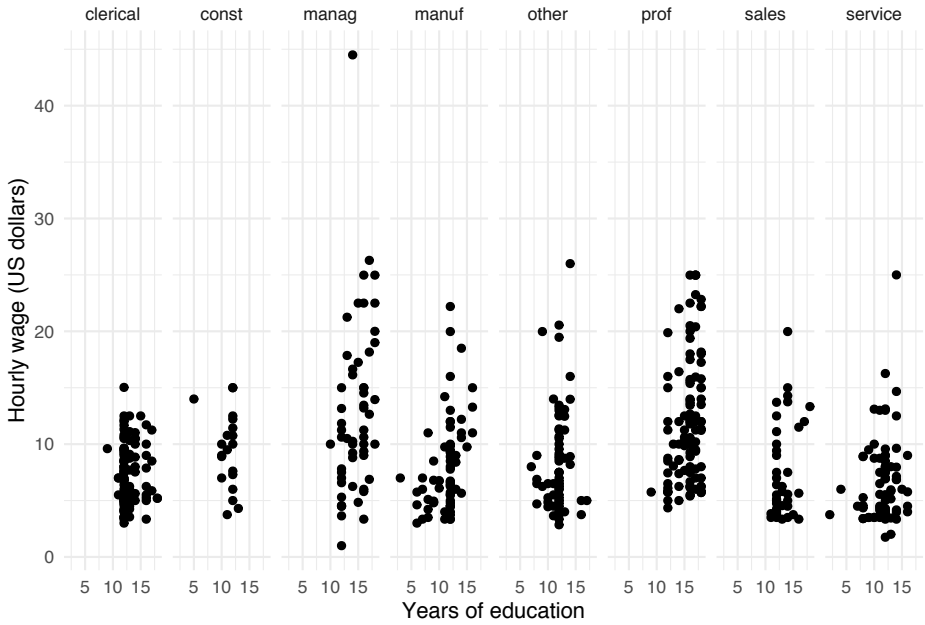


2.9.5 Enriching with panels

Panels (or facets) are a great way to see how relationships differ between levels of a single categorical variable or between combinations of two categorical variables.

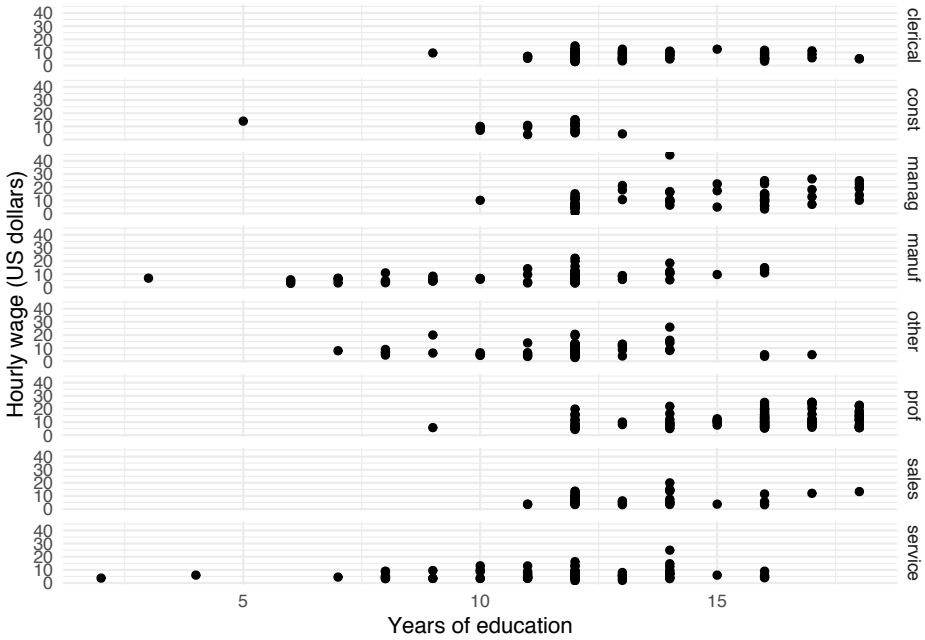
Let's look at the relationship between hourly wage and years of education across job sectors. The following creates a row of plots of this relationship over job sectors.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    facet_grid(. ~ sector) +
    theme_minimal()
```



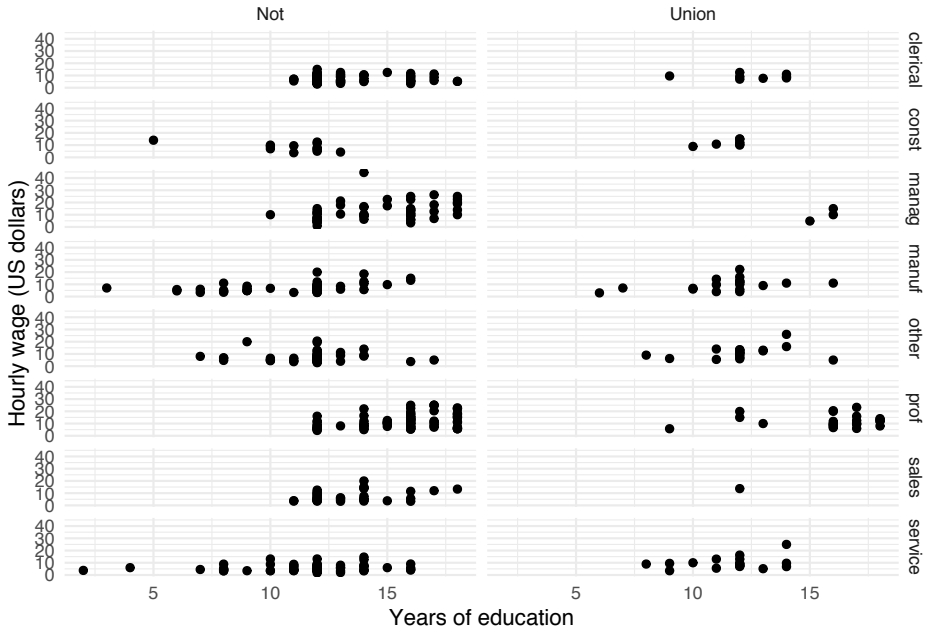
With a small change in notation (`sector ~ .` versus `. ~ sector`), we can create a column of plots.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    facet_grid(sector ~ .) +
    theme_minimal()
```



We can also create panels according to two categorical variables. How do the relationships additionally differ by union status?

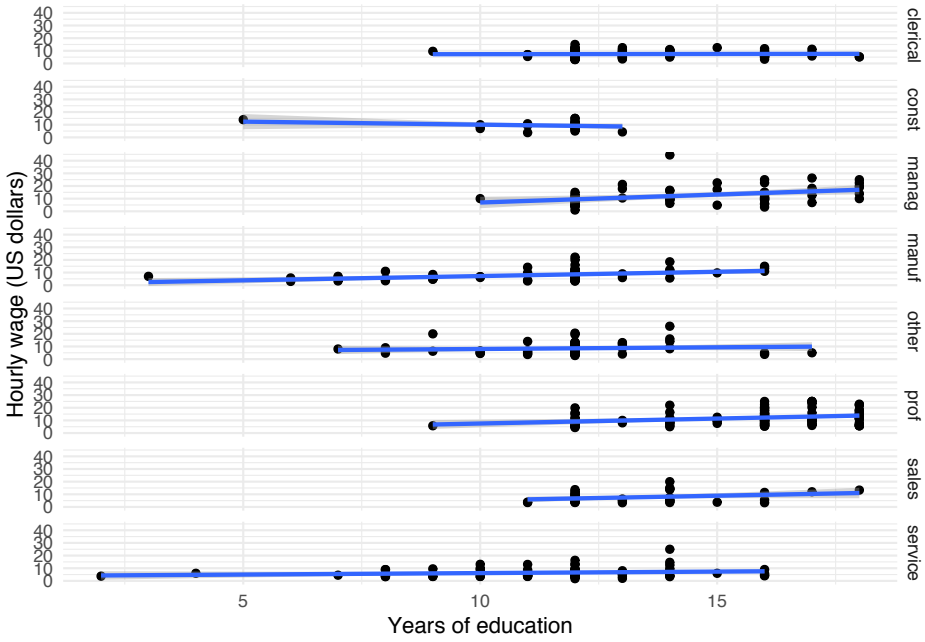
```
CPS85 %>%
    ggplot(aes(x = educ, y = wage)) +
    geom_point() +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    facet_grid(sector ~ union) +
    theme_minimal()
```



2.9.6 Enriching with smoothing

If we have a scatterplot, we may want to get an understanding of the overal relationship between x and y within subsets. We can add `geom_smooth(method = 'lm')` to estimate and plot the linear relationships.

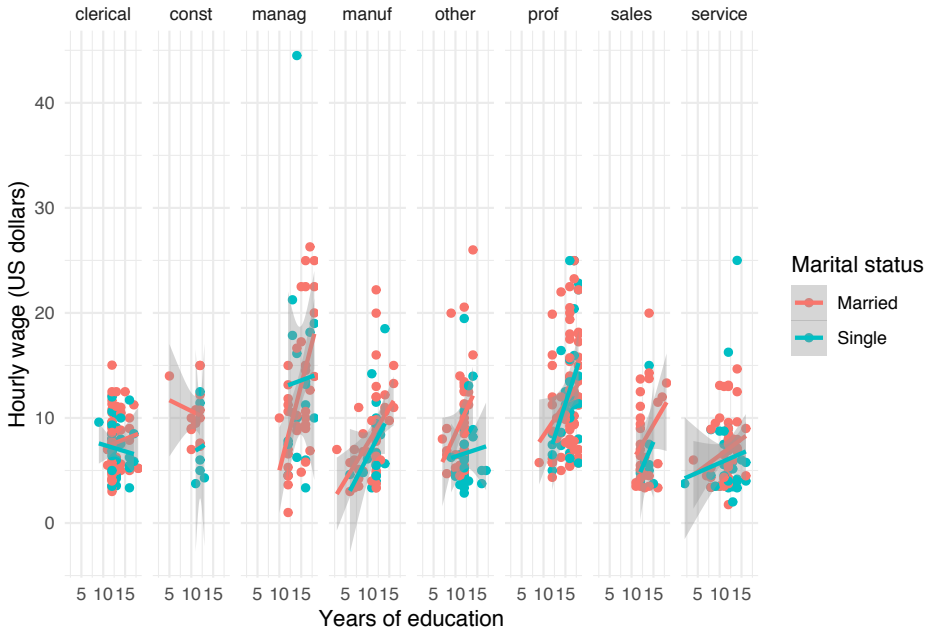
```
CPS85 %>%
    ggplot(aes(x = educ, y = wage)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    facet_grid(sector ~ .) +
    theme_minimal()
```



2.9.7 Putting everything together

The combination of these different visual features can result in powerful visual understanding. Let's combine paneling with color information to explore if there are marital status patterns in these union-job sector subgroups.

```
CPS85 %>%
    ggplot(aes(x = educ, y = wage, color = married)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    xlab("Years of education") +
    ylab("Hourly wage (US dollars)") +
    facet_grid(. ~ sector) +
    guides(color = guide_legend(title = "Marital status")) + # Legend title
    theme_minimal()
```



Creating effective multivariate visualizations takes a lot of trial and error. Some visual elements will better highlight patterns than others, and often times, you'll have to try several iterations before you feel that you are learning something insightful from the graphic. Be tenacious, and keep in mind the good visualization principles outlined at the beginning of this chapter!

2.10 Major Takeaways

1. STOP: Think about whether a variable is categorical or quantitative. This informs the type of graphic and summaries that are appropriate.
2. The shape of a histogram tells you about the relationships between mean and median.
3. If you are describing a histogram, make sure to comment on the shape, center, spread, and outliers.
4. If you are describing a scatterplot, make sure to comment about the direction, form, strength, and unusual features.
5. Be vigilant for unusual points as they could be due to human error in the data collection, but do not automatically throw away unusual or outlying points. Think about whether they are feasible first.
6. Visualizations are just a starting place. Stop to notice what you learn and what questions you have about the data. Let that inform the next visualization you make.

Chapter 3

Regression Models

In our visualization of data, we've seen trends, patterns, and variation. Let's now endeavor to describe those trends, patterns, and variation more *precisely* and *quantitatively* by building statistical *models*.

3.1 Modeling Goals

Broadly, a model is a simplified representation of the world. When we build models, we may have different goals.

One goal when building models is **prediction**. Given data on a **response or outcome variable**, y, and one or more **predictor or explanatory variables**, x, the goal is to find a mathematical function, f, of x that gives good predictions of y. For example, we might want to be able to predict a customer's chest size knowing their neck size. This x may be a single variable, but it is most often a set of variables. We'll be building up to multivariate modeling over the course of this chapter.

> **Reflect:** Can you think of some other concrete examples in which we'd want a model to do prediction? Consider what predictions might be made about you every day...

What are the qualities of a good model and function, f? We want to find an $f(x)$ such that $\hat{y} = f(x)$ is a good predictor of y. In other words, we want the model prediction $\hat{y}$ to be close to the observed response. We want $y - \hat{y}$ to be small. This difference between the observed value and the prediction, $y - \hat{y}$, is called a **residual**. We'll discuss residuals more later.

Another goal when building models is **description**. We want a model, $f(x)$, to "explain" the relationship between the x and y variables. Note that an

overly complicated model may not be that useful here because it can't help us *understand* the relationship. A more complex model may, however, produce better predictions. George Box is often quoted "All models are wrong but some are useful." Depending on our goal, one model may be more useful than another.

> **Reflect:** Can you think of some concrete examples in which we'd want a model to do explain a phenomenon? Consider how policy decisions get made...

To begin, we will consider a simple, but powerful model in which we limit this function, $f(x)$, to be a straight line with a y-intercept, b_0, and slope, b_1.

$$\hat{y} = f(x) = b_0 + b_1 x$$

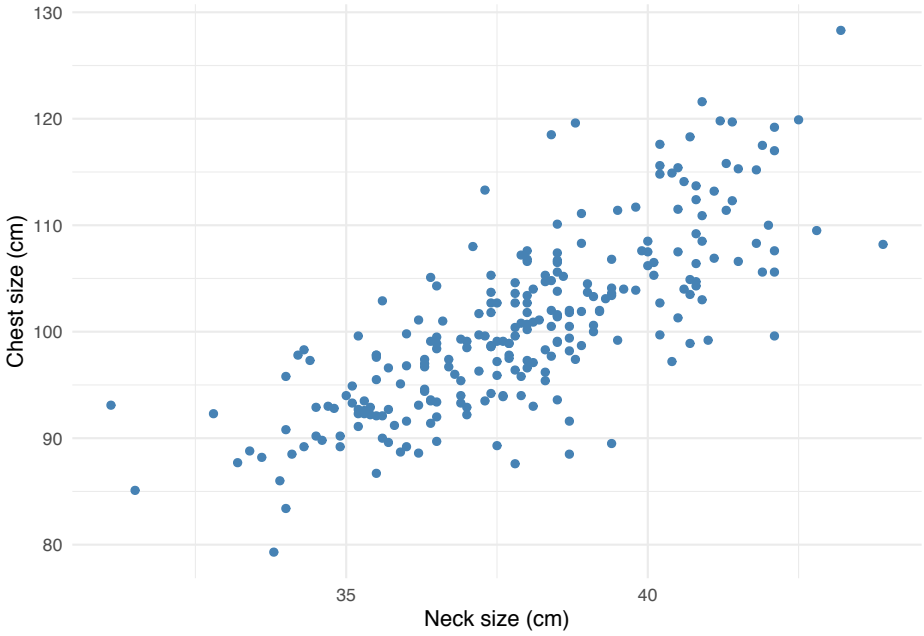
This is a **simple linear regression model**. It is the foundation of many statistical models used in modern statistics and is more flexible than you may think.

> **Math Box:** In the past, you may have seen the equation of a line as
>
> $$y = mx + b$$
>
> where m is the slope and b is the y-intercept. We will be using different notation so that it can generalize to multiple linear regression.
> The y-intercept is the value when $x = 0$ and the slope is change in y for each 1 unit increase of x ("rise over run").

3.2 Lines

Let's return to the thought experiment in which you were a manufacturer of button-down dress shirts.

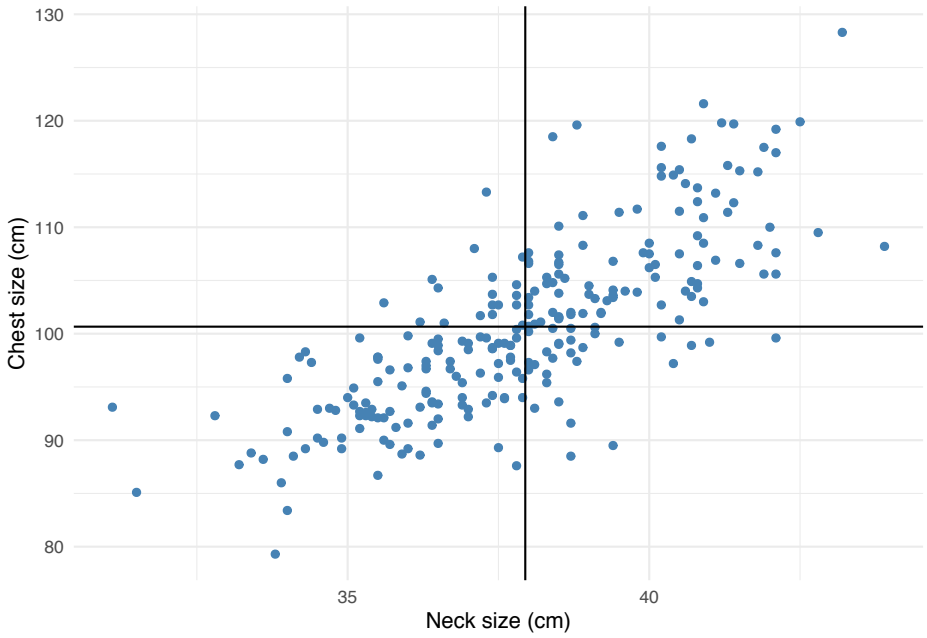
```
body <- read.delim("Data/bodyfat.txt")

body %>%
    ggplot(aes(x = Neck, y = Chest)) +
    geom_point(color = 'steelblue') +
    xlab('Neck size (cm)') +
    ylab('Chest size (cm)') +
    theme_minimal()
```



Reflect: If you were to add one or multiple lines to the plot above to help you make business decisions, where would you want it (or them)?

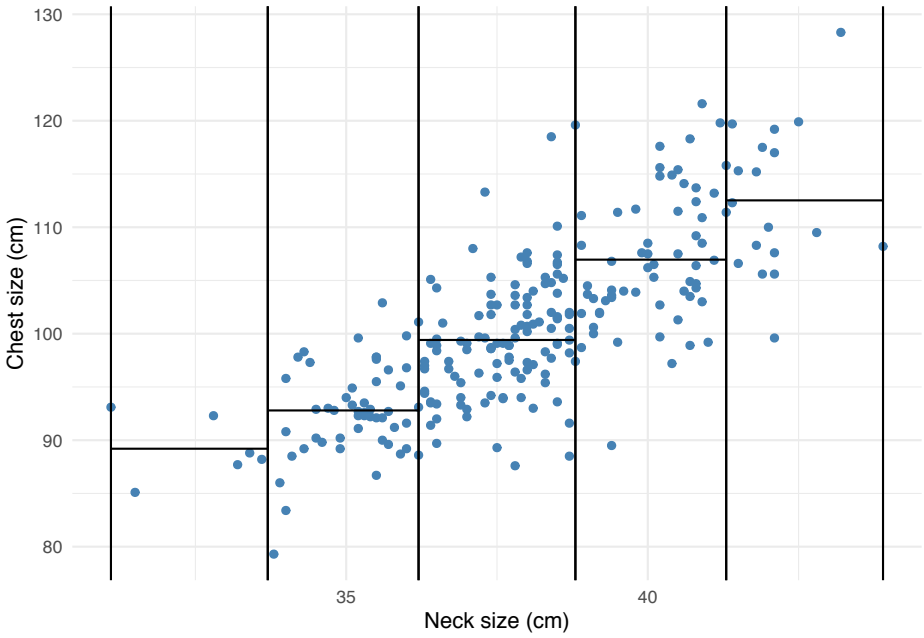
Let's say you were only going to make one size of shirt. You might want to add a horizontal line at the mean Chest size and a vertical line at the mean Neck size.

```
body %>%
    ggplot(aes(x = Neck, y = Chest)) +
    geom_point(color = 'steelblue') +
    geom_hline(yintercept = mean(body$Chest)) +
    geom_vline(xintercept = mean(body$Neck)) +
    xlab('Neck size (cm)') +
    ylab('Chest size (cm)') +
    theme_minimal()
```



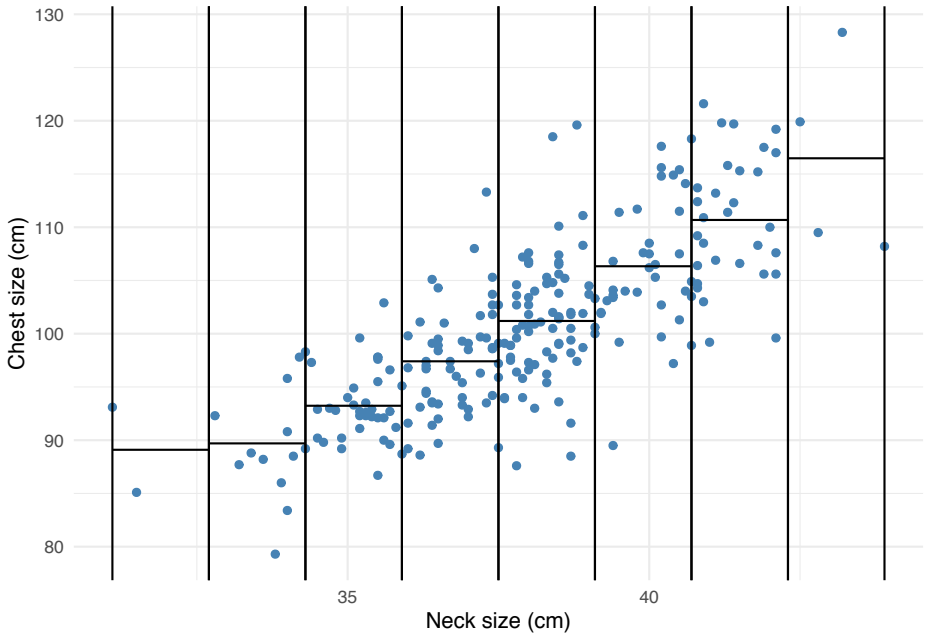
We can see that a shirt made to these specifications would fit the "average person." However, this might not serve your market very well. For many people, the shirt would be too tight because their chest and/or neck sizes would be larger than average. For many people, the shirt would be too large because they chest and/or neck sizes would be smaller than average.

Let's try something else. Let's allow ourselves 5 different sizes (XS, S, M, L, XL). Then, we can cut the Neck sizes variable into 5 groups of equal length and estimate the mean Chest sizes within each of these groups.



Reflect: What do these lines tell us for our business venture?

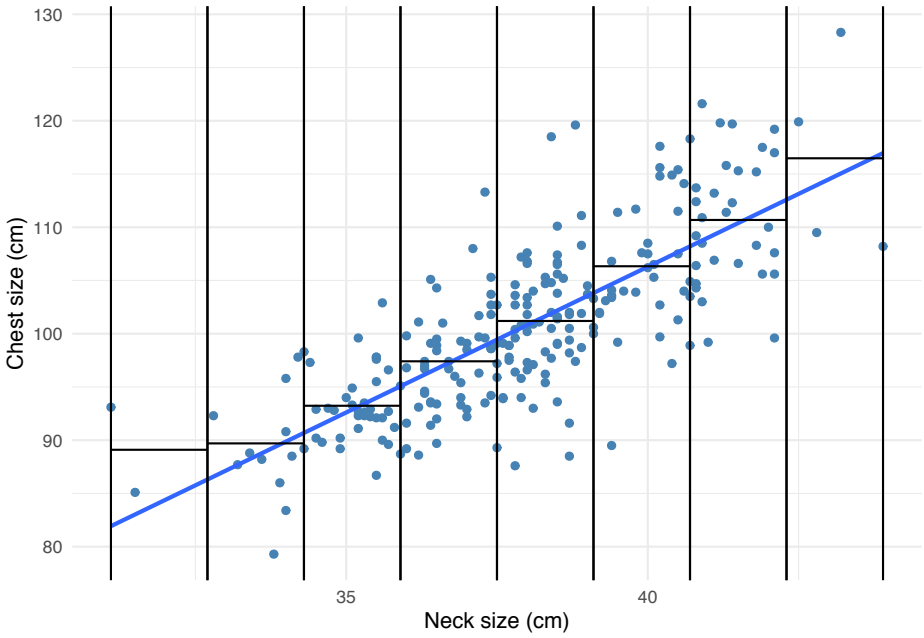
What if we wanted to be able to make more sizes? Could we get a pretty good sense of what the chest sizes should be for a given neck size? Let's try allowing for 8 different sizes.



Reflect: What are the pros and cons of having a larger number of sizes?

Reflect: Stop and think about the data collection process. If you were measuring your own neck size, how precise do you think you could get? What factors might impact that precision?

We notice that there is a generally linear relationship between neck and chest size. Perhaps we find one line to describe the relationship between Neck size and Chest size and use that to decide on sizes later?



Reflect: What does line tell us for our business venture?

If the scatterplot between two quantitative variables **resembles a straight line**,

- a straight line could roughly **describe** the mean of y for each value of x.
- a straight line could **describe** how much we'd *expect* y to change based on a 1 unit change in x.
- a straight line could help us **predict** the y based on a new value of x.

3.3 "Best" fitting line

To choose the "best" fitting line, we need to choose the intercept (b_0) and slope (b_1),

$$\hat{y} = f(x) = b_0 + b_1 x$$

that gives us the "best" fit to the n points on a scatterplot, (x_i, y_i) where $i = 1, ..., n$.

Reflect: What do we mean by "best"? In general, we'd like good predictions and a model that describes the average relationship. But we need to be more precise about what we mean by "best".

3.3.1 First idea

One idea of "best" is that we want the line that minimizes the sum of the **residuals**, $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$. The residual is the error in our prediction, the difference between what you observe and what you predict based on the line.

- **Problem**: We will have positive and negative residuals; they will cancel each other out if we add them together. While a good idea, this definition of "best" won't give us what we want. We'll want an idea that deals with the negative signs.

3.3.2 Second idea

Another idea of "best" is that we want the line that minimizes the sum of the absolute value of the residuals, $\sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |e_i|$.

- **Problem**: This definition of "best" results in a procedure referred to as **Least Absolute Deviations**, but there isn't always one unique line that satisfies this. So, while this is a valid definition of "best," this isn't stable as there isn't always one "best" line.

3.3.3 Third idea

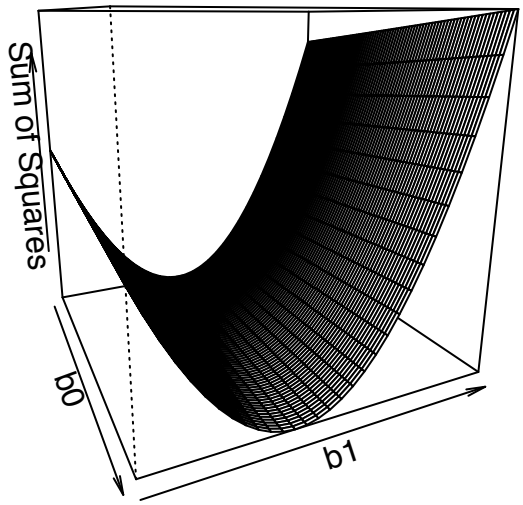
Lastly, another idea of "best" is that we want the line that minimizes the sum of squared residuals, $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$.

- This is referred to as **Least Squares** and has a unique solution. We'll will focus on this definition of "best" in this class. It also has some really nice mathematical properties and connections to linear algebra.

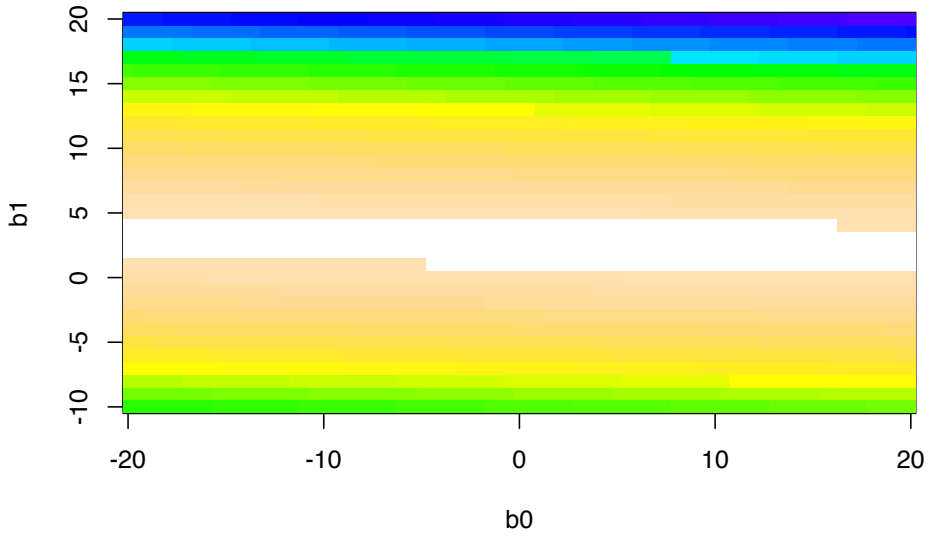
3.4 Least Squares

Let's try to find the line that **minimizes the Sum of Squared Residuals** by searching over a grid of values for (intercept, slope).

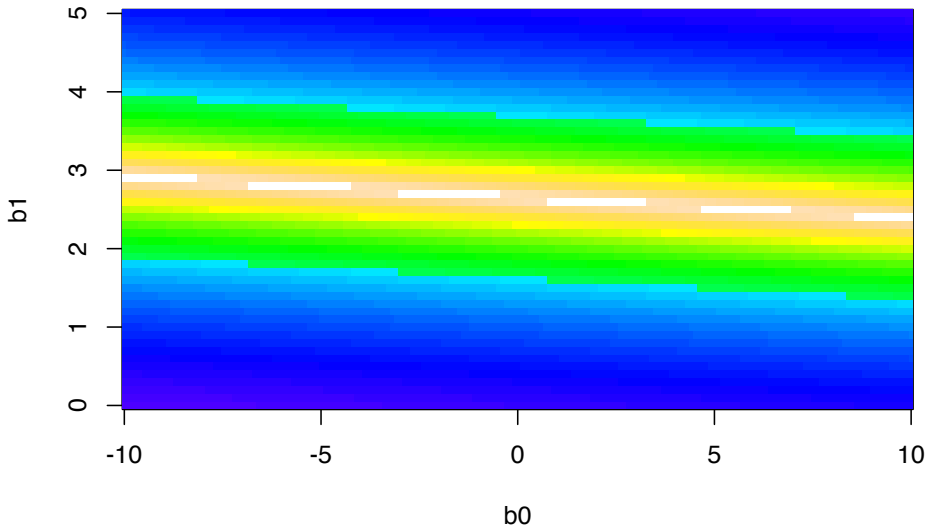
Below is a visual of the sum of squared residuals for a variety of values of the intercept and slope. The surface height is sum of squared residuals for each combination of slope and intercept.



We can see there is valley where the minimum must be. Let's visualize this in a slightly different way. We'll encode the surface height as color (white is lowest).



The large values of the sum of squared residuals are dominating this image, so let's change the color scheme to see more variation in smaller values (white is lowest).



We can limit our search to $b_0 \in (-10, 10)$ and $b_1 \in (2, 3)$.

```
b0 = seq(-10,10,by=.05)
b1 = seq(2,3,by=.05)
b <- expand.grid(b0,b1)
ss <- apply(b,1,f)

b[ss == min(ss),]
```

```
##      Var1 Var2
## 6142 -3.7 2.75
```

We have the minimum point. Over the grid of pairs of values, the minimum sum of squared residuals happens when the intercept is -3.7 and the slope is 2.75.

Math Box: (Optional) Alternative ways (faster than exhaustive search) to find the minimum sum of squared residuals

- We could try a numerical optimization algorithm such as steepest descent.
- We could use multivariable calculus (find partial derivatives, set equal to 0, and solve).

To get started on the calculus, solve the following two equations for the two unknowns (b_0 and b_1):

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i-(b_0+b_1x_i))^2=0$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y_i-(b_0+b_1x_i))^2=0$$

If you are a math/stat/physics/cs major, you should try this by hand and see if you can get the solutions below.

If you find the minimum using calculus (super useful class!), you'll find that we can write the Least Squares solution in an equation format as functions of summary statistics (!), the estimated slope is

$$b_1 = r\frac{s_y}{s_x}$$

and the estimated intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{x}$ is the mean of the variable on the x-axis, $\bar{y}$ is the mean of the variable on the y-axis, s_x is the standard deviation of the variable on the x-axis, s_y is the standard deviation of the variable on the y-axis, and r is the correlation coefficient between the two variables.

Let's do that calculation "by hand" first in R.

```
body %>%
  summarize(sy = sd(Chest), sx = sd(Neck), r = cor(Chest,Neck), ybar = mean(Chest), xbar = mean(N
  mutate(b1 = r*sy/sx, b0 = ybar - b1*xbar) %>%
  select(b0, b1)
```

```
##      b0   b1
## 1 -3.19 2.74
```

Wow. That was quite a bit of coding. From now on, we'll take the shortcut and use the `lm()` function which stands for **l**inear **m**odel. This function gives us the

Least Squares solution to the "best" fitting line, as defined by minimizing the sum of squared residuals.

```
lm(Chest ~ Neck, data = body) # When you see ~, think 'as a function of'
```

```
##
## Call:
## lm(formula = Chest ~ Neck, data = body)
##
## Coefficients:
## (Intercept)         Neck
##       -3.19         2.74
```

3.5 Properties of Least Squares Line

- $(\bar{x}, \bar{y})$ is ALWAYS on the least squares line.
- The residuals from the least squares line ALWAYS sum to 0.
- The mean of the residuals from the least squares line is ALWAYS 0.
- The **standard deviation of the residuals** gives us a sense of how bad our predictions (based on the line) could be.

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-2}}$$

In R: the standard deviation of the residuals, s_e, is referred to as the "residual standard error". Don't confuse this with "Std. Error," that is something else that we'll get to.

```
body %>%
  lm(Chest ~ Neck, data = .) %>% #If you don't want the pipe to pass to the first argu
  summary()
```

```
##
## Call:
## lm(formula = Chest ~ Neck, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.145  -3.333  -0.534   3.074  16.597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.188      5.495   -0.58     0.56
## Neck           2.737      0.145   18.93   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.22 on 248 degrees of freedom
## Multiple R-squared:  0.591,  Adjusted R-squared:  0.589
## F-statistic:  358 on 1 and 248 DF,  p-value: <2e-16
```

This simple linear regression line is

$$\widehat{Chest} = -3.18 + 2.73 * Neck$$

If you have a neck size of 38 cm, then we predict the chest size of an ideal shirt is about 100.5 cm (~ -3.18 + 2.73*38)

```
-3.18 + 2.73*38 #the intercept and slope are rounded here first (not great)
```

```
## [1] 101
```

```
body %>%
  lm(Chest ~ Neck, data = .) %>%
  predict(newdata = data.frame(Neck = 38)) #the intercept and slope are not rounded before predic
```

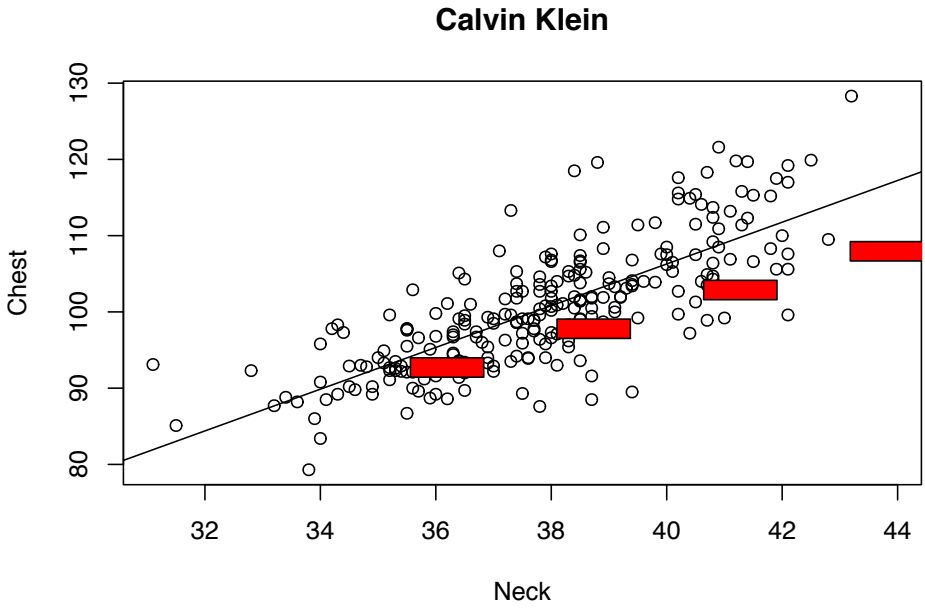
```
##   1
## 101
```

Given your neck size, we can probably predict your chest size within 5 to 10 cm since $s_e = 5.22$ (1 to 2 SD's – recall Section 2.6.5).

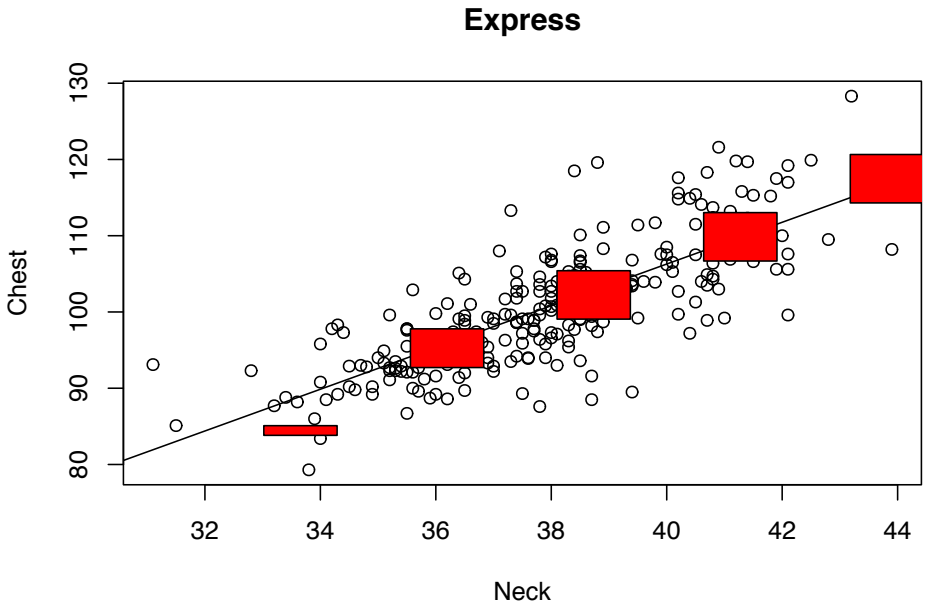
> **Reflect:** If you were a shirt manufacturer, is this a good enough prediction? What is the impact on the customer? Think of if the prediction were an overestimate (loose) or an underestimate (too tight).

3.5.1 Real companies

Let's see how some real companies create shirts. In the plots below, the red boxes represent the advertised range (in cm) for Neck and Chest sizes for each brand.

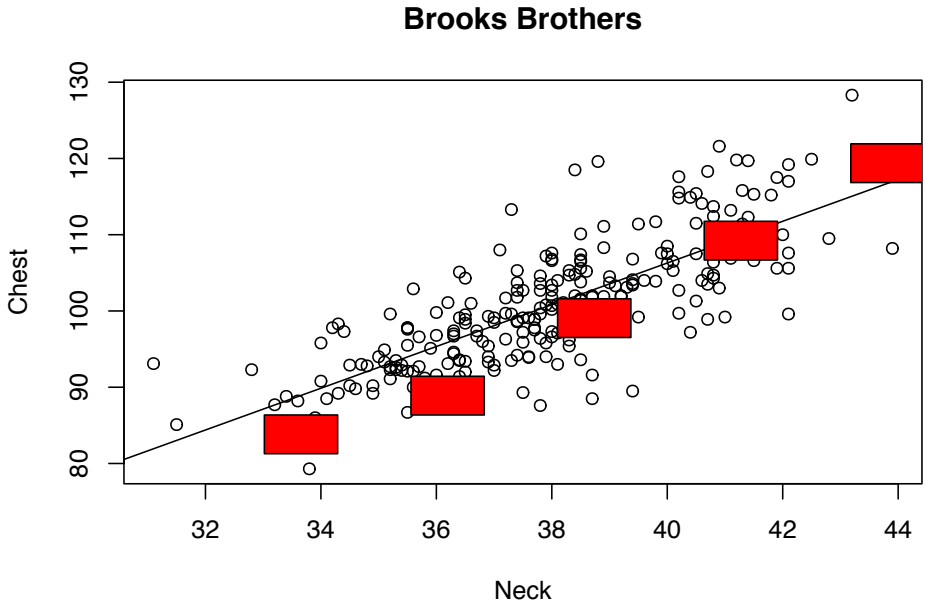


For Calvin Klein, we see that the red boxes are below the least squares line (black line). So for a given neck size, Calvin Klein makes shirts that are a little bit too small at the chest.



For Express, we see that the red boxes are generally on the least squares line, except for the smallest size. This means that Express shirts are generally a good fit at the chest and neck for the 4 largest sizes, but the smallest shirt size is a

bit too small at the chest for the neck size.



For Brooks Brothers, the red boxes are a bit below the least squares line for the 3 smallest sizes and a little above the line for the largest size. This means that the 3 smallest sizes are a bit too small in the chest for our customers (in our data set) with those neck sizes and that the largest shirt is a bit big at the chest for that neck size.

> **Reflect:** We haven't told you how the customer data we've been using was collected. As you compared the brands to this sample data, what assumptions were you making about the population that the sample was drawn from?
> What questions do you have about the sampling procedure?

3.6 Interpretation

Let's look at the summary print out of the `lm()` function in R again. We'll highlight some of the most important pieces of the print out and discuss how you interpret them.

```
body %>%
  lm(Chest ~ Neck, data = .) %>%
  summary()
```

```
##
## Call:
```

```
## lm(formula = Chest ~ Neck, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.145  -3.333  -0.534   3.074  16.597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.188      5.495   -0.58     0.56
## Neck           2.737      0.145   18.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.22 on 248 degrees of freedom
## Multiple R-squared:  0.591,  Adjusted R-squared:  0.589
## F-statistic:  358 on 1 and 248 DF,  p-value: <2e-16
```

3.6.1 Intercept (b_0)

- **Where to find:** In the table called "Coefficients", look for the number under "Estimate" for "(Intercept)". It is -3.1885 for this model.
- **Definition:** The intercept gives the average value of y when x is zero (**think about context**)
- **Interpretation:** If Neck size = 0, then the person doesn't exist. In this context, the intercept doesn't make much sense to interpret.
- **Discussion:** If the intercept doesn't make sense in the context, we might refit the model with the x variable once it is **centered**. That is, the mean of x for all individuals in the sample is computed, and this mean is subtracted from each person's x value. In this case, the intercept is interpreted as the average value of y when x is at its (sample) mean value. See the example below – we get an intercept of 100.66, which is the average Chest size for customers with average Neck size.

```
body %>%
  mutate(CNeck = Neck - mean(Neck)) %>%
  lm(Chest ~ CNeck, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = Chest ~ CNeck, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.145  -3.333  -0.534   3.074  16.597
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  100.662      0.330   304.6   <2e-16 ***
## CNeck          2.737      0.145    18.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.22 on 248 degrees of freedom
## Multiple R-squared:  0.591,  Adjusted R-squared:  0.589
## F-statistic:  358 on 1 and 248 DF,  p-value: <2e-16
```

3.6.2 Slope (b_1)

- **Where to find:** In the table called "Coefficients", look for the number under "Estimate" for the namve of the variable, "Neck" or "CNeck". It is 2.7369 for this model.
- **Definition:** The slope gives the change in average y per 1 unit increase in x (**not for individuals and not causal statement**)
- **Interpretation:** If we compare two groups of individuals with a 1 cm difference in the neck size (e.g. 38 cm and 39 cm OR 40 cm and 41 cm), then we'd expect the average chest sizes to be different by about 2.7cm.
- **Another Interpretation:** We'd expect the average chest size to increase by about 2.7 cm for each centimeter increase in neck size.
- **Discussion:** Note that both interpretations are not written about an individual because we can't easily change our neck size by 1 cm. The slope describes the change in the average in our sample data set, not the change for one person.

3.6.3 Least Squares/Regression Line ($\hat{y} = b_0 + b_1 x$)

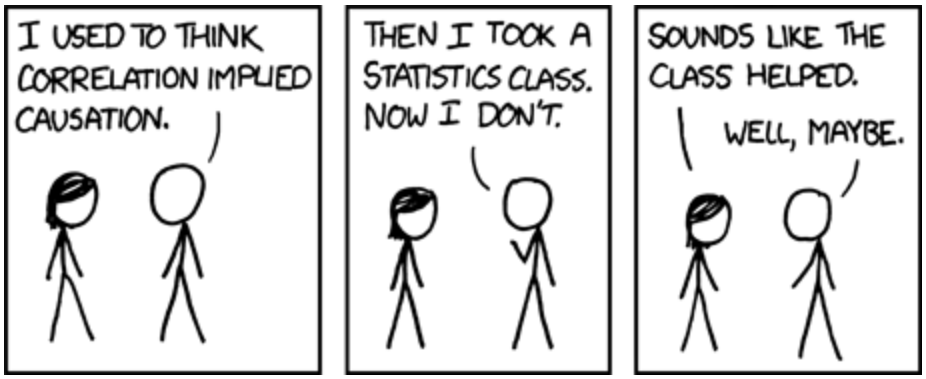
- **Where to find:** In the table called "Coefficients", all of the values you need to write down the line are under "Estimate".

- **Definition:** The line gives the estimated average of y for each value of x (**within the observed range of x**)

- **Interpretation:** The regression line of (Predicted Chest = -3.18 + 2.73*Neck) gives the estimated average Chest size for a given Neck size, based on our sample of data.

- **Discussion:** The "within the observed range of x" phrase is very important. We can't predict values of y for values of x that are very different from what we have in our data. Trying to do so is a big no-no called **extrapolation**. We'll discuss this more in a bit.

3.6.4 Correlation or Association vs. Causation



What is causation? What is a causal effect? Are there criteria for defining a cause?

These are deep questions that have been debated by scientists of all domains for a long time.We are at a point now where there is some consensus on the definition of a causal effect. The causal effect of a variable x on another variable y is the amount that we expect y to change if we **intervene on/manipulate** x by changing it by one unit.

When can we interpret an estimated slope b_1 from a simple linear regression as a causal effect of x on y? Well, in the simple linear regression case, almost never. To interpret the slope causally, there would have to be **no confounding variables** (no variables that impact both x and y). If we performed an experiment, this might be possible, but in most cases, it will not be the possible.

Then what are we to do?

1. Think about the possible causal pathways and try and create a DAG (from the first chapter).
2. Try to **adjust/control for** possible confounders using multiple linear regression (coming up soon).
3. After fitting a model, we should step back and consider other criteria such as Hill's Criteria before concluding there is a cause and effect relationship. Think about whether there is a large effect, whether it can be reproduced in another sample, whether the the effect happens before the cause in time, etc.

Multiple linear regression offers us a way to estimate causal effects of variables *if we use it carefully*. It will be tempting to say: control for everything we can! But we will see that this is not the correct thing to do, as there are very real dangers of over-controlling for variables (known as throwing everything in the model). For now, let it suffice to say that multiple linear regression is as useful as the corn kerneler below. Immensely useful in the right circumstances - but only those circumstances.

Figure 3.1: Source: Walmart

3.7 Model Evaluation

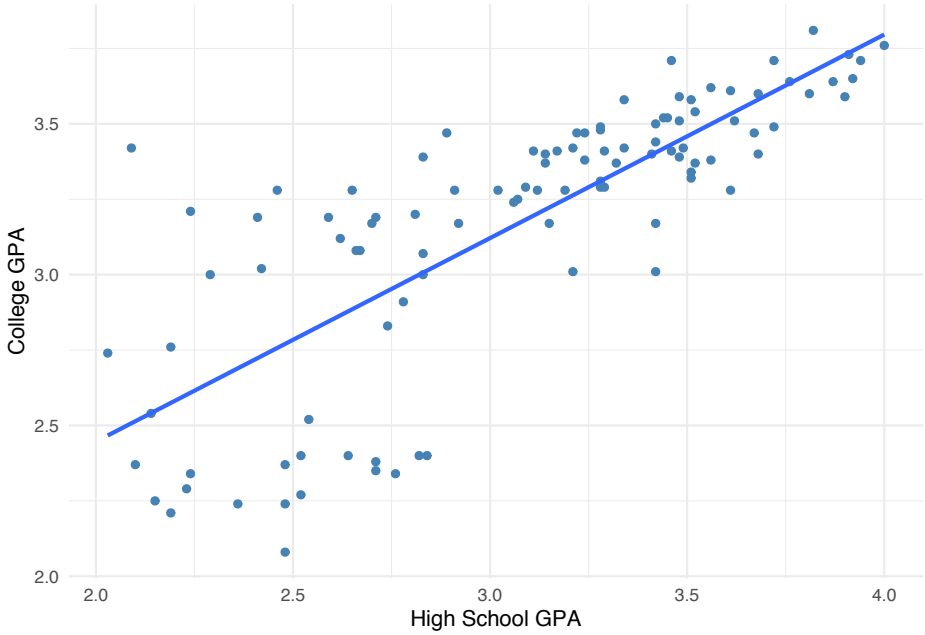
In this section, we consider model evaluation. We seek to develop tools that allow us to answer: is this a "good" model?

3.7.1 Prediction

Let's consider another data example. Can we predict your college GPA based on your high school GPA? (Disclaimer: this is not Macalester data)

```
sat <- read.csv("Data/sat.csv")
```

```
sat %>%
    ggplot(aes(x = high_GPA, y = univ_GPA)) +
    geom_point(color = 'steelblue') +
    geom_smooth(method = 'lm', se = FALSE) +
    xlab('High School GPA') +
    ylab('College GPA') +
    theme_minimal()
```



First things first. Describe the scatterplot.

- **Direction**: Positive relationship (higher high school GPA is associated with higher college GPA)
- **Form**: generally linear
- **Strength**: There is a weak relationship when high school GPA < 3.0 ($r = 0.32$) and a fairly strong relationship when high school GPA > 3.0 ($r = 0.68$).
- **Unusual**: As seen with the strength, there is greater variability in college GPA among individuals with lower high school GPA. That variability decreases with increased high school GPA. We call this pattern of unequal variation as **"thickening"** or **heteroscedasticity** (this terms is used quite a bit in econometrics).

The code below computes the correlation coefficients separately for students with high school GPAs above 3 and for students with high school GPAs less than or equal to 3. We see that the correlation is higher for the high GPA group.

```
sat %>%
    mutate(HighHSGPA = high_GPA > 3) %>%
    group_by(HighHSGPA) %>%
    summarize(Cor = cor(high_GPA,univ_GPA))
```

```
## # A tibble: 2 x 2
##   HighHSGPA   Cor
##   <lgl>     <dbl>
## 1 FALSE     0.316
```

```
## 2 TRUE      0.687
```

Let's build a model to predict college GPA based on high school GPA based on this sample data. Since we noted that there was a linear relationship, let's find the least squares regression line.

```
sat %>%
  lm(univ_GPA ~ high_GPA, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = univ_GPA ~ high_GPA, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6904 -0.1192  0.0327  0.1740  0.9128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0968     0.1666    6.58    2e-09 ***
## high_GPA      0.6748     0.0534   12.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.281 on 103 degrees of freedom
## Multiple R-squared:  0.608,  Adjusted R-squared:  0.604
## F-statistic:  160 on 1 and 103 DF,  p-value: <2e-16
```

The best fitting line is

$$\text{Predicted College GPA} = 1.09 + 0.675 \times \text{High School GPA}$$

Let's plug in a few values.

- If High School GPA $= 2$:

$$\text{Predicted College GPA} = 1.09 + 0.675 \times 2 = 2.44$$

```
## Calculation by hand; rounded before prediction
1.09 + 0.675*2
```

```
## [1] 2.44
```

```
## Calcuation using R's predict() function
sat %>%
  lm(univ_GPA ~ high_GPA, data = .) %>%
  predict(newdata = data.frame(high_GPA = 2))
```

```
##    1
## 2.45
```

- If High School GPA = 3.5:

$$\text{Predicted College GPA} = 1.09 + 0.675 \times 3.5 = 3.45$$

```
1.09 + 0.675*3.5 #rounded before prediction
```

```
## [1] 3.45
```

```
sat %>%
  lm(univ_GPA ~ high_GPA, data = .) %>%
  predict(newdata = data.frame(high_GPA = 3.5)) #rounded after prediction
```

```
##    1
## 3.46
```

- If High School GPA = 4.5:

$$\text{Predicted College GPA} = 1.09 + 0.675 \times 4.5 = 4.13$$

```
1.09 + 0.675*4.5 #rounded before prediction
```

```
## [1] 4.13
```

```
sat %>%
  lm(univ_GPA ~ high_GPA, data = .) %>%
  predict(newdata = data.frame(high_GPA = 4.5)) #rounded after prediction
```

```
##    1
## 4.13
```

Reflect: Does it make sense to use this model for high school GPA's > 4? Some high schools have a max GPA of 5.0 due to the weighting of advanced courses.

- What is the maximum high school GPA in this data set?
- What if your college doesn't allow for GPA's above 4.0?

```
sat %>%
    summarize(max(high_GPA)) # Max high school GPA in this data set
```

```
##   max(high_GPA)
## 1             4
```

Making predictions beyond the observed range of values is called **extrapolation** and is generally a risky thing to do. If you make prediction for values of x beyond the minimum or maximum of the observed values, then you are assuming that the relationship you observe can be extended into the new prediction range.

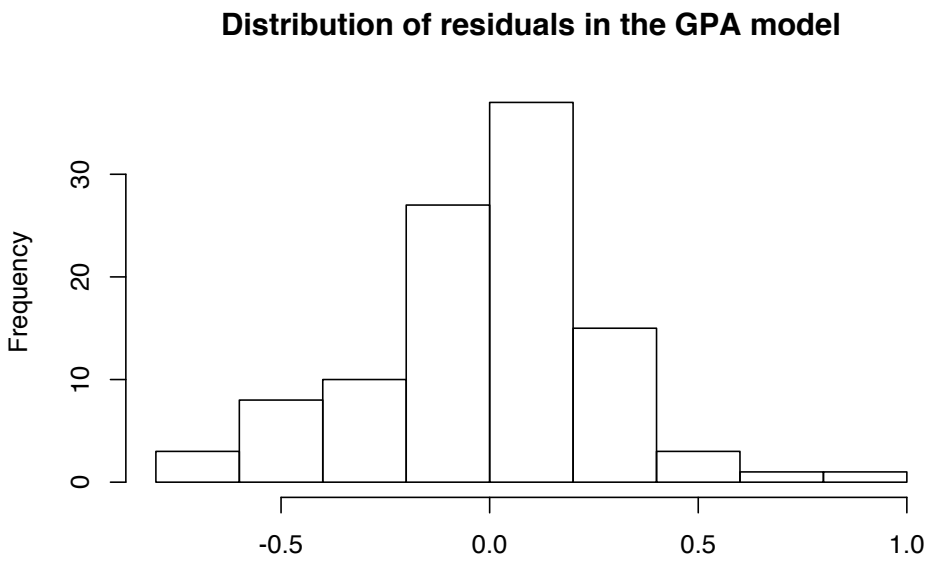
This is the main issue of **forecasting**, making predictions in the future. You have to assume that the trend that you observe now will continue in the future and that the current state of affairs will stay the same. For an infamous case of extrapolation, check out this article that appeared in the journal Nature.

3.7.2 Prediction Errors

Recall that a residual, e_i, for the ith data point is the difference between the actual and predicted values: $e_i = y_i - \hat{y}_i$.

If the residuals were approximately unimodal and symmetric, we expect about 95% of the residuals to be within 2 standard deviations of 0 (the mean residual). (Recall Section 2.6.5.)

```
sat %>%
  lm(univ_GPA ~ high_GPA, data = .) %>%
  residuals() %>%
  hist(main = "Distribution of residuals in the GPA model")
```



.

Below we calculate SSE (the sum of squared errors/residuals), and the standard deviation of the residuals (s_e).

```
lm.gpa <- sat %>%
  lm(univ_GPA ~ high_GPA, data = .)
SSE <- sum(residuals(lm.gpa)^2)
n <- length(residuals(lm.gpa)) # Sample size
s <- sqrt(SSE/(n-2)) #sd of residuals
s
```

```
## [1] 0.281
```

```
2*s
```

```
## [1] 0.563
```

Using this model (that is, using your high school GPA), we can predict your college GPA within about 0.56 GPA points. Is this useful? Is predicting within a margin of 0.56 GPA points good enough? Let's compare this margin of error with the margin of error about the mean (the standard deviation):

```
sd(sat$univ_GPA)
```

```
## [1] 0.447
```

```
2*sd(sat$univ_GPA)
```

```
## [1] 0.894
```

Without knowing your high school GPA, we could have just guessed your college GPA as the overall mean college GPA in the sample, and this guess would probably be within ± 0.89 of your actual college GPA. This is a higher margin of error than the approximate 0.56 if we did use your high school GPA (in the simple linear regression model). We are able to predict your college GPA with a smaller margin of error than if we just guessed your college GPA with the mean. Our model has *explained* some (but not all) of the variation in college GPA.

The standard deviation of college GPA is based on the sum of squared total variation, SSTO (variation around the mean),

$$SSTO = \sum (y_i - \bar{y})^2$$

$SSTO$ is the numerator of the standard deviation of y (without knowing anything about x).

```
(SSTO = sum((sat$univ_GPA - mean(sat$univ_GPA))^2))
```

```
## [1] 20.8
```

We define $SSTO$ here because it will help to compare $SSTO$ to SSE (sum of squared residuals from the model) to obtain a measure of how well our models are **fitting** the data - how well they are predicting the outcome.

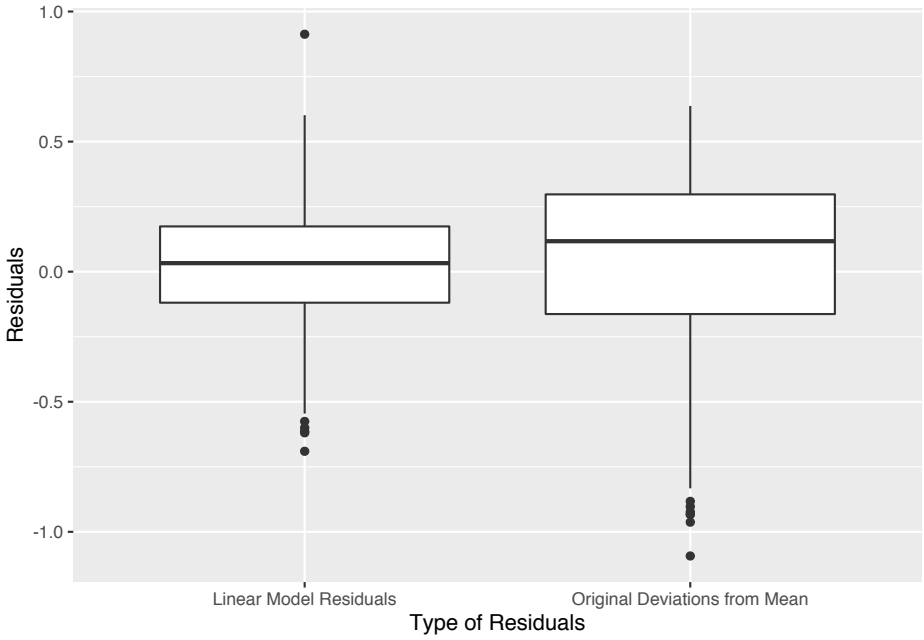
3.7.3 R^2

Let's study how *models reduce unexplained variation.*

- Before a model is fit, the unexplained variation is given by $SSTO$. It is the overall variability in the outcome. Think back to interpreting standard deviation and variance as measures of spread. We used these to describe broadly how much the outcome varys

- Unexplained variation in the outcome that remains after modeling is given by SSE, the sum of squared residuals.

So to study how models reduce unexplained variation, we compare the magnitude of the residuals from a linear regression model (which uses the predictor x) with the original deviations from the mean (which do not use the predictor x).



We started with the sum of the deviations from the mean $SSTO = \sum (y_i - \bar{y})^2$ before we had info about high school GPA (x).

- Now, with our knowledge of x, we have $SSE = \sum (y_i - \hat{y}_i)^2$
- SSE should be smaller than $SSTO$ (!)

Two extreme cases:

- If the error SSE goes to zero, we'd have a "perfect fit".
- If $SSE = SSTO$, x has told us nothing about y.
- So we define a measure called **R-squared**, which is the *fraction* or *proportion* of the total variation in y "accounted for" or "explained" by the model in x.

Math Box:

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

In R, `lm()` will calculate R-Squared (R^2) for us, but we can also see that it equals the value from the formula above.

```
1 - SSE/SSTO
```

```
## [1] 0.608
```

```
glance(lm.gpa) #r.squared = R^2, sigma = s_e (ignore the rest)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.608         0.604 0.281      160. 1.18e-22     2  -14.9  35.7  43.7
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

- Is there a "good" value of R^2? Same answer as correlation – no.
- R^2 doesn't tell you the direction or the form of the relationship.
- Note: $R^2 = r^2$ for simple linear models with one x variable (where r is the correlation coefficient).

3.8 Diagnostics

Residuals are what's left over from a model. We can actually learn a lot by studying what is left over, what is left unexplained by the model. INSERT SCATOLOGY JOKE.

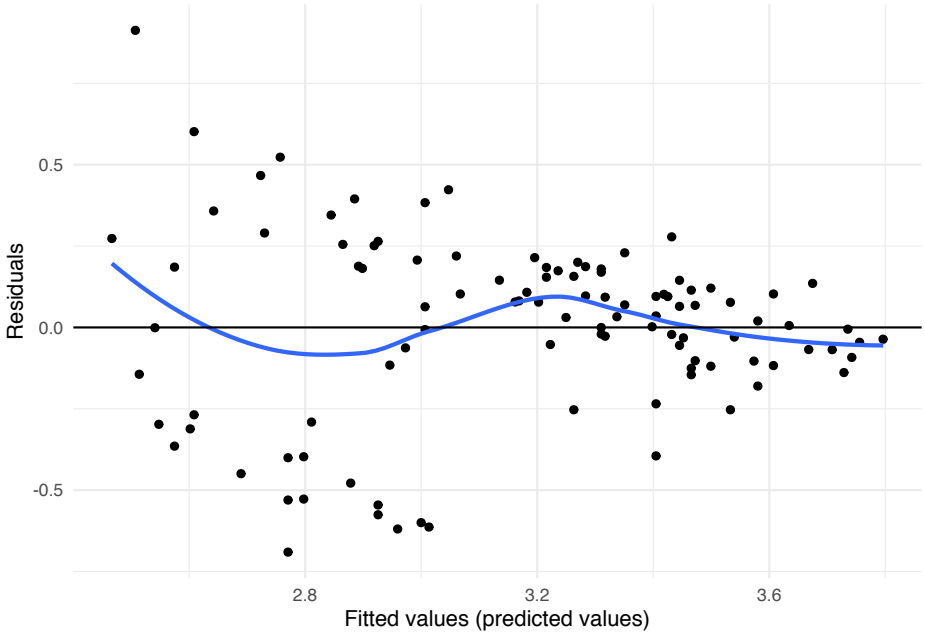
What do we need for a simple linear model to make sense?

- Variables are both **Quantitative**
- Relationship is **Straight Enough**
- There are no extreme **Outliers**
- Spread is roughly same throughout – **No Thickening**

To check these, we look at the original scatterplot as well as a plot of the *residuals* against *fitted or predicted values*,

```
augment(lm.gpa, data = sat) %>%
    ggplot(aes(x = .fitted, y = .resid)) +
    geom_point() +
```

```
    geom_smooth(se = FALSE) +
    geom_hline(yintercept = 0) +
    labs(x = "Fitted values (predicted values)", y = "Residuals") +
    theme_minimal()
```



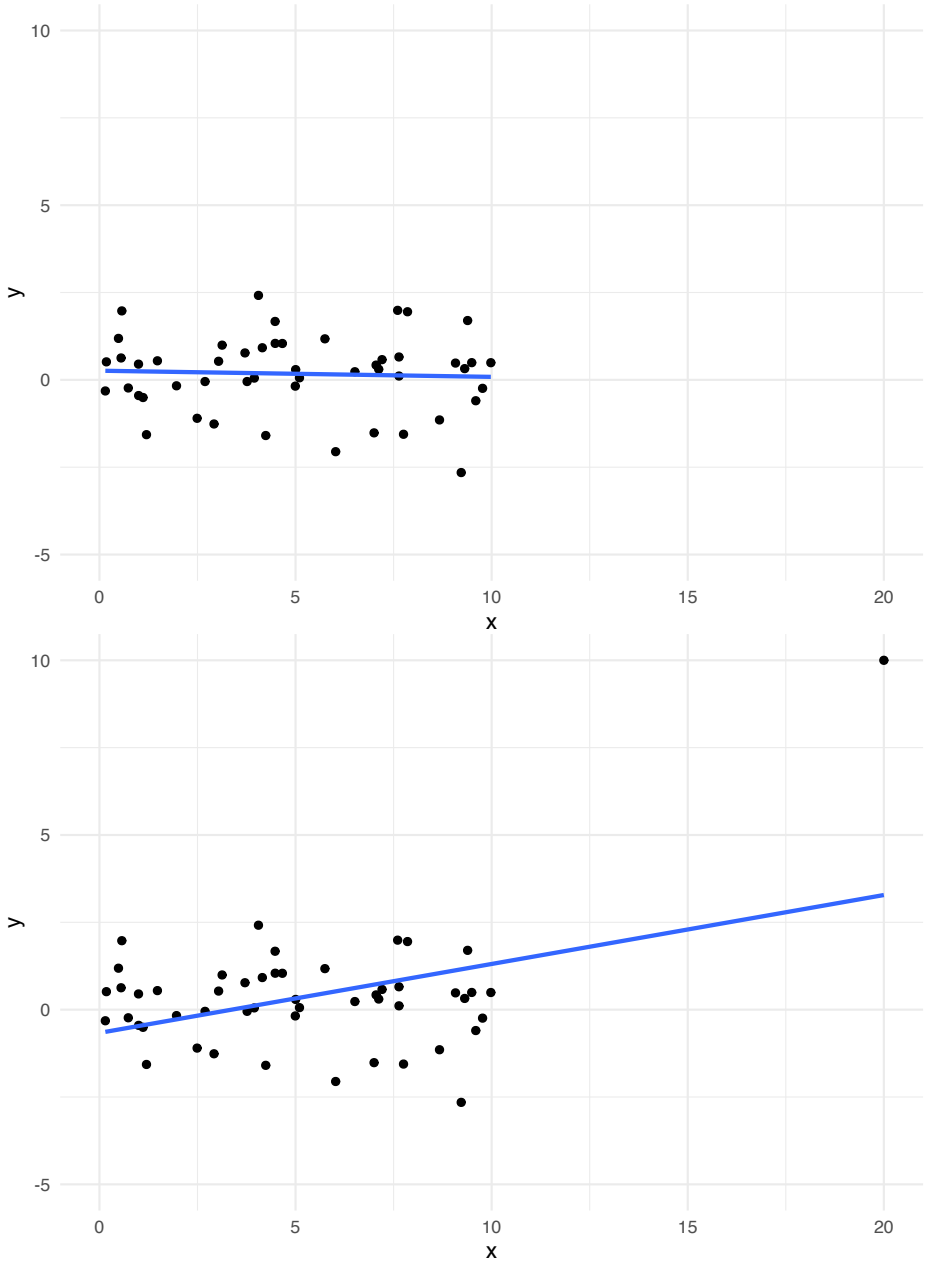
What do you think?

- Is there any pattern in the residuals? (Is the original scatterplot straight enough?)
- Is there equal spread in residuals across prediction values?

Studying the residuals can highlight subtle non-linear patterns and thickening in the original scatterplot. Think of it as a magnifying glass that helps you see these patterns.

Additionally, we want to avoid extreme outliers because points that are both far from the mean of x and do not fit the overall relationship have **leverage** or the ability to change the line.

See the example below. See how the relationship changes with the addition of one point, one extreme outlier.



Check out this interactive visualization to get a feel for outlier points and their potential leverage: http://omaymas.github.io/InfluenceAnalysis/

3.8.1 Solutions to Regression Issues

If the observed data do not satisfy the conditions above, what can we do? Should we give up using a statistical model? No!

- Problem: Both variables are NOT **Quantitative**
  - If your x-variable is categorical, we'll turn it into a quantitative variable using **indicator variables** (coming up)
  - If you have a binary variable (exactly 2 categories) that you want to predict as your y variable, we'll use **logistic regression** (coming up)
- Problem: Relationship is NOT **Straight Enough**
  - If the plot does not thicken, we can add higher degree (e.g. x^2, x^3, x^4, etc.) terms to the model (**multiple linear regression** - coming up).
  - If the plot does thicken, see solutions below.
- Problem: **Spread** is NOT the same throughout
  - You may be able to transform the y-variable using mathematical functions ($log(y)$, y^2, etc.) to make the spread more consistent (one approach is to use **Box-Cox Transformation** – take more statistics classes to learn more)
  - Be careful in interpreting the standard deviation of the residuals; be aware of the units. (For example, if you take a log transformation of the outcome, the units of the standard deviation of the residuals will be on the log scale.)
- Problem: You have **extreme outliers**
  - Look into the outliers. Determine if they could be due to human error or explain their unusual nature using the context. Think carefully about them, dig deep.
  - Do a **sensitivity analysis**: Fit a model with and without the outlier and see if your conclusions drastically change (see if those points had leverage to change the model).

3.9 Multiple Linear Regression

We can generalize the idea of a simple linear model by including many explanatory variables (x's). A **multiple linear regression model** can be written as:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

- Each coefficient b_j can be interpreted as the increase in the predicted/average y associated with a 1 unit increase in x_j, **keeping all other variables constant**. (*There are some exceptions - we'll get there.)

- These explanatory variables can be:
  - Quantitative variables (or transformations of them)
  - Indicator variables for categorical variables (only need $l-1$ indicators for a variable with l categories)
  - Interaction terms (product of two variables, which allows for *effect modification*)

Let's talk about a new data example: home prices. We want to build a model to predict the price of a home based on its many characteristics. Here we have a data set of homes recently sold in New England with many variables such as the age of the home, the land value, whether or not it has central air conditioning, the number of fireplaces, the sale price, and more...

```
homes <- read.delim('Data/Saratoga.txt')
head(homes)
```

```
##    Price Lot.Size Waterfront Age Land.Value New.Construct Central.Air
## 1 132500     0.09          0  42      50000             0           0
## 2 181115     0.92          0   0      22300             0           0
## 3 109000     0.19          0 133       7300             0           0
## 4 155000     0.41          0  13      18700             0           0
## 5  86060     0.11          0   0      15000             1           1
## 6 120000     0.68          0  31      14000             0           0
##   Fuel.Type Heat.Type Sewer.Type Living.Area Pct.College Bedrooms
## 1         3         4          2         906          35        2
## 2         2         3          2        1953          51        3
## 3         2         3          3        1944          51        4
## 4         2         2          2        1944          51        3
## 5         2         2          3         840          51        2
## 6         2         2          2        1152          22        4
##   Fireplaces Bathrooms Rooms
## 1          1       1.0     5
## 2          0       2.5     6
## 3          1       1.0     8
## 4          1       1.5     5
## 5          0       1.0     3
## 6          1       1.0     8
```

*The exception to the interpretation comment above is if some of our x variables are strongly correlated. In this case, we cannot keep all other variables constant because if you increase the value of one, then a variable with high correlation will also likely change in value.

3.9.1 Indicator Variables

In New England, fireplaces are often used as a way to provide supplementary heat to the house. Let's study the impact of a fireplace has on the sale price of

a home. In particular, we only care if the home has 1 or more fireplaces or no fireplaces. So we make a new variable that is TRUE if there are more than 0 fireplaces in a home and FALSE otherwise.

```
homes <- homes %>%
  mutate(AnyFireplace = Fireplaces > 0)
```

In order to include this information in our linear regression model, we need to turn that categorical variable (`AnyFireplace` with values of TRUE or FALSE) into an **indicator variable**, which has a numeric value of 0 or 1:

$$1_{AnyFireplaceTRUE} = \begin{cases} 1 & \text{if a home has at least one fireplace} \\ 0 & \text{if a home does not have a fireplace} \end{cases}$$

In fact, R creates this indicator for you when you include a categorical variable as an x variable the `lm()` function.

```
lm.home <- lm(Price ~ AnyFireplace, data = homes)
summary(lm.home)
```

```
##
## Call:
## lm(formula = Price ~ AnyFireplace, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -234914  -59653  -18784   42145  585347
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        174653       3419    51.1   <2e-16 ***
## AnyFireplaceTRUE    65261       4522    14.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93000 on 1726 degrees of freedom
## Multiple R-squared:  0.108,  Adjusted R-squared:  0.107
## F-statistic:  208 on 1 and 1726 DF,  p-value: <2e-16
```

Our "best fitting line" is

$$\text{Predicted Price} = 174653.35 + 65260.61 \times 1_{AnyFireplaceTRUE}$$

What does this mean?

Let's think about two types of homes: a home with one or more fireplaces and a home without a fireplace. Let's write out the equations for those two types of

homes.

- Home with fireplace (indicator = 1):

$$\text{Predicted Price} = 174653.35 + 65260.61 \times 1 = \$239,914$$

```
174653.35 + 65260.61*1
```

```
## [1] 239914
```

- Home without fireplace (indicator = 0):

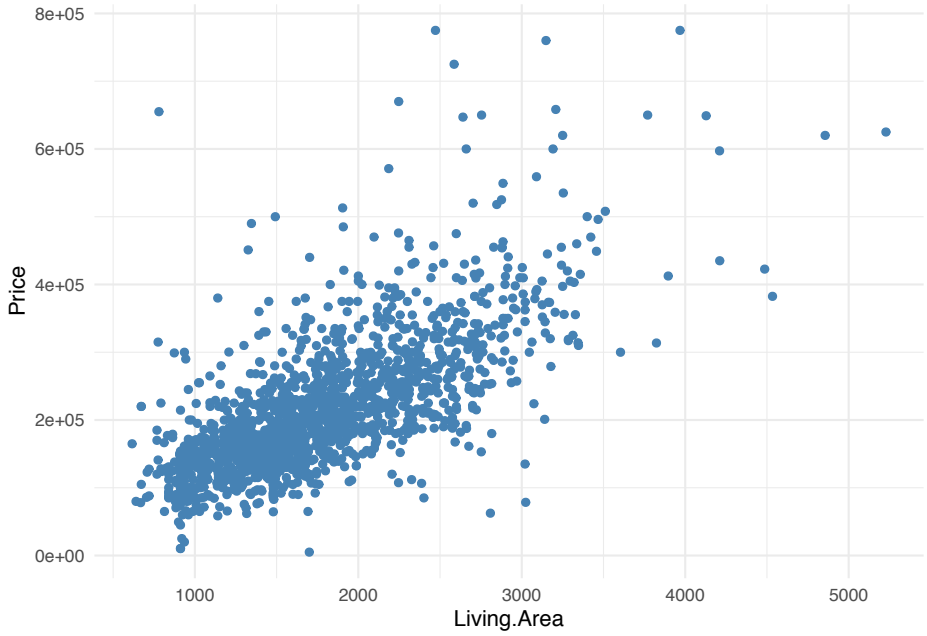
$$\text{Predicted Price} = 174653.35 + 65260.61 \times 0 = \$174,653.35$$

The difference between these predicted prices is $65,260.61, the values of the "slope" for the indicator variable.

> **Reflect:** So is this how much a fireplace is worth? If I installed a fireplace in my house, should the value of my house go up $65,260?
> **No**, because we should not make causal statements based on observational data without thinking deeply about the context. What could be confounding this relationship? What third variable may be related to both the price and whether or not a house has a fireplace?

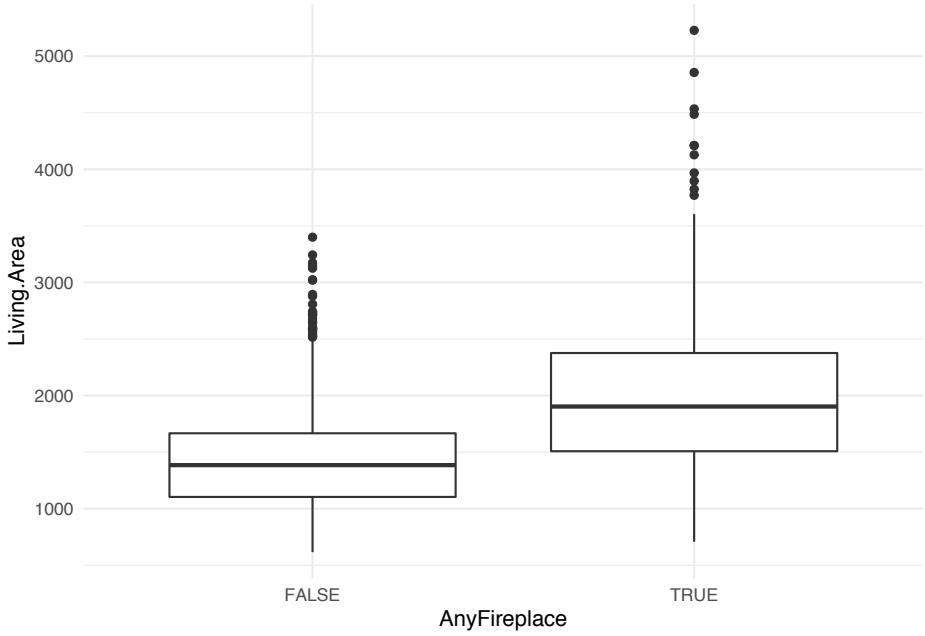
Let's consider the size of the house. Is price related to the area of living space (square footage)?

```
homes %>%
    ggplot(aes(x = Living.Area, y = Price)) +
    geom_point(color = 'steelblue') +
    theme_minimal()
```



Is the presence of a fireplace related to area of living space?

```
homes %>%
    ggplot(aes(x = AnyFireplace, y = Living.Area)) +
    geom_boxplot() +
    theme_minimal()
```



We see that the amount of living area differs between homes with fireplaces and homes without fireplaces. Thus, Living Area could confound the relationship between AnyFireplace and Price because it is related to both variables.

Let's put Living Area in the model along with AnyFireplace to account for it (to control/adjust for it).

```
lm.home2 <- lm(Price ~ AnyFireplace + Living.Area, data = homes)
summary(lm.home2)
```

```
##
## Call:
## lm(formula = Price ~ AnyFireplace + Living.Area, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -271421  -39935   -7887   28215  554651
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13599.16    4991.70    2.72   0.0065 **
## AnyFireplaceTRUE  5567.38    3716.95    1.50   0.1344
## Living.Area        111.22       2.97   37.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 69100 on 1725 degrees of freedom
## Multiple R-squared:  0.508,  Adjusted R-squared:  0.508
## F-statistic:  891 on 2 and 1725 DF,  p-value: <2e-16
```

Our "best fitting line" is

$$\text{Predicted Price} = 13599.16 + 5567.37 \times 1_{AnyFireplaceTRUE} + 111.21 \times \text{Living.Area}$$

Note that the slope for the indicator variable is very different. This suggests that Living Area was confounding the relationship between Price and AnyFireplace.

What does this mean?

Let's think about two types of homes: a home with one or more fireplaces and a home without a fireplace.

- Home with fireplace (indicator = 1):

$$\begin{aligned}\text{Predicted Price} &= 13599.16 + 5567.37 \times 1 + 111.21 \times \text{Living.Area} \\ &= \$19,166.53 + \$111.21 \times \text{Living.Area}\end{aligned}$$

Among homes with a fireplace, we have one linear relationship between living area and price.

- Home without fireplace (indicator = 0):

$$\begin{aligned}\text{Predicted Price} &= 13599.16 + 5567.37 \times 0 + 111.21 \times \text{Living.Area} \\ &= \$13,599.16 + \$111.21 \times \text{Living.Area}\end{aligned}$$

Among homes without a fireplace, we have a different linear relationship between living area and price.

- For either type of home, \$111.21 is the increase in the predicted or average Price associated with a 1 square footage increase in Living.Area, **keeping the number of fireplaces constant**.

Now let's compare homes that are the same size.

If we keep Living.Area constant by considering two equally sized homes, then we'd expect the home with the fireplace to be worth \$5567.37 more than a home without a fireplace.

We see this by taking the difference between the two equations:

$$\text{Predicted Price (with Fireplace)} - \text{Predicted Price (without Fireplace)}$$

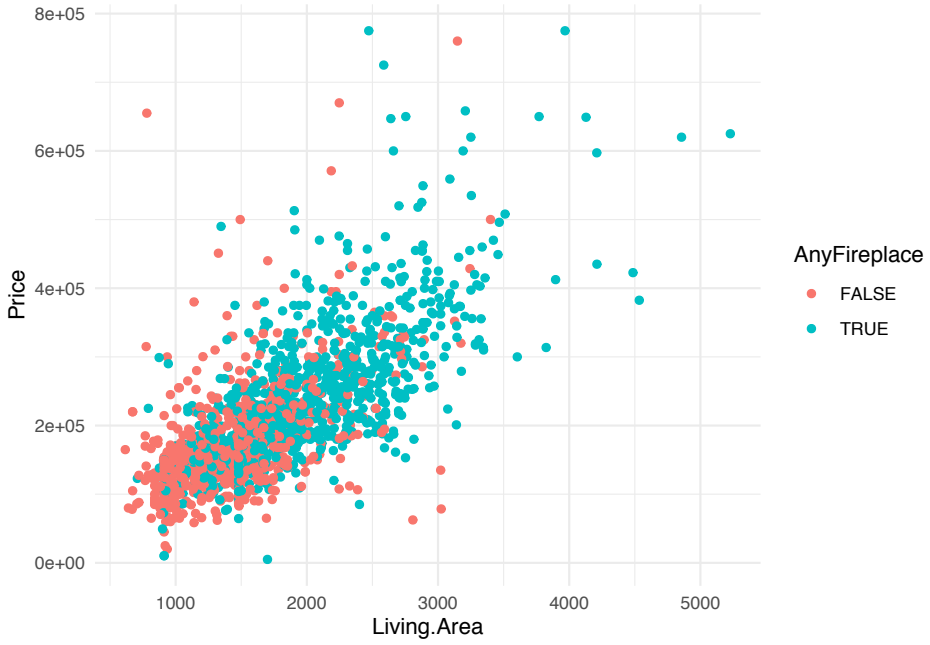
$$= (\$19,166.53 + \$111.21 \times \text{Living.Area}) - (\$13599.16 + \$111.21 \times \text{Living.Area}) \quad = \$19,166.53 - \$13,5$$

The difference between the intercepts is 5567.37.

- This was the estimated coefficient or slope for AnyFireplaceTRUE.
- So the $5567.37 is the increase in the predicted or average Price associated with a 1 unit increase in AnyFireplace (going from FALSE to TRUE), **keeping all other variables (Living.Area) constant**.

Let's look back at the relationship between Living.Area and Price and color the scatterplot by AnyFireplace. So we are now looking at three variables at a time. The above model with AnyFireplace and Living.Area results in two lines for Living.Area v. Price, with different intercepts but the same slope (parallel lines).

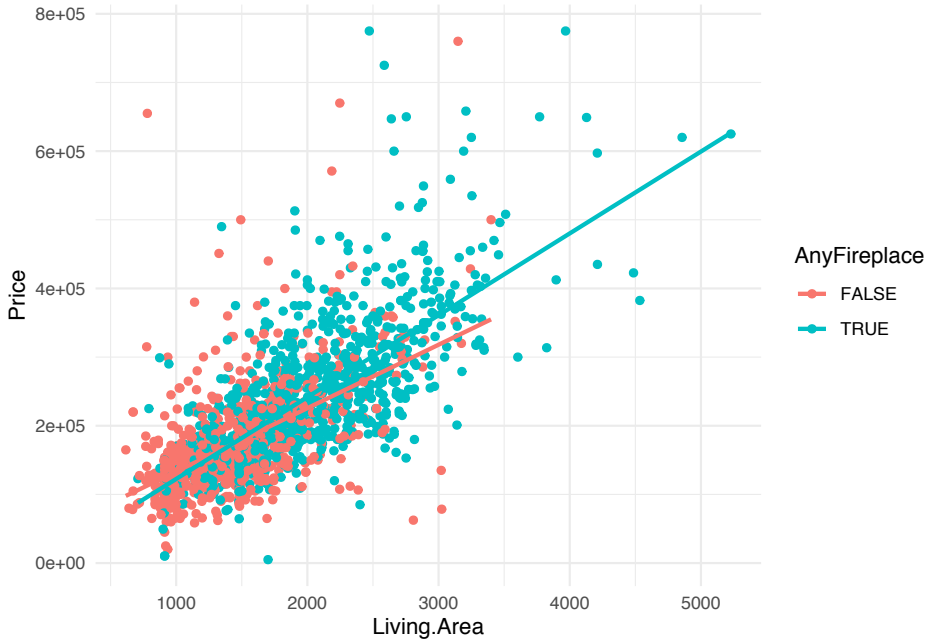
```
homes %>%
    ggplot(aes(x = Living.Area, y = Price, color = AnyFireplace)) +
    geom_point() +
    theme_minimal()
```



Let's try and fit two separate lines to these two groups of homes, home with any fireplaces and home with no fireplaces. Do these lines have the same intercepts? Same slopes?

```
homes %>%
    ggplot(aes(x = Living.Area, y = Price, color = AnyFireplace)) +
```

```
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE) +
    theme_minimal()
```



In this case, it look as though having a fireplace in your house slightly changes the relationship between Living.Area and Price. In fact, having a fireplace in your house, the increase in your price for every 1 square foot is greater than that for homes without fireplaces (slopes are different).

3.9.2 Interaction Variables

We can allow for different slopes within one regression model (!), rather than fitting two separate models.

> **Reflect:** When should we fit only one model; when should we fit separate models?
> If we fit separate models, we are **stratifying** and then modeling. But what if some of the strata are small?
> Fitting one model allows us to "borrow information across groups."
> There is no one right answer. Researchers struggle with this decision to stratify or not to stratify.

- If we add a variable in the model (without an interaction), it only changes

the intercept.

- We can achieve different slopes by allowing a variable x_1 to affect the slope for another variable x_2. That is, x_1 impacts the effect of x_2 on the outcome y. (Fireplace presence impacts the effect of living area on house price.)

  $$b_2 = a + bx_1$$

  This is called **effect modification** (when one variable can modify the effect of another variable on the outcome).

- A model with effect modification looks like this:

  $$\hat{y} = b_0 + b_1x_1 + b_2x_2 = b_0 + b_1x_1 + (a + bx_1)x_2 = b_0 + b_1x_1 + ax_2 + bx_1x_2$$

  The model above has an **interaction term,** which is the product of two variables. Here we have $x_1 * x_2$.

Let's build a model with effect modification for our housing data. Let's include an interaction term between AnyFireplace and Living.Area to allow for different slopes.

```
lm.home3 <- lm(Price ~ AnyFireplace*Living.Area, data = homes)
summary(lm.home3)
```

```
##
## Call:
## lm(formula = Price ~ AnyFireplace * Living.Area, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241710  -39588   -7821   28480  542055
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    40901.29    8234.66    4.97  7.5e-07 ***
## AnyFireplaceTRUE              -37610.41   11024.85   -3.41  0.00066 ***
## Living.Area                      92.36       5.41   17.07  < 2e-16 ***
## AnyFireplaceTRUE:Living.Area     26.85       6.46    4.16  3.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68800 on 1724 degrees of freedom
## Multiple R-squared:  0.513,  Adjusted R-squared:  0.512
## F-statistic:  605 on 3 and 1724 DF,  p-value: <2e-16
```

What does this mean?

Let's think about two types of homes: a home with one or more fireplaces and a home without a fireplace.

- Home with fireplace (indicator = 1):

$$\begin{aligned}\text{Predicted Price} &= 40901.29 + -37610.41 \times 1 + 92.36391 \times \text{Living.Area} + 26.85 \times \text{Living.Area} \times 1 \\ &= \$3,290.88 + \$119.21 \times \text{Living.Area}\end{aligned}$$

```
40901.29 + -37610.41*1
```

```
## [1] 3291
```

```
92.36391 + 26.85*1
```

```
## [1] 119
```

- Home without fireplace (indicator = 0):

$$\begin{aligned}\text{Predicted Price} &= 40901.29 + -37610.41 \times 0 + 92.36391 \times \text{Living.Area} + 26.85 \times \text{Living.Area} \times 0 \\ &= \$40,901.29 + \$92.36 \times \text{Living.Area}\end{aligned}$$

We note different slopes and different intercepts for these two groups.

3.9.3 Causation

We alluded earlier that multiple linear regression could provide estimates of causal effects in the right circumstances. What are those circumstances? When we include **all** confounding variables.

Remember, a confounder is a **common cause** of both the causal variable of interest and the outcome (e.g. living area could be a confounder of fireplace presence and house price).

We also alluded earlier that we should not just throw every variable we have into a multiple regression model. Why? Imagine a scenario for understanding how smoking affects lung cancer development. It is very important to consider whether a variable is a **mediator** of the relationship between the cause and the outcome. A mediator is a variable in a chain or along the path (Section 1.6) between the cause and the outcome. In the smoking and lung cancer example, one mediator could be tar in the lungs. Suppose that smoking only affects lung cancer risk by creating tar on the lungs. If we adjust for tar (by holding it constant), then we also effectively hold smoking constant too! If smoking is held constant, then we cannot estimate its effect on cancer risk because it is not varying!

Wait - we could never possibly know of or measure all confounding variables, could we!? This is true, but that doesn't mean that our endeavor to understand causation is fruitless. As long as we can describe the relationship between known

confounders as precisely as possible with writing down a DAG (causal model), we have a starting ground for moving forward.

We collect data, analyze how well our model predicts that data, and collect more data based on that, perhaps measuring more potential confounders as our scientific knowledge grows. We can also conduct sensitivity analyses by asking: how strongly must a confounder affect the variable of causal interest and the outcome to completely negate or reverse the association we see? Such endeavors and more are the subject of the field of *causal inference*.

Reminder: If you want a "gentle" but mathematical introduction to Causal Inference, we suggest taking the class at Macalester and/or reading "Causal Inference in Statistics: A Primer" by Judea Pearl, Madelyn Glymour, Nicholas P. Jewell). Fun Fact: Nicholas Jewell was Prof. Heggeseth's PhD advisor!

3.9.4 Conditions for Multiple Linear Regression

In order for a multiple linear regression model to make sense,

1. Relationships between each pair of quantitative x's and y are **straight enough** (check scatterplots and residual plot)
2. There is about **equal spread** of residuals across fitted values (check residual plot)
3. There are no extreme outliers (points far away in x's can have **leverage** to change the line)

3.9.5 Is the Difference Real?

We could ask: is there *really* a difference in the slopes for Living Area and Price between homes with and without a fireplace?

```
lm.home4 <- lm(Price ~ Living.Area*AnyFireplace, data = homes)
summary(lm.home4)
```

```
##
## Call:
## lm(formula = Price ~ Living.Area * AnyFireplace, data = homes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241710  -39588   -7821   28480  542055
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   40901.29    8234.66    4.97  7.5e-07 ***
## Living.Area                      92.36       5.41   17.07  < 2e-16 ***
## AnyFireplaceTRUE             -37610.41   11024.85   -3.41  0.00066 ***
## Living.Area:AnyFireplaceTRUE     26.85       6.46    4.16  3.4e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68800 on 1724 degrees of freedom
## Multiple R-squared:  0.513,  Adjusted R-squared:  0.512
## F-statistic:  605 on 3 and 1724 DF,  p-value: <2e-16
```

If we ask ourselves this question, we are assuming a few things:

1. We would like to make a general statement about a **target population of interest**.

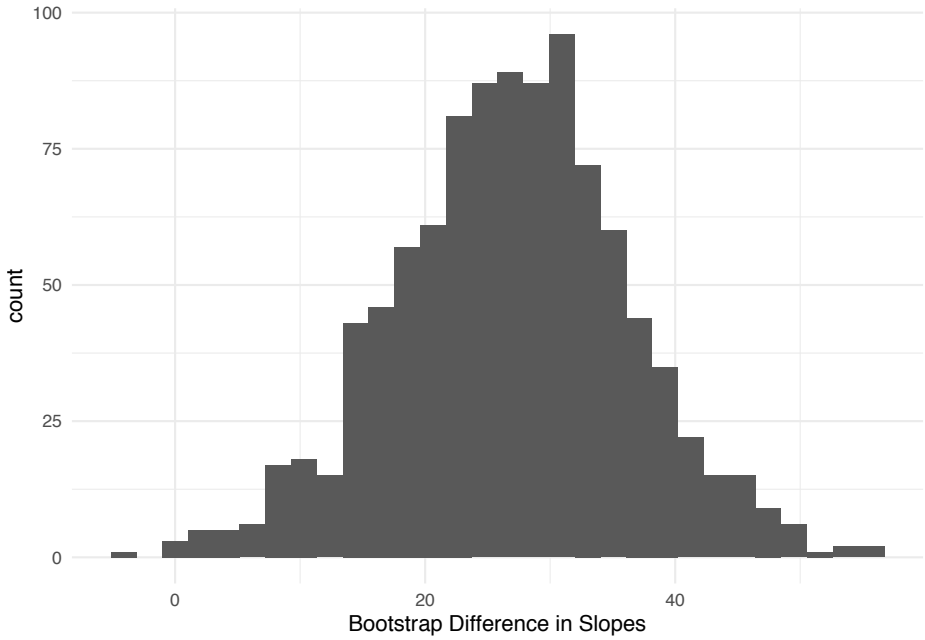
2. We don't have data for everyone in our population (we don't have a **census**).

3. Depending on who **randomly** ends up in our **sample**, the relationship may change a bit.

4. We want to know how much the relationship may change based on **sampling variation.**

- Let's treat our sample (of size n) as a 'fake' population (we don't have the full population but if the sample is representative, then it will be a good proxy).
  - Randomly resample from our sample (with replacement) a new sample of size n
- Estimate the least squares regression line.
- Repeat.

```
set.seed(333) ## Setting the seed ensures that our results are reproducible
## Repeat the sampling and regression modeling 1000 times
boot <- do(1000)*lm(Price ~ Living.Area*AnyFireplace, data = resample(homes))

## Plot the distribution of the 1000 slope differences
boot %>%
    ggplot(aes(x = Living.Area.AnyFireplaceTRUE)) +
    geom_histogram() +
    xlab('Bootstrap Difference in Slopes') +
    theme_minimal()
```



We see that if we were to have a slightly different sample (drawn from our "fake" population), then the difference in the slope could be as long as 0 and as large as 50.

This process of resampling from the sample is called **Bootstrapping** and it is used to:

1. Measure the variability in the estimate (here we are interested in the difference in slopes) between random samples and
2. Provide an interval of plausible values for the estimate (the difference in slopes here).

Let's first look at the variability of the difference in slopes across the bootstrap samples. The standard deviation of the slopes will be similar to the `Std. Error` from the linear model output.

```
boot %>%
  summarize(sd(Living.Area.AnyFireplaceTRUE)) #this is going to be of similar magnitud
```

```
##   sd(Living.Area.AnyFireplaceTRUE)
## 1                             9.33
```

```
summary(lm.home4)
```

```
##
## Call:
## lm(formula = Price ~ Living.Area * AnyFireplace, data = homes)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241710  -39588   -7821   28480  542055
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   40901.29    8234.66    4.97  7.5e-07 ***
## Living.Area                      92.36       5.41   17.07  < 2e-16 ***
## AnyFireplaceTRUE             -37610.41   11024.85   -3.41  0.00066 ***
## Living.Area:AnyFireplaceTRUE     26.85       6.46    4.16  3.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68800 on 1724 degrees of freedom
## Multiple R-squared:  0.513,  Adjusted R-squared:  0.512
## F-statistic:  605 on 3 and 1724 DF,  p-value: <2e-16
```

This standard deviation is somewhat close to the 6.459 in the Std. Error column of the `summary(lm.home4)` output above.

To get an interval of plausible values for the difference in the slopes, we look at the histogram and take the middle 95%. The lower end will be the 2.5th percentile and the upper end will be the 97.5th percentile.

```
boot %>%
  summarize(lower = quantile(Living.Area.AnyFireplaceTRUE, 0.025), upper = quantile(Living.Area.A
```

```
##   lower upper
## 1  7.69  45.4
```

> **Reflect:** Based on this evidence, do you think it is possible that the slopes are the same for the two types of homes (with and without fireplaces)? How would you justify your answer?

3.9.6 Dealing with Non-Linear Relationships

If we notice a curved relationship between two quantitative variables, it doesn't make sense to use a straight line to approximate the relationship.

What can we do?

3.9.6.1 Transform Variables

One solution to deal with non-linear relationships is to transform the explanatory (x) variables or transform the outcome variable (y).

Guideline #1: If there is unequal spread around the curved relationship, focus

first on transforming y. If the spread is roughly the same around the curved relationship, focus on transforming x.

When we say transform a variable, we are referring to taking the values of a variable and plugging them into a mathematical function such as $\sqrt{x}$, $\log(x)$ (which represents natural log, not log base 10), x^2, $1/x$, etc.

Guideline #2: To choose the mathematical function, we focus on power functions and organize them in a **Ladder of Powers** of y (or x):

$$\vdots \tag{3.1}$$
$$y^3 \tag{3.2}$$
$$y^2 \tag{3.3}$$
$$\mathbf{y = y^1} \tag{3.4}$$
$$\sqrt{y} \tag{3.5}$$
$$y^{1/3} \tag{3.6}$$
$$y^0 \quad (we\ use\ \log(y)\ here; naturallog) \tag{3.7}$$
$$y^{-1/3} \tag{3.8}$$
$$1/\sqrt{y} \tag{3.9}$$
$$1/y \tag{3.10}$$
$$1/y^2 \tag{3.11}$$
$$\vdots \tag{3.12}$$

To choose a transformation, we start at y (power = 1) and think about going up or down the ladder.

But which way?

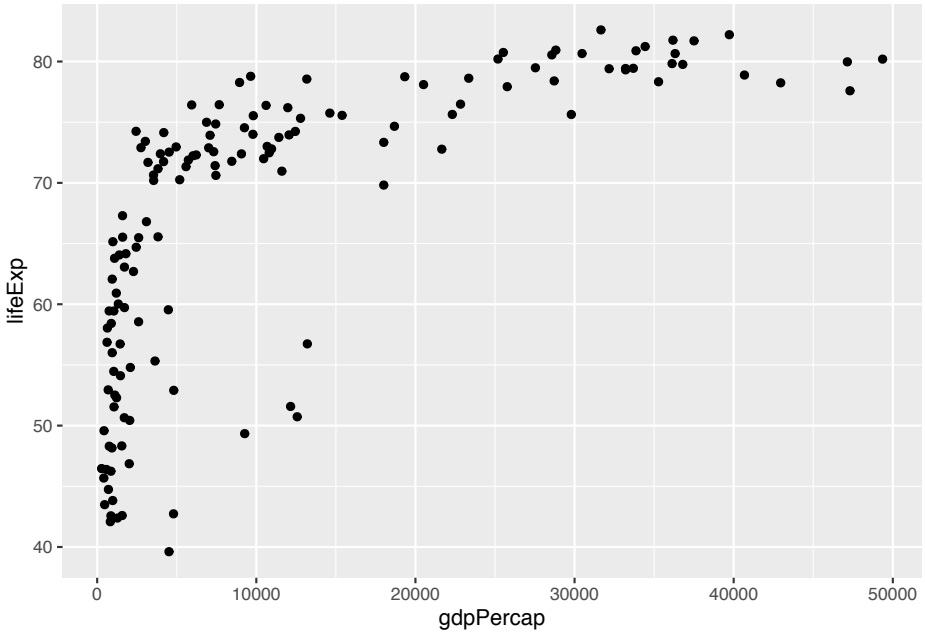
Our friend J.W. Tukey (the same guy who invented the boxplot) came up with an approach to help us decide. You must ask yourself: Which part of the circle does the scatterplot most resemble (in terms of concavity and direction)? Which quadrant?

Guideline #3: The sign of x and y in the quadrant tells you the direction to move on the ladder (positive = up, negative = down).

Practice: Which quadrant does this relationship below resemble?

```
require(gapminder)

gapminder %>%
  filter(year > 2005) %>%
```

```
  ggplot(aes(y = lifeExp, x =  gdpPercap)) +
  geom_point()
```

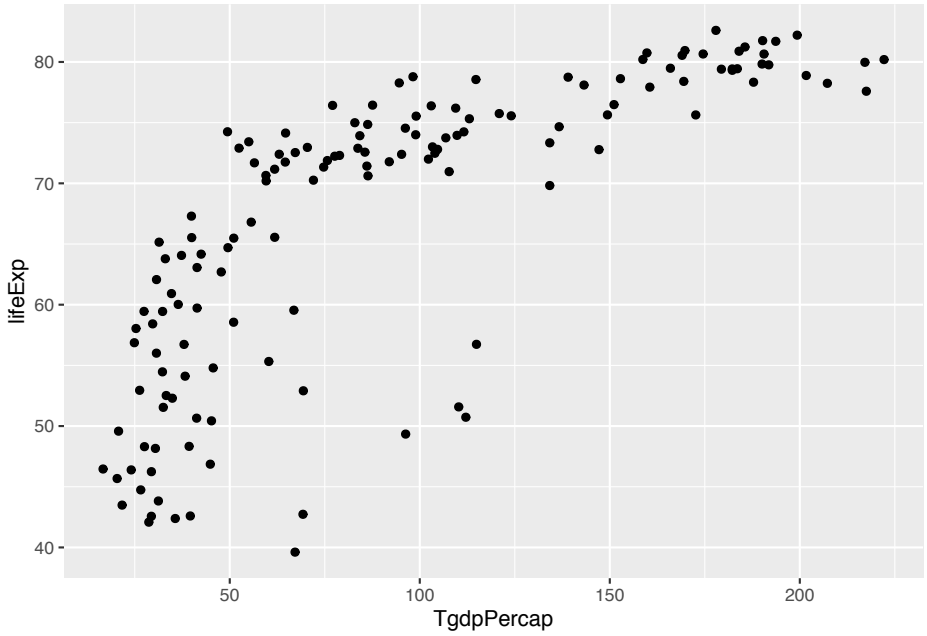


Based on this plot, we see that the spread is roughly equal around the curved relationship (-> focus on transforming x) and that it is concave down and positive (quadrant 2: top left). This suggests that we focus on going down the ladder with x.

Try these transformations until you find a relationship that is roughly straight. If you go too far, the relationship will become more curved again.

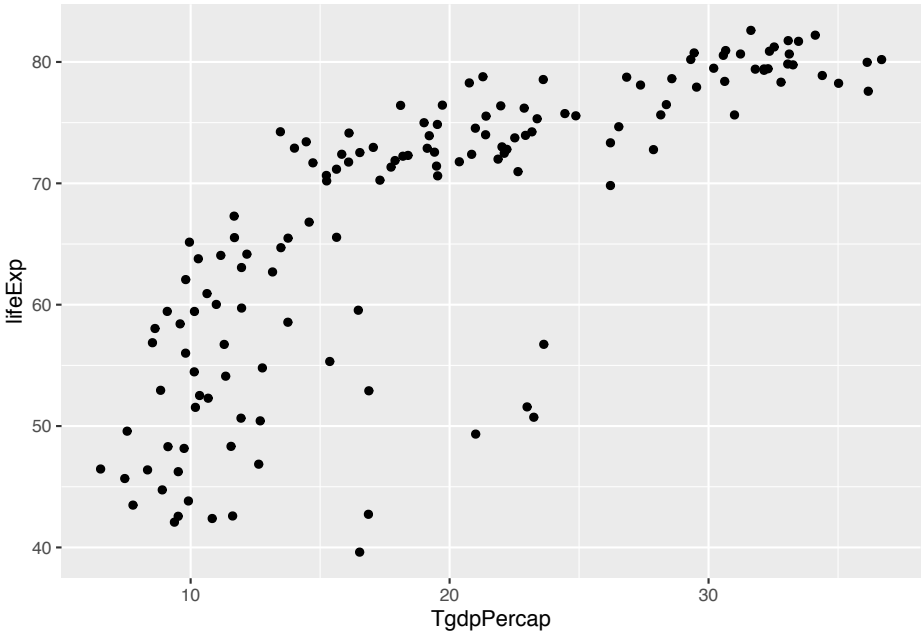
Let's try going down the ladder.

```
gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = sqrt(gdpPercap)) %>% #power = 1/2
  ggplot(aes(y = lifeExp, x =  TgdpPercap)) +
  geom_point()
```



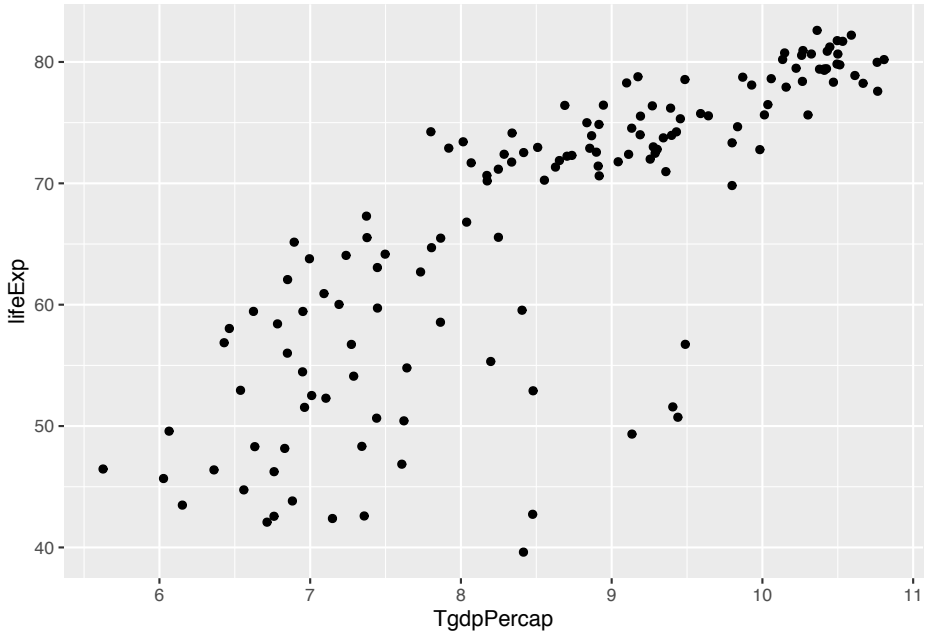
Not quite straight. Let's keep going.

```
gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = gdpPercap^(1/3)) %>% #power = 1/3
  ggplot(aes(y = lifeExp, x =  TgdpPercap)) +
  geom_point()
```



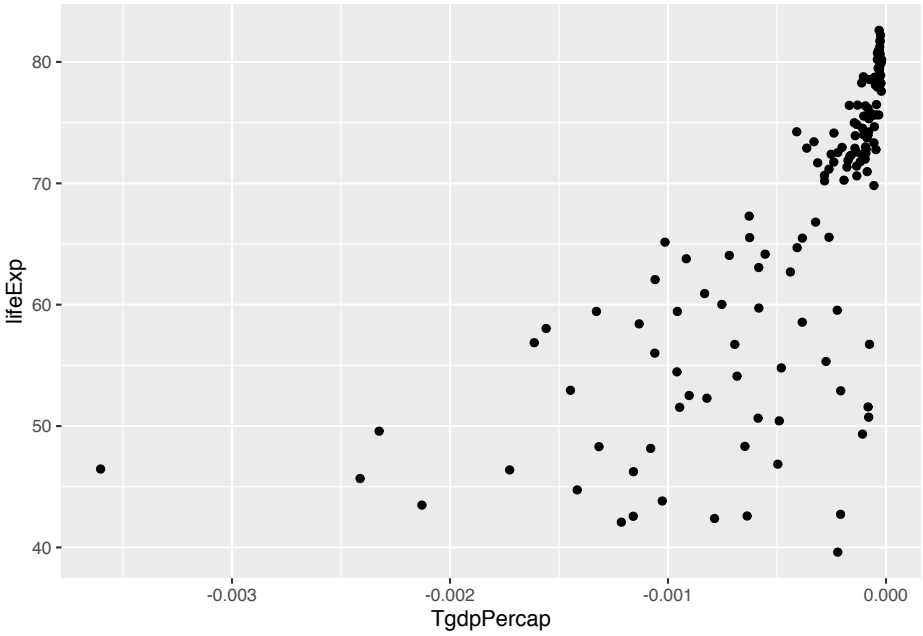
Not quite straight. Let's keep going.

```
gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = log(gdpPercap)) %>% #power = 0
  ggplot(aes(y = lifeExp, x =  TgdpPercap)) +
  geom_point()
```



Getting better. Let's try to keep going.

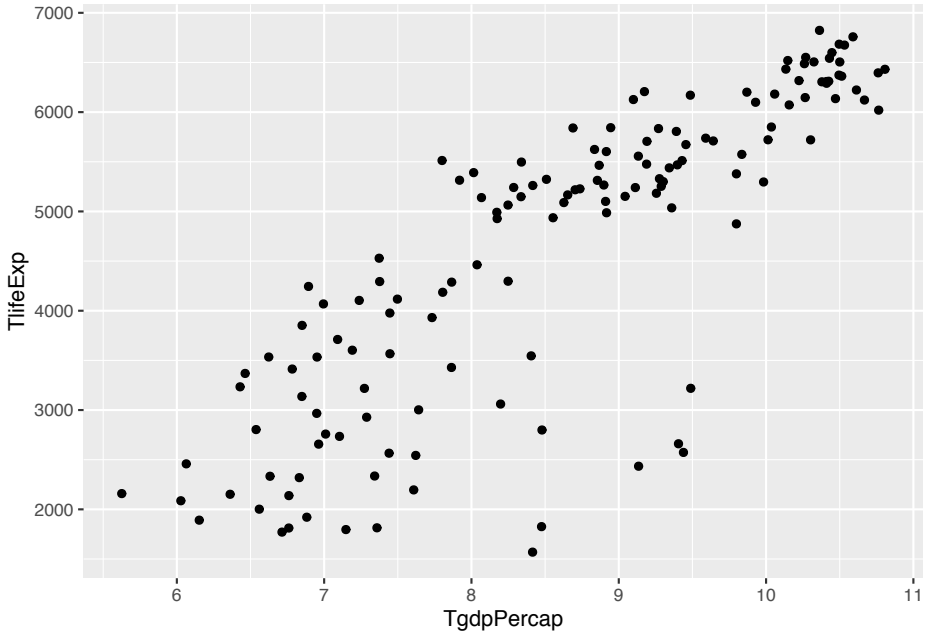
```
gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = -1/gdpPercap) %>%  #power = -1 (added a negative sign to maintai
  ggplot(aes(y = lifeExp, x =  TgdpPercap)) +
  geom_point()
```



TOO FAR! Back up. Let's stick with log(gdpPercap).

Now we see some unequal spread so let's also try transforming y.

```
gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = log(gdpPercap)) %>%
  mutate(TlifeExp = lifeExp^2) %>%
  ggplot(aes(y = TlifeExp, x =  TgdpPercap)) +
  geom_point()
```



That doesn't change it much. Maybe this is as good as we are going to get.

Transformations can't make relationships look exactly linear with equal spread, but sometimes we can make it closer to that ideal.

Let's try and fit a model with these two variables.

```
lm.gap <- gapminder %>%
  filter(year > 2005) %>%
  mutate(TgdpPercap = log(gdpPercap)) %>%
  lm(lifeExp ~ TgdpPercap, data = .)

summary(lm.gap)
```

```
##
## Call:
## lm(formula = lifeExp ~ TgdpPercap, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25.95  -2.66   1.22   4.47  13.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.950      3.858    1.28      0.2
## TgdpPercap     7.203      0.442   16.28   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.12 on 140 degrees of freedom
## Multiple R-squared:  0.654,  Adjusted R-squared:  0.652
## F-statistic:  265 on 1 and 140 DF,  p-value: <2e-16
```

Interpretations

What does the slope of this model, $b_1 = 7.2028$, mean in this data context?

$$\widehat{LifeExp} = b_0 + b_1 log(Income)$$

- The slope is the the additive increase in the predicted $\widehat{LifeExp}$ when $log(Income)$ increases to $log(Income) + 1$.
- Let's think about $log(Income) + 1$. Using some rules of logarithms:

$$log(Income)+1 = log(Income)+log(e^1) = log(e*Income) = log(2.71*Income)$$

So adding 1 to $log(Income)$ is equivalent to multiplying Income by 2.71.

In our model, we note that if a country's income measured by GDP increased by a multiplicative factor of 2.71, the predicted average life expectancy of a country increases by about 7.2 years.

For the sake of illustration, imagine we fit a model where we had transformed life expectancy with a log transformation. What would the slope, b_1, mean in this context?

$$\widehat{log(LifeExp)} = b_0 + b_1 Income$$

- The slope is the the additive increase in $\widehat{log(LifeExp)}$ when $Income$ increases to $Income + 1$.
- Let's think about $\widehat{log(LifeExp)} + b_1$. Using rules of logarithms:

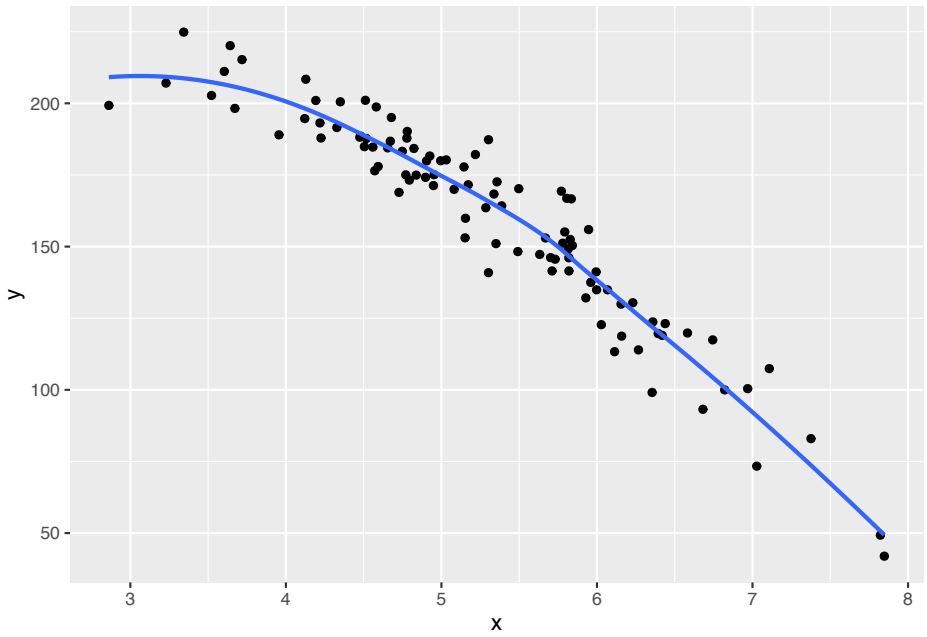
$$\widehat{log(LifeExp)} + b_1 = \widehat{log(LifeExp)} + log(e^{b_1}) = log(\widehat{LifeExp} * e^{b_1})$$

The additive increase of b_1 units in $\widehat{log(LifeExp)}$ is a multiplicative increase of $\widehat{LifeExp}$ by a factor of e^{b_1}.

3.9.6.2 Alternative Solutions

We could also model non-linear relationships by including higher degree terms in a linear model like the example below. By using poly(), we now include x and x^2 as variables in the model.

```
x <- rnorm(100, 5, 1)
y <- 200 + 20*x - 5*x^2 + rnorm(100,sd = 10)
dat <- data.frame(x,y)
dat %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(se = FALSE)
```



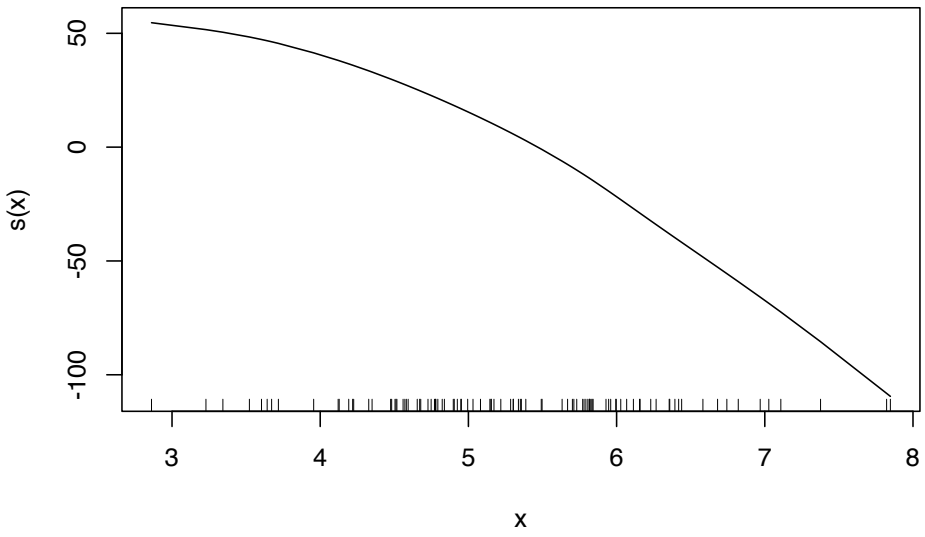

```
lm(y ~ poly(x, degree = 2, raw = TRUE), data = dat)
```

```
##
## Call:
## lm(formula = y ~ poly(x, degree = 2, raw = TRUE), data = dat)
##
## Coefficients:
##                       (Intercept)  poly(x, degree = 2, raw = TRUE)1
##                            192.81                             23.30
## poly(x, degree = 2, raw = TRUE)2
##                             -5.38
```

A more advanced solution (which is not going to be covered in class) is a **generalized additive model** (GAM), which allows you to specify which variables have non-linear relationships with y and estimates that relationship for you using spline functions (super cool stuff!). We won't talk about how this model is fit or how to interpret the output, but there are other cool solutions out there!

```
require(gam)
plot(gam(y ~ s(x), data = dat))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```



3.10 Logistic Regression

If you want to predict a binary categorical variable (only 2 possible outcomes), the standard linear regression models don't apply. If you let the two possible outcomes be 0 and 1, you'll never get a straight line relationship with any x variable.

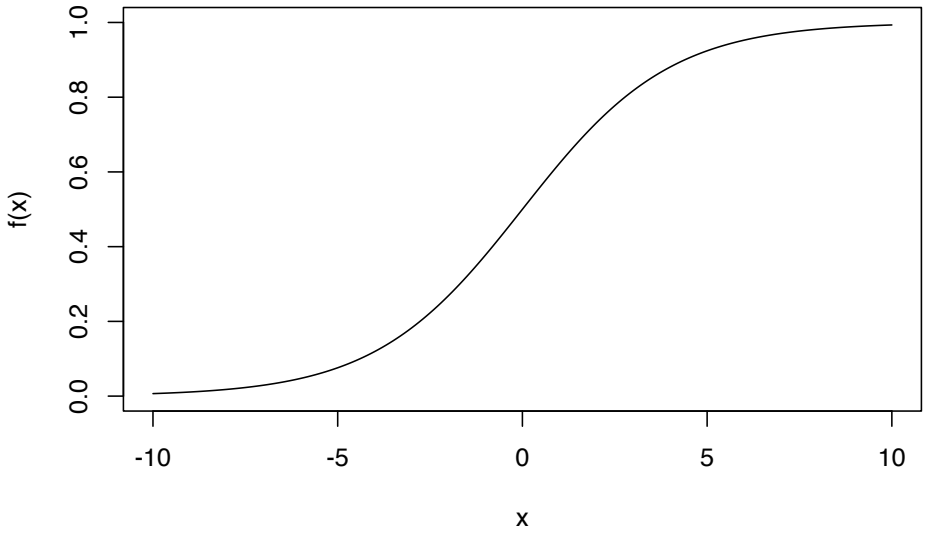
Throughout this section, we will refer to one outcome as 'success' (denoted 1) and 'failure' (denoted 0). Depending on the context of the data, the success could be a negative thing such as 'heart attack' or '20 year mortality' or it could be a positive thing such as 'passing a course.'

We will denote p to be the chance of success and $1 - p$ be the chance of failure. We want to build a model to explain why the chance of one outcome (success) may be higher for one group of people in comparison to another.

3.10.1 Logistic and Logit

The **logistic function** is an S shaped curve (sigmoid curve). For our purposes, the function will take the form

$$f(x) = \frac{1}{1 + e^{b_0 + b_1 x}}$$



For any real value x, this function, $f(x)$, will be a value between 0 and 1. This is perfect for us since probabilities or chances should also be between 0 and 1.

In fact, we'll let the chance of the other outcome (failure), $1 - p$, be modeled by this S function.

$$1 - p = \frac{1}{1 + e^{b_0 + b_1 x}}$$

Math Box: With a bit of algebra and rearranging terms, we can write this equation in terms of p, the chance of success.

$$1 - p = \frac{1}{1 + e^{b_0 + b_1 x}}$$

$$p = 1 - \frac{1}{1 + e^{b_0 + b_1 x}}$$

$$p = \frac{1 + e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} - \frac{1}{1 + e^{b_0 + b_1 x}}$$

$$p = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

Let's define one more term. The **odds** of an success outcome is the ratio of the chance of success to the chance of failure, odds $= p/(1 - p)$.

Math Box: With a bit more algebra and rearranging terms, we can write the above model as a linear regression model.

$$p/(1-p) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} / \frac{1}{1+e^{b_0+b_1x}}$$
$$p/(1-p) = e^{b_0+b_1x}$$
$$\log(p/(1-p)) = b_0 + b_1x$$

This is a **simple logistic regression model**. On the left hand side, we have the natural log of the odds, called the **logit** function. Think of this as a transformed version of our outcome. On the right hand side, we have a familiar linear equation.

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1x$$

Like a linear regression model, we can extend this model to a **multiple logistic regression model** by adding additional x variables,

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$$

3.10.2 Fitting the Model

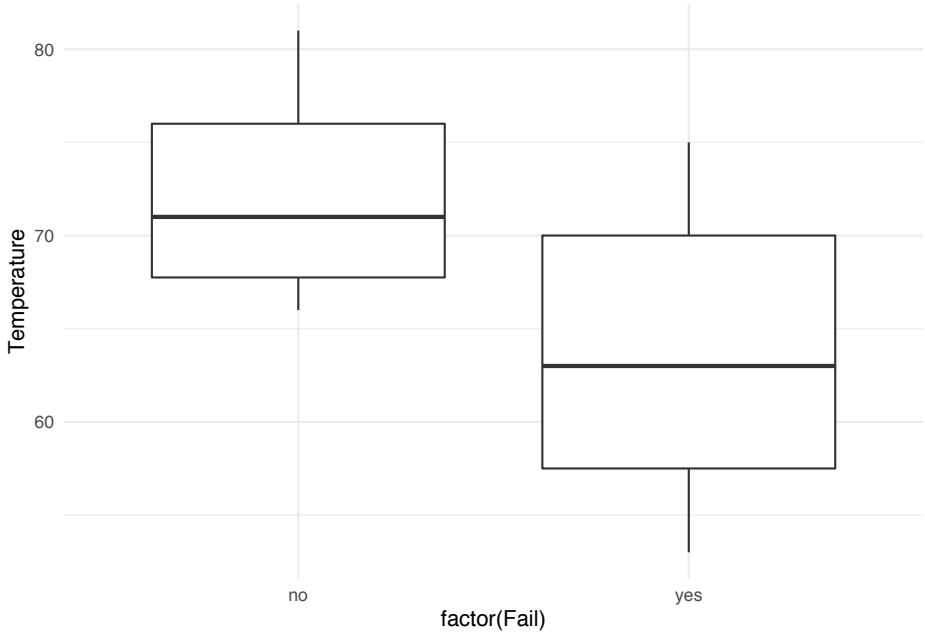
Based on observed data that includes an indicator variable for the responses (1 for success, 0 for failure) and predictor variables, we need to find the slope coefficients, b_0,...,b_k that best fits the data. The way we do this is through a technique called **maximum likelihood estimation**. We will not discuss the details in this class; we'll save this for an upper level stats class such as Mathematical Statistics.

In R, we do this with the **g**eneral **l**inear **m**odel function, `glm()`.

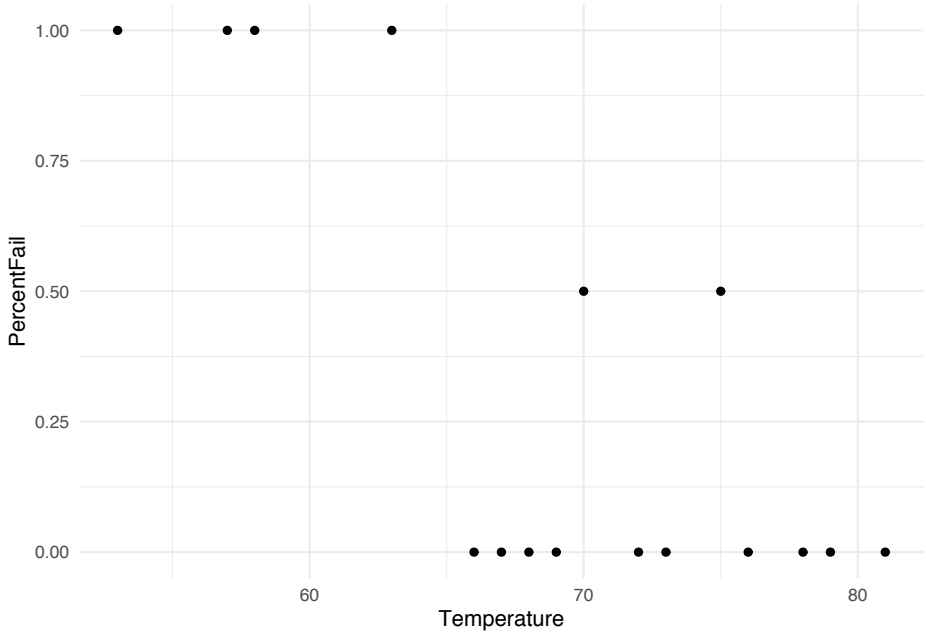
For a data set, let's go back in history to January 28, 1986. On this day, the U.S. space shuttle called Challenger took off and tragically exploded about minute after the launch. After the fact, scientists ruled that the disaster was due to an o-ring seal failure. Let's look at experimental data on the o-rings prior to the fateful day.

```
require(vcd)
data(SpaceShuttle)

SpaceShuttle %>%
```

```
  filter(!is.na(Fail)) %>%
  ggplot(aes(x = factor(Fail), y = Temperature) )+
  geom_boxplot() +
  theme_minimal()
```




```
SpaceShuttle %>%
  filter(!is.na(Fail)) %>%
  group_by(Temperature) %>%
  summarise(PercentFail = mean(Fail == 'yes')) %>%
  ggplot(aes(x = Temperature, y = PercentFail)) +
  geom_point() +
  theme_minimal()
```



Reflect: What are the plots above telling us about the relationship between chance of o-ring failure and temperature?

Let's fit a simple logistic regression model to predict the chance of o-ring failure (which is our "success" here – we know it sounds morbid) based on the temperature using the experimental data.

```
model.glm <- glm(Fail ~ Temperature, data = SpaceShuttle, family = binomial)
summary(model.glm)
```

```
##
## Call:
## glm(formula = Fail ~ Temperature, family = binomial, data = SpaceShuttle)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.061  -0.761  -0.378   0.452   2.217
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   15.043      7.379    2.04    0.041 *
## Temperature   -0.232      0.108   -2.14    0.032 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 24.32
##
## Number of Fisher Scoring iterations: 5
```

Based on a logistic regression model, the predicted log odds of an o-ring failure is given by

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 15.0429 - 0.2322 \cdot Temperature$$

3.10.3 Interpretation

Let's take a look at the estimates from the model. What do they mean?

```
summary(model.glm)
```

```
##
## Call:
## glm(formula = Fail ~ Temperature, family = binomial, data = SpaceShuttle)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.061  -0.761  -0.378   0.452   2.217
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   15.043      7.379    2.04    0.041 *
## Temperature   -0.232      0.108   -2.14    0.032 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 24.32
##
## Number of Fisher Scoring iterations: 5
```

The slope coefficient (-0.232) tells you how much the predicted log odds increase with an increase of 1 degree increase in temperature. But what does a 1 unit

increase in log odds mean? We need to do a bit of algebra.

Math Box: Our estimated model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1 x$$

where $b_0 = 15.04$ and $b_1 = -0.232$.
If we imagine increasing x by 1, then we get a different set of predicted chances of success, $\hat{p}^*$,

$$\log\left(\frac{\hat{p}^*}{1-\hat{p}^*}\right) = b_0 + b_1(x+1)$$

Let's find the difference between these two equations,

$$\log\left(\frac{\hat{p}^*}{1-\hat{p}^*}\right) - \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = (b_0 + b_1(x+1)) - (b_0 + b_1 x)$$

and simplify the right hand side (don't you love it when things cancel!),

$$\log\left(\frac{\hat{p}^*}{1-\hat{p}^*}\right) - \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_1$$

and then simplify the left hand side (using our rules of logarithms),

$$\log\left(\frac{\hat{p}^*/(1-\hat{p}^*)}{\hat{p}/(1-\hat{p})}\right) = b_1$$

Let's exponentiate both sides,

$$\left(\frac{\hat{p}^*/(1-\hat{p}^*)}{\hat{p}/(1-\hat{p})}\right) = e^{b_1}$$

We find that $e^{b_1} = e^{-0.232} = 0.793$ is the **odds ratio** based on increasing x by 1 unit (it is a ratio of odds). The odds ratio is the ratio of the odds of success between two groups of units (those with temperature = T and those with temperature = T + 1 such as 28 v 29 degrees)

If a ratio is greater than 1, that means that the denominator is less than the numerator or equivalently the numerator is greater than denominator (odds of success are greater with T+1 as compared to T –> positive relationship between x variable and odds of success). If a ratio is less than 1, that means that the denominator is greater than the numerator or equivalently the numerator is less than denominator (odds of success are less with T+1 as compared to T –>

negative relationship between x variable and odds of success). If the ratio is equal to one, the numerator equals the denominator (odds of success are equal for the two groups –> no relationship).

In this case, we have an odds ratio < 1 which means that the estimated odds of o-ring failure is lower for increased temperatures (in particular by increasing by 1 degree). This makes sense since we saw that the chance of o-ring failure decrease with warmer temperatures.

3.10.4 Prediction

On January 28, 1986, the temperature was 26 F degrees. Let's predict the chance of "success," which is a failure of o-rings in our data context, at that temperature.

```
predict(model.glm, newdata = data.frame(Temperature = 26), type = 'response') #type =
```

```
## 1
## 1
```

They didn't have any experimental data testing o-rings at this low of temperatures, but even based on the data collected, we predict the chance of failure to be nearly 1 (near certainty).

3.10.4.1 Hard Predictions/Classifications

These predicted chances of "success" are useful to give us a sense of uncertainty in our prediction. If the chance of o-ring failure were 0.8, we would be fairly certain that a failure were likely but not absolutely. If the chance of o-ring failure were 0.01, we would be very certain that a failure were not likely to occur.

But if we had to decide whether or not we should let the shuttle launch go, how high should the predicted chance be to delay the launch (even if it would cost a lot of money to delay)?

It depends.

If we used a threshold of 0.8, then we'd say that for any experiment with a predicted chance of o-ring failure 0.8 or greater, we'll predict that there will be o-ring failure. As with any predictions, we may make an error. With this threshold, what is our **accuracy** (# of correctly predicted/# of data points)?

In the table below, we see that there were three data points in which we correctly predicted o-ring failure (using a threshold of 0.8). There were 4 data points in which we erroneously predicted that it wouldn't fail when it actually did and correctly predicted no failure for 16 data points. So in total, our accuracy is (16+3)/(16+4+3) = 0.82 or 82%.

The only errors we made were **false negatives**; we didn't predict o-ring failure but the o-rings did actual happen in the experiment. In this data context, false

negatives have real consequences on human lives because a shuttle would launch and potentially explode because we had predicted there would be no o-ring failure.

The **false negative rate** is the number of false negatives divided by the false negatives + true positives (denominator should be total number of experiments with actual o-ring failures). With a threshold of 0.80, our false negative rate is $4/(3+4) = 0.57 = 57\%$. We failed to predict 57% of the o-ring failures. This is fairly high when there are lives on the line.

A **false positive** (predicting failure when it doesn't happen) would delay launch but have minimal impact on human lives.

The **false positive rate** is the number of false positives divided by the false positives + true negatives (denominator should be total number of experiments with no o-ring failures). With a threshold of 0.80, our false positive rate is $0/16 = 0$. We always accurately predicted the o-rings would not fail.

```
augment(model.glm, type.predict ='response') %>%
  mutate(predictOFail = .fitted >= 0.8) %>%
  count(Fail, predictOFail)
```

```
## # A tibble: 3 x 3
##   Fail  predictOFail     n
##   <fct> <lgl>        <int>
## 1 no    FALSE           16
## 2 yes   FALSE            4
## 3 yes   TRUE             3
```

What if we used a lower threshold to reduce the number of false negatives (those with very real human consequences)? Let's lower it to 0.25 so that we predict o-ring failure more easily. Let's find our accuracy: $(10+4)/(10+3+6+4) = 0.61$. Worse than before, but let's check false negative rate: $3/(3+4) = 0.43$. That's lower. But now we have a non-zero false positive rate: $6/(6+10) = 0.375$. So of the experiments with no o-ring failure, we predicted wrong 37.5% of the time.

```
augment(model.glm, type.predict ='response') %>%
  mutate(predictOFail = .fitted >= 0.25) %>%
  count(Fail, predictOFail)
```

```
## # A tibble: 4 x 3
##   Fail  predictOFail     n
##   <fct> <lgl>        <int>
## 1 no    FALSE           10
## 2 no    TRUE             6
## 3 yes   FALSE            3
## 4 yes   TRUE             4
```

Where the threshold goes depends on the real consequences of making those two

types of errors.

3.10.5 Model Evaluation

In deciding whether a logistic regression model is a good and useful model, we need to consider the accuracy, the false positive rate, and the false negative rate. Depending on the context, we may focus on maximizing the overall accuracy or we may focus on minimizing the false negative rate or minimizing the false positive rate.

There are other measures of model fit for a logistic regression such as AIC and BIC (lower is better), but those are beyond the scope of this course. You'll learn about them in future statistics courses.

3.10.6 Alternative Classification Models

Logistic regression is a very useful model to predict a binary outcome. However, it has its limitations. We are assuming a linear relationship between explanatory variables and the log odds of success. This is hard to check because we don't have a variable for odds that we could quickly plot.

Other methods out there are more flexible but also more complex. Here is a list of some of the most popular classification methods.

- Classification Trees can predict a binary outcome and choose the variables that are most important by recursively partitioning the data into groups that are more similar in terms of the outcome as well as in chosen predictor variables.
- Random Forests are an ensemble of classification tress that together are more stable than any one classification tree.
- Boosted Trees are classification trees that are sequentially created to target the errors from the last tree.
- Neural Networks are a type of classification algorithm that creates new features based on the original data that are the best predictors of the outcome.

Take Statistical Machine Learning to learn more about these methods. But keep in mind that sometimes for a task, complex is not necessary: see https://www.huffingtonpost.com/2014/02/10/klemens-torggler-evolution-door_n_4762261.html.

3.11 Major Takeaways

1. All models are wrong, but some are useful and fair for our goals (prediction or description).

2. We want a model with small residuals or a low number of prediction errors.
3. To determine if a model is useful and fair, we study what is left over (the residuals).
4. We use models to describe phenomena by interpreting slope coefficients. But make sure you are talking about the average or predicted outcome! If you have multiple variables, you are keeping all others fixed (if possible).
5. We also use models to prediction values, but be careful about predicting outside the observed range of our explanatory (x) variables. That is called extrapolation.

Chapter 4

Random Variability

Up until this point, we have thought about

1. Data collection process (sampling and study design) and Data Quality (issues of bias)
2. Data visualization (the first step of any data analysis)
3. Modeling (to explain observed variation and provide predictions)

Throughout the past three chapters, we have also sprinkled in the idea that **the sample we observe is just one random representation of the true population or phenomenon.** Therefore, any **statistic**, any numerical summary of the sample, would be an estimate the true numerical summary of the population (which we refer to as the corresponding unknown **parameter**).

4.1 Sampling Variability

If we could repeat the random sampling process, each sample we would get would be slightly different. The composition of units in the sample would differ every time. Sometimes we would randomly over-represent one group of people and another time we would randomly over-represent another group of people, just by chance. The statistic of interest (e.g. a mean, a median, a regression coefficient) that we calculate would vary between across the different samples.

4.2 Randomization Variability

Another source of random variation in a statistic that can arise is that of random treatment/control group assignments in an experiment. If we could repeat the randomization process in an experiment, each treatment group would be slightly different. The individual composition of the units in the treatment and control groups would differ every time. Sometimes we would randomly over-represent

one group of people in the treatment group, and another time we would randomly over-represent another group of people in the treatment group, just by chance. The statistic that we calculate and compare between groups (e.g. a difference in means, a difference in medians, a regression coefficient) would change for every reshuffling of the individuals.

These descriptions cover the idea of **random variability**. This random variability is typically due to random sampling or randomized group assignments. The sample composition can vary, and therefore, the statistics that we calculate vary between different sample.

The important thing to keep in mind is that we only get to see one sample, one particular composition of individuals. But we need to put this one observed sample and statistic in the context of the random variability.

Let's explore this concept a bit more before we formally talk about probability and chance.

4.3 Simulating Random Sampling from a Population

The data set that we will work with contains ALL flights leaving New York City in 2013. This data represents a full census of the target population of flights leaving NYC in a particular year.

We'll start by creating two new variables, `season` defined as winter (Oct - March) or summer (April - Sept) and `day_hour` defined as morning (midnight to noon) or afternoon (noon to midnight).

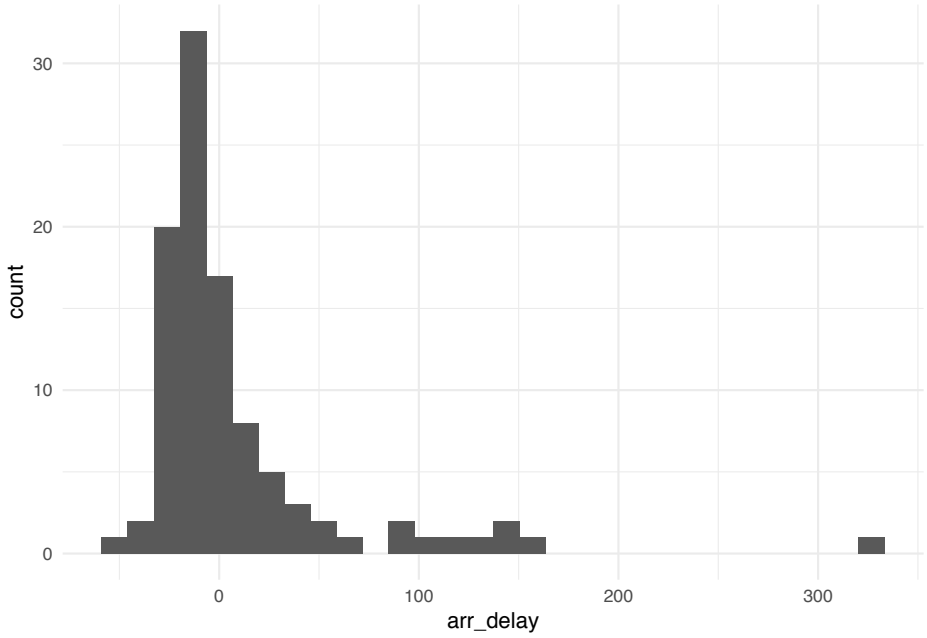
```
data(flights)
flights <- flights %>%
    na.omit() %>%
    mutate(season = case_when(
      month %in% c(10:12, 1:3) ~ "winter",
      month %in% c(4:9) ~ "summer"
    )) %>%
    mutate(day_hour = case_when(
      between(hour, 1, 12) ~ "morning",
      between(hour, 13, 24) ~ "afternoon"
    )) %>%
    select(arr_delay, dep_delay, season, day_hour, origin, carrier)
```

Since we have the full population of flights in 2013, we could just describe the flights that happened in that year. Having data on the full population is very rare in practice. Instead, we are going to use this population to illustrate sampling variability.

Let's take one random sample of 100 flights from the data set, using a simple

random sampling strategy. Let's look at the arrival delay (in minutes) `arr_delay` with a histogram and calculate the median and mean arrival delay.

```
flights_samp1 <- flights %>%
    sample_n(size = 100) ## Sample 100 flights randomly

flights_samp1 %>%
    ggplot(aes(x = arr_delay)) +
    geom_histogram() +
    theme_minimal()
```



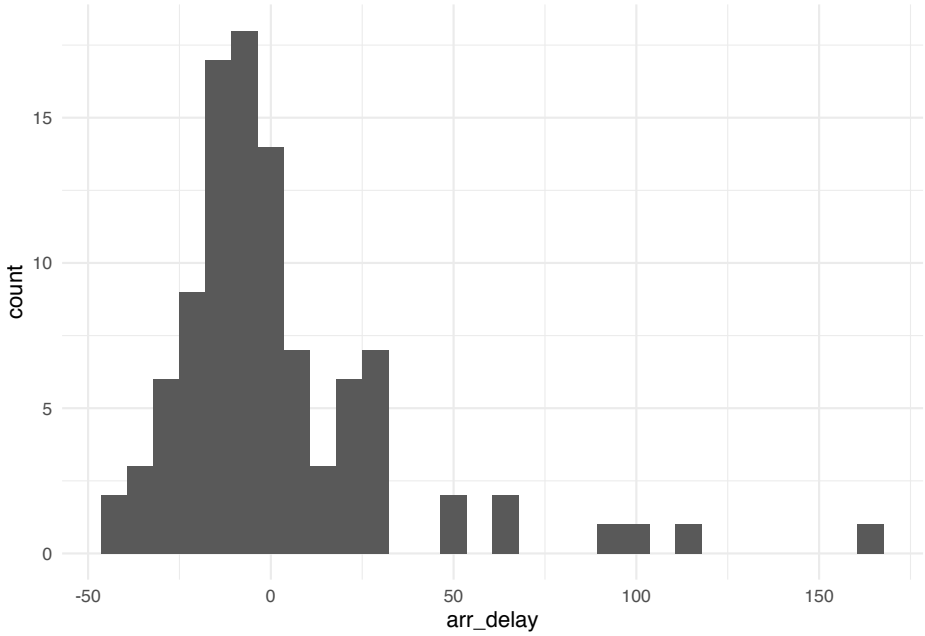

```
flights_samp1 %>%
    summarize(medians = median(arr_delay), means = mean(arr_delay))
```

```
## # A tibble: 1 x 2
##   medians means
##     <dbl> <dbl>
## 1      -9  7.09
```

Reflect: At this point, we haven't looked at the entire population of flights from 2013. Based on a sample of 100 flights, what do you think the distribution of arrival delays looks like for the entire population? Shape? Center? Spread?
Why do you think that?

Now, let's take another random sample of 100 flights from the full population of flights.

```
flights_samp2 <- flights %>%
    sample_n(size = 100) ## Sample 100 flights randomly

flights_samp2 %>%
    ggplot(aes(x = arr_delay)) +
    geom_histogram() +
    theme_minimal()
```




```
flights_samp2 %>%
    summarize(medians = median(arr_delay), means = mean(arr_delay))
```

```
## # A tibble: 1 x 2
##   medians means
##     <dbl> <dbl>
## 1    -4.5  2.14
```

Reflect: How does the second sample differ from the first sample? What do they have in common?

We could keep the process going. Take a sample of 100 flights, look at the histogram, and calculate the median and mean. Repeat many, many times.

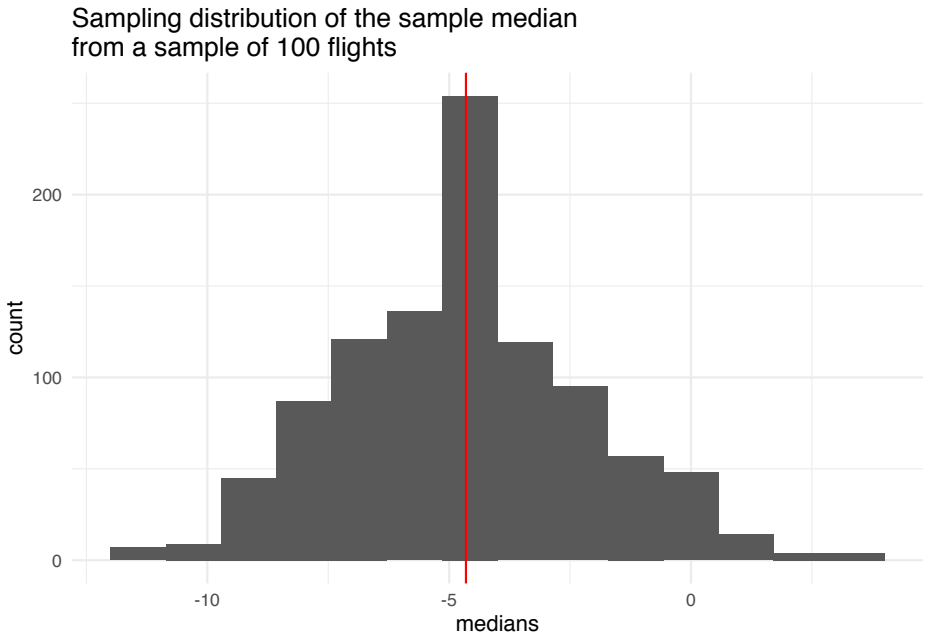
We can add a little bit of code to help us simulate this sampling process 1000

times and calculate the median and mean arrival delay for each random sample of 100 flights.

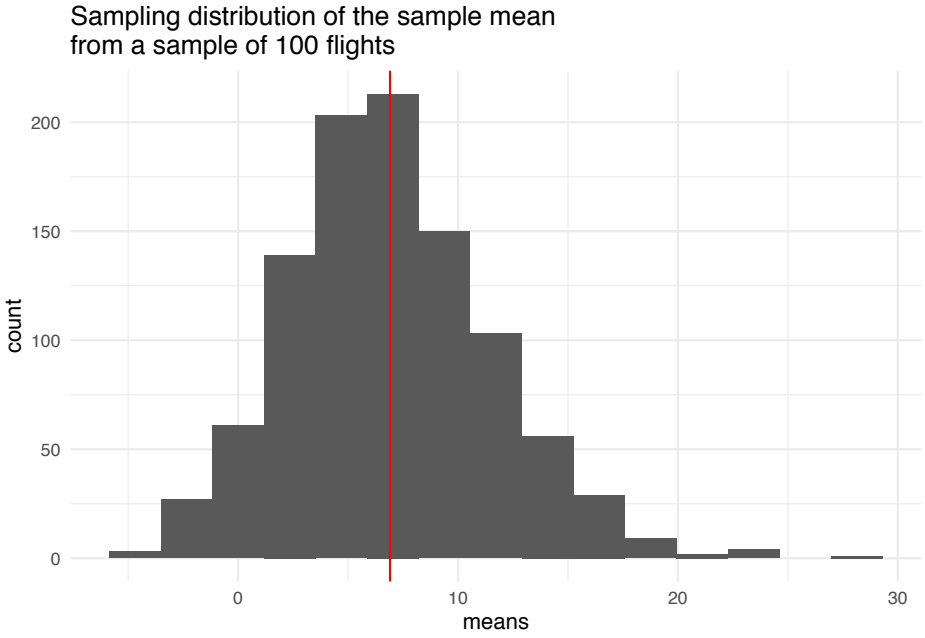
```
sim_data <- mosaic::do(1000)*(
    flights %>%
      sample_n(size = 100) %>%  # Generate samples of 100 flights
      summarize(medians = median(arr_delay), means = mean(arr_delay)) # Calculate the median and
)
```

Now we have 1000 medians and 1000 means, each corresponding to one random sample of 100 flights from the population. Let's summarize and visualize this simulation.

```
# Summarize
summary_sim <- sim_data %>%
    summarize(
        mean_medians = mean(medians),
        mean_means = mean(means),
        sd_medians = sd(medians),
        sd_means = sd(means))

# Visualize M
sim_data %>%
    ggplot(aes(x = medians)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_sim, aes(xintercept = mean_medians), color = "red") +
    labs(title = "Sampling distribution of the sample median\nfrom a sample of 100 flights") +
    theme_minimal()
```




```
# Visualize
sim_data %>%
    ggplot(aes(x = means)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_sim, aes(xintercept = mean_means), color = "red") +
    labs(title = "Sampling distribution of the sample mean\nfrom a sample of 100 flight
    theme_minimal()
```



These histograms approximate the **sampling distribution** of sample median arrival delay and the **sampling distribution** of sample mean arrival delay, both of which describe the variability in the sample statistic *across all possible random samples from the population.*

> **Reflect:** Describe the shape, center, and spread of the sampling distribution for the sample mean.

4.4 IRL: Bootstrapping

In real life (IRL), we don't have a full target population from which we can repeatedly draw samples. We only have one **sample** that was already drawn from the larger target population.

To get a sense of the sampling variability, we could try to mimic this process of sampling from the full target population using our best stand-in for the population: our sample. We will call the sample our "fake population" for the moment. This process of resampling our sample is called **bootstrapping**.

We bootstrap our sample in order to 1) estimate the variability of the statistic and 2) get a range of plausible values for the true population parameter.

There are four steps to bootstrapping. They are very similar to simulating the sampling process from a population.

1. Generate

To generate different random samples of the same size (100 flights) from our "fake population", we have to draw sample of 100 flights WITH REPLACEMENT, meaning that we have to put a flight back into the pool after drawing them out.

> **Reflect:** What would happen if we drew WITHOUT REPLACEMENT?

2. Calculate

In our simulation above, we calculated the median and mean arrival delay. In theory, we could calculate any numerical summary of data (e.g. the mean, median, SD, 25th percentile, etc.)

```
boot_data <- mosaic::do(1000)*(
    flights_samp1 %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      summarize(medians = median(arr_delay), means = mean(arr_delay)) # Calculate stat
)
```

3. Summarize

Let's summarize these 1000 medians and 1000 means generated from resampling (with replacement) *from our sample* (our "fake population").

```
# Summarize
summary_boot <- boot_data %>%
    summarize(
        mean_medians = mean(medians),
        mean_means = mean(means),
        sd_medians = sd(medians),
        sd_means = sd(means)
    )
summary_boot
```

```
## # A tibble: 1 x 4
##   mean_medians mean_means sd_medians sd_means
##          <dbl>      <dbl>      <dbl>    <dbl>
## 1        -9.13       7.25       2.17     5.09
```

Let's compare this to the summaries from the simulation of randomly sampling *from the population.*

```
summary_sim
```

```
## # A tibble: 1 x 4
##   mean_medians mean_means sd_medians sd_means
##          <dbl>      <dbl>      <dbl>    <dbl>
```

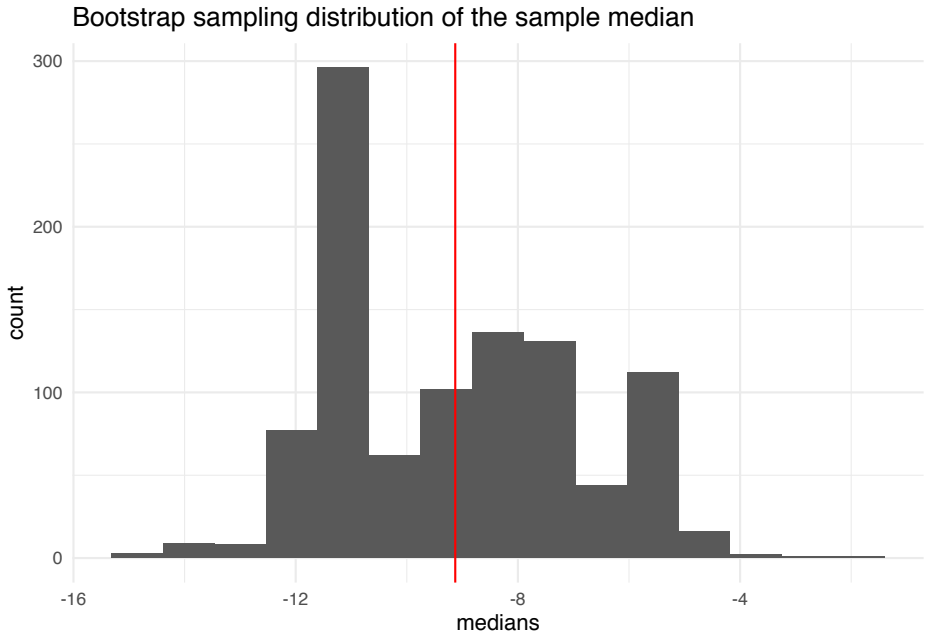
```
## 1         -4.65        6.91        2.58      4.60
```

They won't be exactly the same, but they should be of roughly similar magnitude.

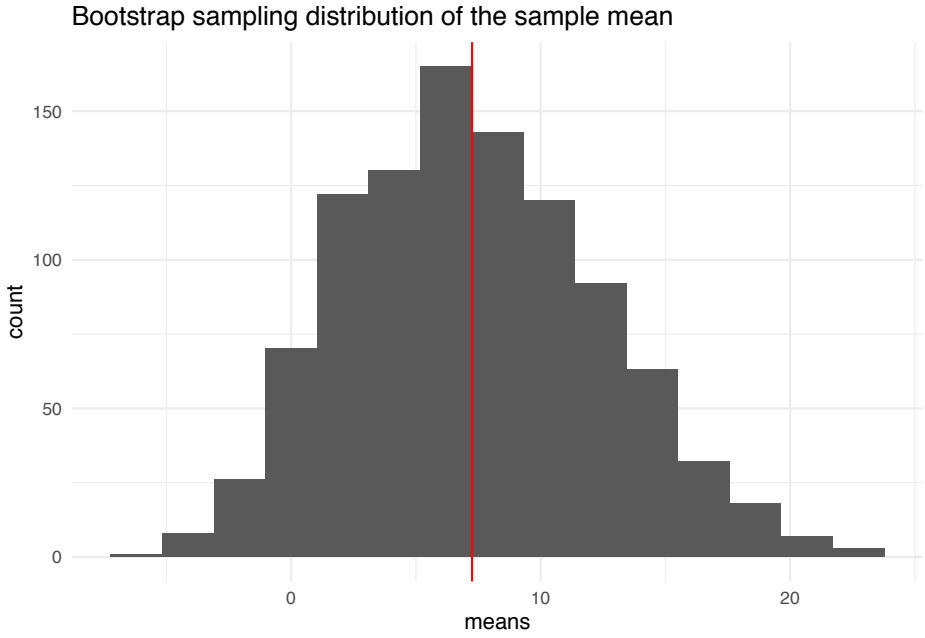
4. Visualize

Let's visualize these 1000 medians and 1000 means generated from resampling (with replacement) from our sample (our "fake population").

```
# Visualize
boot_data %>%
    ggplot(aes(x = medians)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_boot, aes(xintercept = mean_medians), color = 'red') +
    labs(title = "Bootstrap sampling distribution of the sample median") +
    theme_minimal()
```

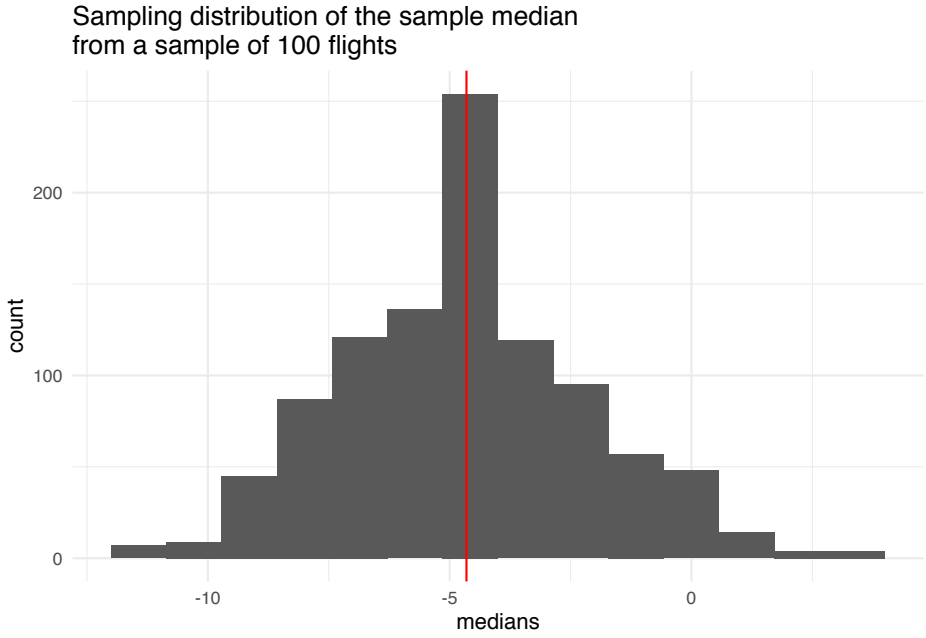



```
# Visualize
boot_data %>%
    ggplot(aes(x = means)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_boot, aes(xintercept = mean_means), color = 'red') +
    labs(title = "Bootstrap sampling distribution of the sample mean") +
    theme_minimal()
```

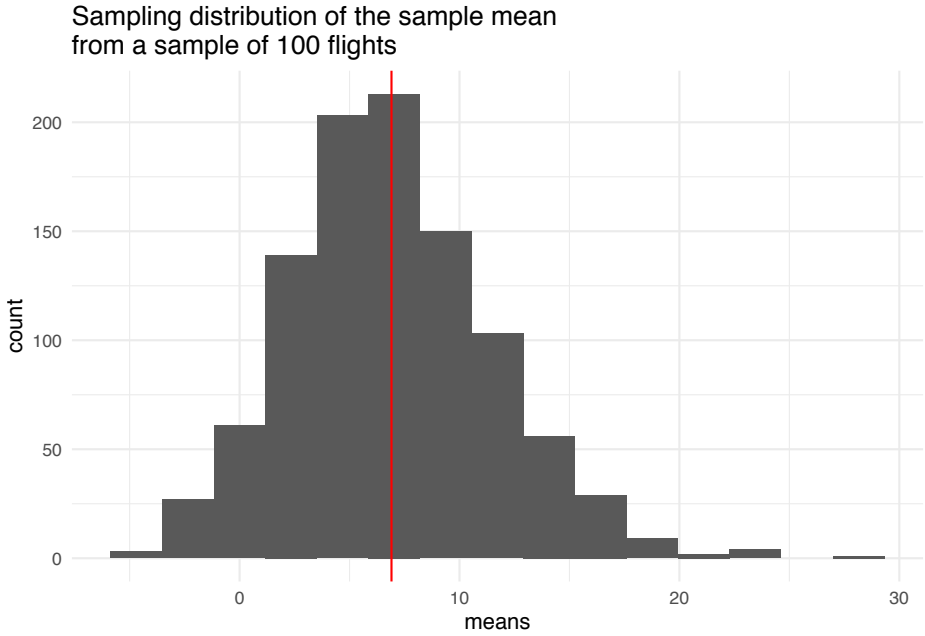


Let's compare these to the visuals from the simulation from the population.

```
# Visualize
sim_data %>%
    ggplot(aes(x = medians)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_sim, aes(xintercept = mean_medians), color = 'red') +
    labs(title = "Sampling distribution of the sample median\nfrom a sample of 100 flig
    theme_minimal()
```




```r
# Visualize
sim_data %>%
    ggplot(aes(x = means)) +
    geom_histogram(bins = 15) +
    geom_vline(data = summary_sim, aes(xintercept = mean_means), color = 'red') +
    labs(title = "Sampling distribution of the sample mean\nfrom a sample of 100 flights") +
    theme_minimal()
```



The process of resampling from our sample, called **bootstrapping**, is becoming the one of main computational tools for estimating sampling variability in the field of Statistics.

> **Reflect:** How well does bootstrapping do in mimicking the simulations from the population? What could we change to improve bootstrap's ability to mimic the simulations?

This is a really important concept in Statistics! We'll come back to the ideas of sampling variability and bootstrapping throughout the rest of the course.

> **Reflect:** What do you think the mean arrival delay is in the population? If you had to give an interval of plausible values for the population mean, what range would you give? Why?

4.5 Simulating Randomization into Groups

We have been thinking about arrival delays in general. Let's move now to comparing arrival delays between groups. If you were planning a trip, you may be able to choose between two flights that leave at different times of day. Do morning flights have shorter arrival delays on average than afternoon flights? If so, book the flight earlier in the day!

Let's look at the data! We use a random sample of 500 flights from the population to investigate this question.

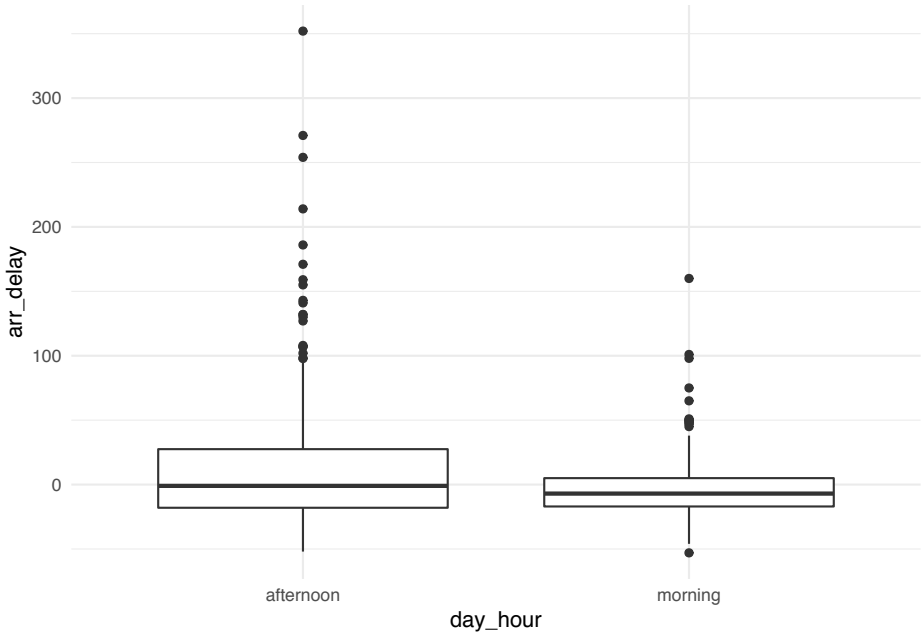
```
flights_samp500 <- flights %>%
    sample_n(size = 500)
```

Let's summarize and visualize the relationship between hour of the day (morning or afternoon) and the arrival delay.

```
flights_samp500 %>%
    group_by(day_hour) %>%
    summarize(median = median(arr_delay), mean = mean(arr_delay))
```

```
## # A tibble: 2 x 3
##   day_hour  median  mean
##   <chr>      <dbl> <dbl>
## 1 afternoon     -1 12.6
## 2 morning       -7 -2.39
```

```
flights_samp500 %>%
    ggplot(aes(x = day_hour, y = arr_delay)) +
    geom_boxplot() +
    theme_minimal()
```



Reflect: Based solely on the visual and numerical summaries, are arrival delays less in the morning than in the afternoon?

We don't know the exact reason why some flights were scheduled in the morning or the afternoon and why one flight might be delayed (it's probably due to a complex combination of factors). Let's imagine that a randomization process was used to decide when particular flights were scheduled (morning or afternoon); a flip of a coin to decide morning or afternoon.

We want to compare the mean arrival delays in morning flights and in afternoon flights.

If there were **no difference** in arrival delays between morning and afternoon flights, then it wouldn't matter whether a flight left in the morning or afternoon. That is, the `day_hour` variable would be irrelevant to the arrival delay `arr_delay`. If that were true, then we could reshuffle the values of `day_hour` and it wouldn't change our conclusions.

Wouldn't it be great if we could see how the mean arrival delays might change if we shuffled the flights between the "morning" group and "afternoon" group, randomly?

In fact, wouldn't it be great if we could look at every permutation of flights between two groups?

4.6 IRL: Randomization Tests

In real life, we don't often consider every possible permutation (reshuffling of group members) due to the immensely large number of permutations. However, we can randomly reshuffle flights about 1000 times to try to approximate many of those permutations. Such a procedure is called a **randomization or permutation procedure**. There are three main steps:

1. Hypothesis

Our **null hypothesis** (a hypothesis that is conservative/not interesting/does not elicit action) is that there is no relationship between time of day and the arrival delay.

2. Generate and Calculate

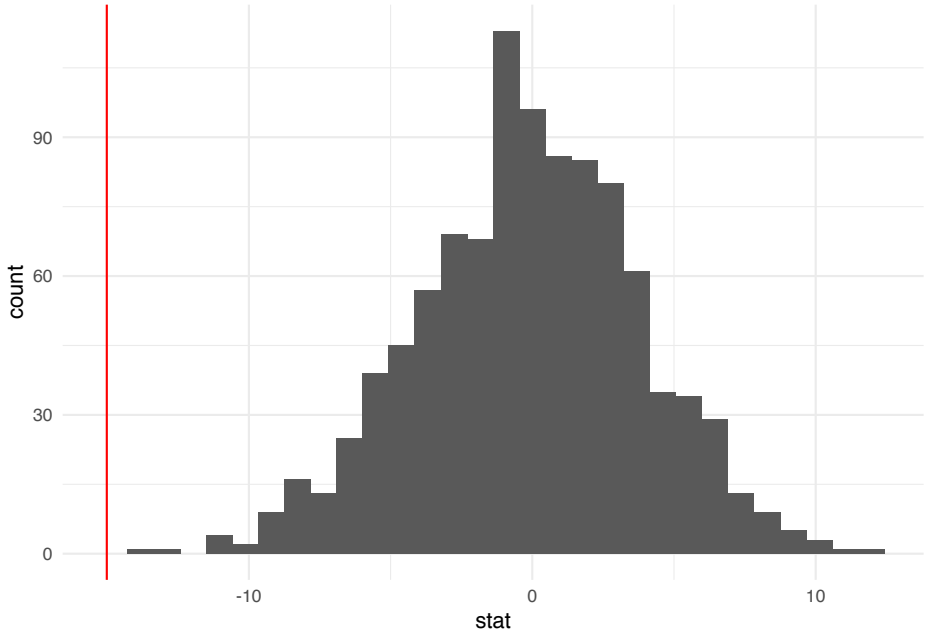
We can generate 1000 new data sets based on randomly reshuffling the labels of `day_hour`. For each of these data sets, we calculate the difference in mean arrival delay between morning and afternoon groups.

```
require(infer) #install.packages('infer')
null_dist <-  flights_samp500 %>%
    specify(arr_delay ~ day_hour) %>%
    hypothesize(null = "independence") %>% # Hypothesize
    generate(reps = 1000, type = "permute") %>% # Generate permutations
    calculate(stat = "diff in means", order = c("morning", "afternoon")) #calculate di
```

3. Visualize

The histogram below shows the histogram of differences in means if the null hypothesis were true. The vertical line shows the observed difference in means.

```
obs <- flights_samp500 %>%
    specify(arr_delay ~ day_hour) %>%
    calculate(stat = "diff in means", order = c("morning", "afternoon"))

null_dist %>%
    ggplot(aes(x = stat)) +
    geom_histogram() +
    geom_vline(data = obs, aes(xintercept = stat), color = "red") +
    theme_minimal()
```



Reflect: Do you think that the mean arrival delay is different for morning and afternoon? Is the observed difference in means likely to have occurred if there were no relationship?

We will return to the ideas of testing hypotheses later in the course.

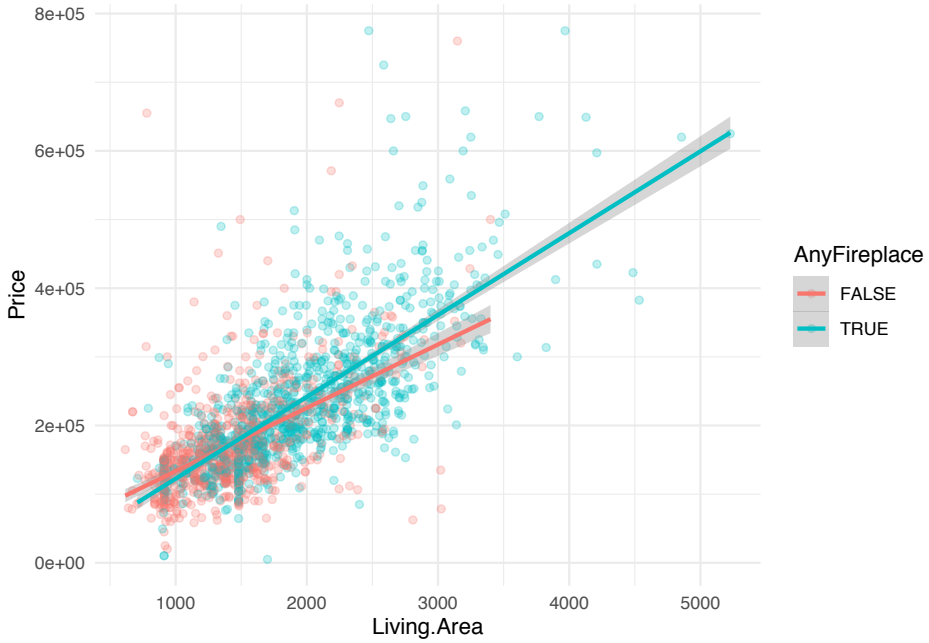
Chapter 5

Randomness and Probability

Now that we have some intuition about random variability, we will formalize some of the concepts of probability and chance.

Recall the regression model we built to predict home price as a function of square footage and fireplaces in Multiple Linear Regression.

```
homes <- read.delim("http://sites.williams.edu/rdeveaux/files/2014/09/Saratoga.txt")
homes <- homes %>%
    mutate(AnyFireplace = Fireplaces > 0)
homes %>%
    ggplot(aes(x = Living.Area, y = Price, color = AnyFireplace)) +
    geom_point(alpha = 0.25) +
    geom_smooth(method = "lm") +
    theme_minimal()
```



To allow for different slopes among homes with and without a fireplace, we used an interaction term between Living.Area and AnyFireplace. Based on our one sample, we observed a difference in the slopes of $26.85 per square foot.

But is this true for the larger population of homes in the area? Is each square foot worth exactly $26.85 more, on average, in homes with a fireplace than in homes without a fireplace?

```
lm.home3 <- lm(Price ~ AnyFireplace*Living.Area, data = homes)
tidy(lm.home3)
```

```
## # A tibble: 4 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     40901.    8235.        4.97 7.47e- 7
## 2 AnyFireplaceTRUE               -37610.   11025.       -3.41 6.61e- 4
## 3 Living.Area                       92.4      5.41      17.1  1.84e-60
## 4 AnyFireplaceTRUE:Living.Area      26.9      6.46       4.16 3.38e- 5
```

Probably not. In fact, our sample is just one random sample from the larger population of homes in upstate New York. If we had gotten a slightly different sample of homes, then we would have different estimates for each of the regression coefficients. We explored this sampling variation in Chapter 4 using bootstrapping.

Let's connect our goals to the terms **statistic**, **estimate**, and **parameter**.

The interaction coefficient is a **statistic**. It (as well as the other coefficient

estimates) is a numerical summary of our data that was estimated from a **sample**. The actual numerical value of the interaction coefficient in the R output is called the **estimate**. If we had a census, a full set of data on all homes in upstate New York, we could fit the same linear regression model. The coefficient estimates given to us by R would represent population **parameters** because they are computed from the whole population, rather than a sample. By understanding **how much** statistics vary from sample to sample, we can start to **quantify the amount of uncertainty** in our estimates. Because the process of obtaining a sample is a type of random process, we will spend some time discussing formal probability.

This chapter briefly discusses the theory of formal probability so that we have terminology and basic concepts to understand and discuss random events. This framework provides a way of thinking about **uncertainty**, **random variability**, and **average behavior in the long run.**

A **random process** or **random event** is any process/event whose outcome is governed by chance. It is any process/event that cannot be known with certainty. Examples range from the outcome of flipping a coin to the estimated model coefficients from randomly selected samples.

We've used the term "chances" up until now. We are now going to use "probability" as an equivalent word for "chance".

For a much more in-depth discussion of probability (calculus-based), take STAT/MATH 354 (Probability).

5.1 Three Types of Probability

Reflect: What is the probability of rolling a 1 on a six-sided die? How do you know this? How can you justify that number?

There are three types of probability.

1. **Empirical Probability:** If you could repeat a random process over and over again, you'd get a sense of the possible outcomes and their associated probabilities by calculating their relative frequency in the long run. If you repeatedly tossed a balanced die, then the relative frequency of 1's after tossing the die MANY times would be the empirical probability. If you repeatedly got a sample of 100 people, the relative frequency of estimated odds ratios below 1 would be the empirical probability of getting an odds ratio below 1.

2. **Theoretical Probability:** If you don't have time to toss a die a million times, you could calculate probabilities based on mathematical theory and assumptions. When tossing a balanced die, you would assume that each

side is equally likely to land face-up. Thus the chance of rolling a 1, is 1/6 for a six-sided die.

> **Reflect:** What is the probability that you'll talk to someone you do not know this week? What does that number represent? How can you justify that number?

3. **Subjective Probability:** If you use a number between 0 and 1 (100%) to reflect your uncertainty in an outcome (rather than based on empirical evidence or mathematical theory), then you are using subjective probability.

In this class, we'll focus on theoretical and empirical probability. In particular, we will use computational tools to estimate empirical probabilities using simulations (such as bootstrapping and randomization tests) and mathematical tools to estimate theoretical probabilities.

5.2 Theoretical Probability Rules

To understand theoretical probability, we need to define a few terms and set some rules for working with probabilities (known as axioms).

The **sample space**, S, is the set of all possible outcomes of a random process.

- Example: If you flip two coins (each coin has one side Heads and one side Tails), then the sample space contains four possible outcomes: Heads and Heads (HH), Heads and Tails (HT), Tails and Heads (TH), and Tails and Tails (TT). That is, $S = \{HH, HT, TH, TT\}$.

A subset of outcomes is called an **event**, denoted as A.

- Example: If you flip two coins, an event A could be that exactly one of the coins lands Heads, $A = \{HT, TH\}$.

For events A and B and sample space S, the probability of an event A, notated as $P(A)$, follows the rules below:

- Rule 1: $0 \leq P(A) \leq 1$ (probability has to be between 0 and 1)
- Rule 2: $P(S) = 1$ (one of the possible outcomes has to happen)
- Rule 3: $P(\text{not } A) = 1 - P(A)$ (if we know the chance of something happening, we also know the chance that it doesn't happen)
- Rule 4: $P(A \text{ or } B) = P(A) + P(B)$ if A and B are disjoint events.
  - A and B are **disjoint** if A occuring prevents B from occurring (they both can't happen at the same time).
- Rule 5: $P(A \text{ and } B) = P(A) \times P(B)$ if A and B are independent.
  - A and B are **independent** if B occurring doesn't change the probability of A occurring.
- Rule 4*: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ in general
- Rule 5*: $P(A \text{ and } B) = P(A \mid B)P(B) = P(B \mid A)P(A)$ in general

- The **conditional probability** of A **given** that event B occurs, $P(A \mid B)$, is equal to the probability of the joint event (A and B) divided by the probability of B.

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Intuition: Given that B happened, we focus on the subset of outcomes in S in which B occurs and then figure out what the chance of A happening within that subset.

For more details on theoretical probability, please see [Appendix A]. This material is optional but available for those of you who want to understand the mathematical reasoning behind the rest of the chapter.

5.2.1 Diagnotic Testing and Probability

Let's start by taking a moment to consider a recent Washington Post article that discusses the role of probability in medical diagnostics. Before you read the whole article, consider a question.

Reflect: Say that Disease X has a prevalence of 1 in 1,000 (meaning that 1 out of every 1,000 people will have it).
The test to detect Disease X has a false-positive rate of 5 percent (meaning that out of every 100 subjects who do not have Disease X, 5 will falsely test positive for it).
The test's accuracy is 99 percent (meaning that out of every 100 who do have Disease X, 99 will correctly test positive for it).
If a patient's test result comes back positive, what is the probability that this patient actually has the disease?

If you said the probability is 95%, then you are wrong, but almost half of the doctors surveyed in 2014 thought exactly the same thing.

We can use the rules of probability to get a sense of what the desired probability actually is. We want to know the probability that they have the disease GIVEN that they got a positive test result, $P(D \mid +)$, where D stands for disease and $+$ stands for positive test result.

Based on the definition of conditional probability, we must consider only those that got a positive test result back and look at the proportion of them that have the disease. In mathematical notation, that is equal to

$$P(D \mid +) = \frac{P(D \text{ and } +)}{P(+)}$$

What information were we given again?

- The prevalence of the disease is 1 in 1,000, so $P(D) = 1/1000$. Using Rule 3, the probability of no disease is $P(\text{no } D) = 999/1000$. In 1000 people, 1 will actually have the disease and 999 won't have the disease.
- The false-positive rate is 5 percent, so given that you don't have the disease, the probability of getting a false positive is $P(+ \mid \text{ no } D) = 0.05$. So of the 999 that don't have the disease, about $0.05 \times 999 = 49.95$ (about 50) of them will get a false positive test result.
- While it is not stated directly in the Washington Post article, most medical tests have a fairly high accuracy in catching the disease. For example: $P(+ \mid D) = 0.99$. Therefore, the 1 person who actually has the disease will most likely get a positive test result back ($0.99 * 1 = 0.99$).

Remember that our interest is in $P(D \mid +)$. By the definition of conditional probability, we consider only those with positive test results (about 50 who are disease free and 1 who has the disease). So the probability of actually having the disease GIVEN a positive test result is about $1/51 = 0.019$. This is not close to 95%!

In mathematical notation, that looks like this

$$
\begin{aligned}
P(D \mid +) &= \frac{P(D \text{ and } +)}{P(+)} && \text{Rule 5*} \\
&= \frac{P(D \text{ and } +)}{P(+ \text{ and } D) + P(+ \text{ and no } D)} && \text{2 ways you can get } + \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid \text{ no } D)P(\text{no } D)} && \text{Rule 5*} \\
&= \frac{0.99 * 1/1000}{0.99 * 1/1000 + 0.05 * 999/1000} && \text{Plug in values} \\
&= \frac{0.99 * 1}{0.99 * 1 + 0.05 * 999} && \text{Simplify and evaluate} \\
&= 0.019
\end{aligned}
$$

The third line above ($P(D \mid +) = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+| \text{ no } D)P(\text{no } D)}$) is often called **Bayes' Rule**. The important idea to take from this is that what we condition on can make a big difference in the resulting probability.

Now, take some time to read the full Washington Post article.

5.2.2 Court Arguments and Probability

The concept of conditional probability also plays an important role in the judicial system in the U.S. The foundation of the judicial system is the concept of "innocent until proven guilty". Decisions are supposed to be based from a point

of view of assuming that the defendant is innocent. Thus, jurors are supposed to decide the chances of seeing this evidence assuming innocence. That is, evidence is presented to jurors as the conditional probability: $P($ evidence $\mid$ innocent $)$.

Unfortunately, many prosecutors try to make the wrong argument by flipping the conditional probability, whether maliciously or due to a lack of statistical knowledge. They sometimes mistakenly try to argue that it is unlikely that a person is innocent given the evidence that is presented, $P($ innocent $\mid$ evidence $)$.

This can be dangerous. We know that $P($ evidence $\mid$ innocent $) \neq P($ innocent $\mid$ evidence $)$ based on the disease testing example above. Generally, $P(A \mid B)$ is not equal to (and can be very different from) $P(B \mid A)$.

This is known as the prosecutor's fallacy. You can read more about it here.

5.3 Random Variable

With a basic understanding of theoretical probability rules, we can introduce the most important concept from probability for our uses in this class: a random variable.

A **random variable** (X) is variable whose outcome (the value it takes) is governed by chance. In other words, it is a variable (something capable of taking different values) whose value is random. Examples include:

- $X =$ age of the next person to walk into the building
- $X =$ the number of dots on the side that lands face up on a balanced 6-sided die

When considering data analysis and modeling, the random variables we will be considering will be estimated regression coefficients, estimated odds ratios, etc. Why are these random variables? Because their values depend on the random samples that we draw. To establish our understanding, let's start with a simple example.

You are going to flip a fair coin 3 times (the coin has 2-sides, we'll call one side Heads and the other Tails).

- Assume there are only 2 possible outcomes and $P(\text{Heads}) = P(\text{Tails}) = 0.5$ (the coin can't land on its side).
- Below are three possible random variables based on the same random process (flipping a 2-sided coin 3 times):
- **Example 1**: $X =$ the number of heads in 3 coin flips
  - What are the possible values of X? 0, 1, 2, or 3.
- **Example 2**: Say you get 3 dollars for each Head
  - $Y =$ the amount of money won from 3 coin flips, $Y = 3 * X$

  - The possible values of Y are 0, 3, 6, or 9.

- **Example 3**: Z = the number of heads on the last of the 3 coin flips
  - The possible values are 0 or 1.

What might you want to know about these random variables? In general, we'd like to know the probability model (what values it takes and the associated chances), the expected value (long-run average), and the variance (a measure of how much the values vary). Let's talk about each of these next.

5.4 Probability Models

The **probability model** for a random variable X gives the possible values of X and the associated probabilities.

- What is the probability model for X: the number of heads in 3 coin flips?

$P(X = 0) = P(\text{three tails}) = 0.5^3$ (using Rule 5: independence)

$P(X = 1) = P(\text{HTT or THT or TTH }) = 3 * 0.5^3$ (using Rule 4: disjoint events & Rule 5)

$P(X = 2) = P(\text{HHT or HTH or THH }) = 3 * 0.5^3$ (using Rule 4 & 5)

$P(X = 3) = P(\text{three heads}) = 0.5^3$ (using Rule 5)

- What is the probability model for $Y = 3 * X$? (The total number of dollars earned when \$3 is paid for each head.)

$P(Y = 0) = P(\text{three tails}) = 0.5^3$

$P(Y = 3) = P(\text{HTT or THT or TTH }) = 3 * 0.5^3$

$P(Y = 6) = P(\text{HHT or HTH or THH }) = 3 * 0.5^3$

$P(Y = 9) = P(\text{three heads}) = 0.5^3$

- What about Z? (The number of heads on the last of the 3 coin flips)

$P(Z = 0) = P(\text{HHT or TTT or HTT or THT}) = 4 * 0.5^3 = 0.5$ (using Rules 4 & 5)

$P(Z = 1) = P(\text{HHH or TTH or HTH or THH}) = 4 * 0.5^3 = 0.5$ (using Rules 4 & 5)

For most situations in our class, we won't use the probability rules to calculate chances by hand. Rather, we will use a **named probability model** such that the we can calculate probabilities for particular values of the random variable using either:

- a **probability mass function (pmf)** (for a finite number of possible values) or

- a **probability density function (pdf)** (for a infinite number of possible values)

5.4.1 Using probability mass functions

Let's say that we are working with a random variable X that represents the result of spinning the arrow on a spinner that has 3 regions labeled 1, 2, and 3. X can only takes the values 1, 2, or 3.

The **probability mass function (pmf)** for X gives the probabilities that $X = 1$, $X = 2$, and $X = 3$ which are determined by the relative areas of the regions on the spinner. The pmf for X is frequently denoted as $p(x)$, which is shorthand for $P(X = x)$. Based on the construction of our spinner, the pmf can be denoted with a table like the one below:

x	1	2	3
$p(x)$	0.4	0.5	0.1

The first row indicates the values that X can take. The second row indicates the associated probabilities, the values of the probability mass function. Note that this row adds up to 1 because one of these 3 outcomes must happen.

We can use the pmf and the probability rules introduced earlier to calculate probabilities of different events.

Example 1: $P(X = 1 \text{ or } 2)$

Because the events $X = 1$ and $X = 2$ are disjoint (mutually exclusive/can't happen simultaneously), this probability is equal to $P(X = 1) + P(X = 2) = 0.4 + 0.5 = 0.9$. (Rule 4)

Example 2: $P(X \neq 1)$

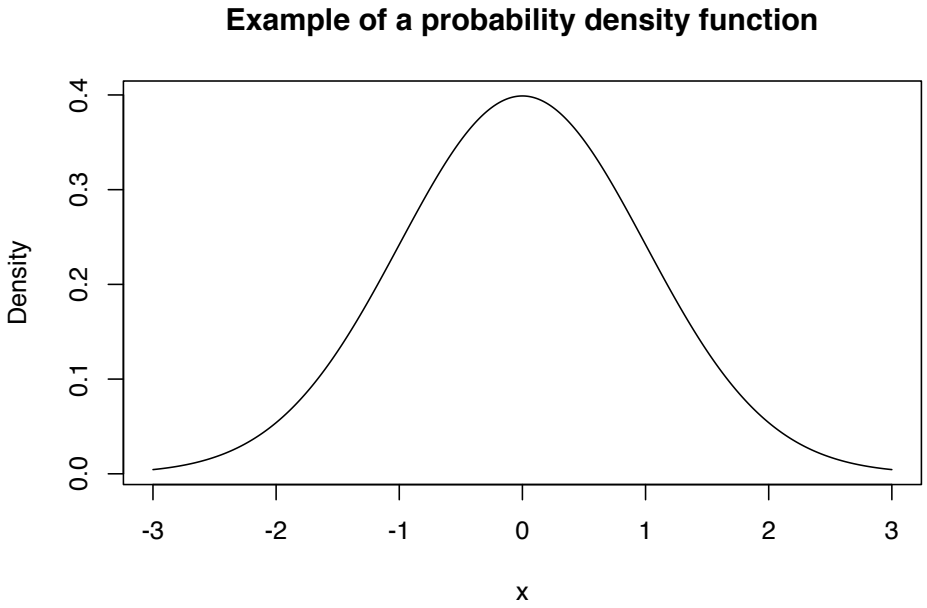
By Rule 3, $P(X \neq 1) = 1 - P(X = 1) = 1 - 0.4 = 0.6$. Another way to arrive at this would be to see that $P(X \neq 1) = P(X = 2 \text{ or } 3) = P(X = 2) + P(X = 3) = 0.5 + 0.1$.

5.4.2 Using probability density functions

When a random variable can take infinitely many values (i.e. a quantitative variable), using a probability mass function will not work because we would have to specify infinitely many probabilities. A **probability density function (pdf)** serves an analogous role to probability mass functions but works for quantitative variables that can take infinitely many values.

We have looked at density plots previously when we learned about data visualization. These were smooth curves that showed us the distribution of quantitative variable. In this class, the probability density functions that we will look at will

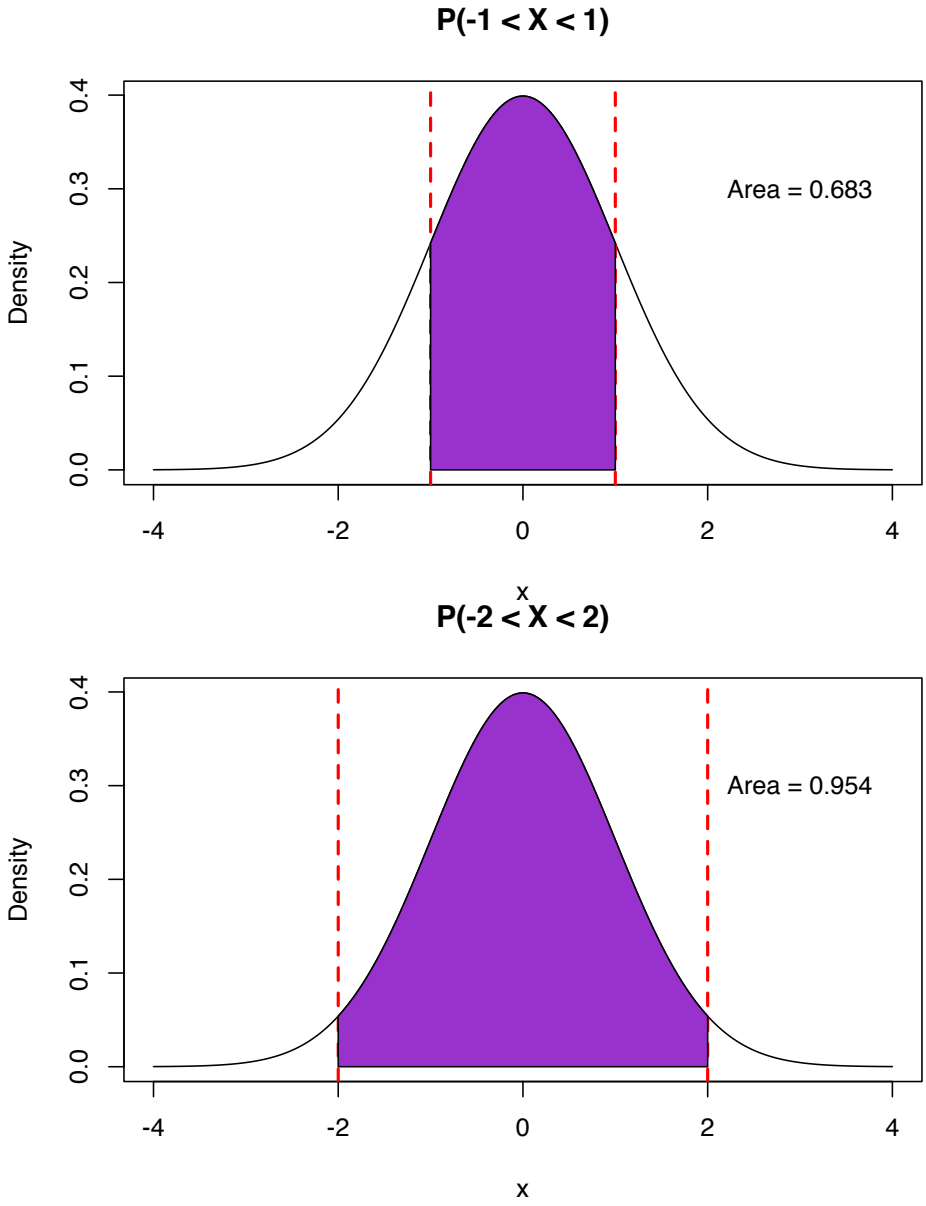
be smooth curves. One example of a pdf for a famous named probability model (called the Normal distribution) is shown below:

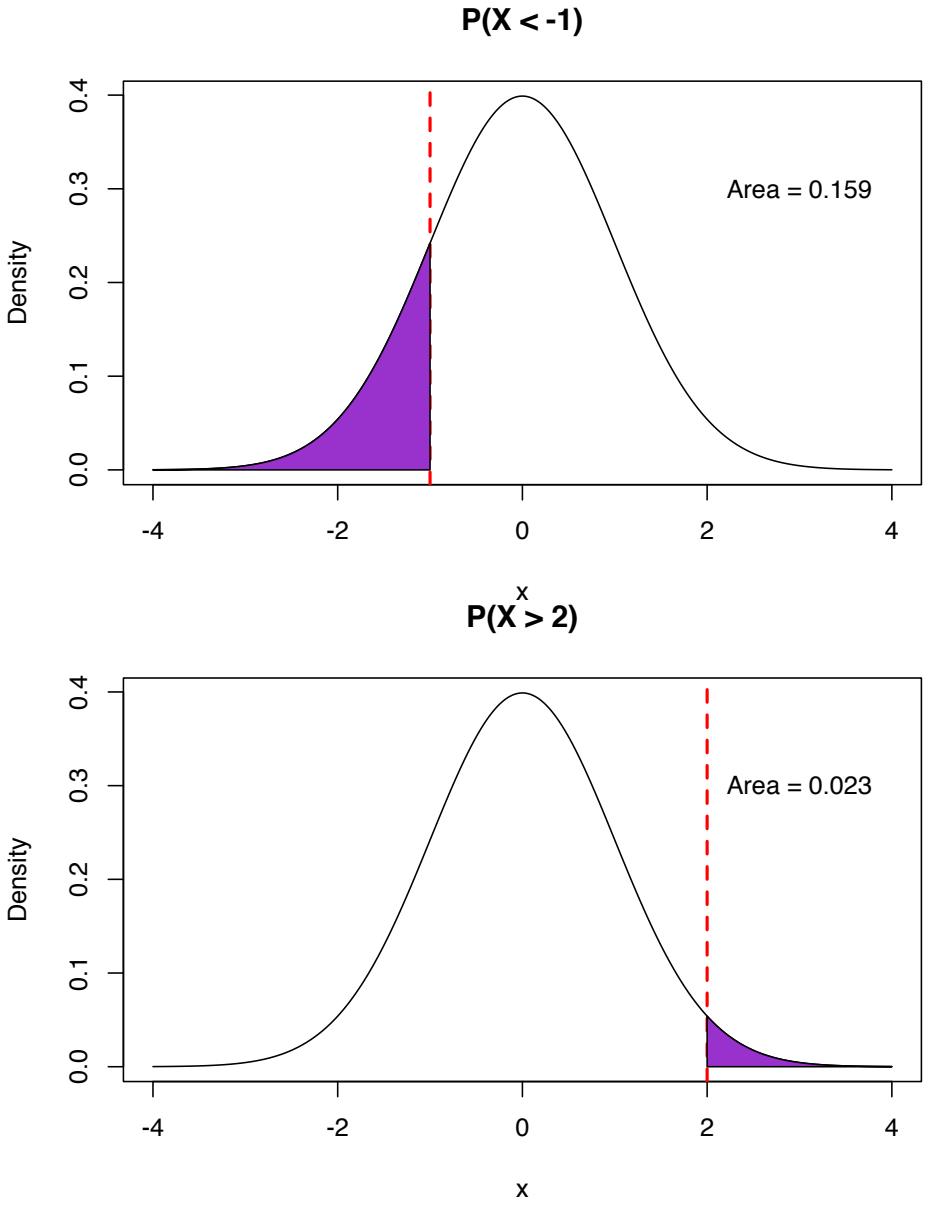


There are two main ideas that you should be comfortable with when working with quantitative random variables:

1. With a quantitative random variable X, we are interested in the probability that X falls in a certain range. Examples include $P(X > 3), P(X < -5), P(-1 < X < 1)$.
2. The calculation of such probabilities is achieved by looking at areas underneath the probability density function within that range ("the area under the curve").

The pictures below illustrate how different probability statements correspond to different areas beneath the pdf curve.


P(-1 < X < 1)
Area = 0.683
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
P(-2 < X < 2)
Area = 0.954
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x



5.4.3 Expected value and variance

Two important properties of a random variable X are its **expected value** and **variance**.

The **expected value** of a random variable is a real number and gives a measure of the typical value that the random variable takes or the long-run average if you could repeat the random process many times. For a mathematical definition,

you can look at the [probability appendix][Appendix] in these notes. Generally speaking, an expected value is a weighted average of the values that the random variable can take, weighted by the probability of getting that value. If, for example, X has a very high probability to take the value 9, then 9 will be weighed more heavily in the average.

Why do we care about expected value? Remember that the random variables that we will be concerned with are quantities such as estimated slopes from linear regression. This estimated slope is a (quantitative) random variable because its value depends on the random sample that we drew. Because it is a random variable, it has an associated probability density function which tells us how this estimated slope would vary across different samples of the same size. This distribution is called a **sampling distribution**, and expected value tells us the mean (center) of this distribution. When we take simple random samples from the population, the expected value is usually equal to the true target population value that we are trying to estimate from our sample. *This is a powerful idea because the statistic we compute from our one sample is, on average, a good estimate for the population parameter.*

The **variance** of a random variable is very much related to the variance of a set of numbers that was introduced as a measure of spread/variation/variability. Recall that variance as a measure of spread is approximately the average of the squared distances of each data point to the mean. The variance of a random variable follows in very much the same spirit. It gives the average squared distance of the random variable from its expected value. The variance of a random variable, like the variance of a set of numbers, measures how spread out the values of the random variable could be, but weighted according to the probability of the value. The standard deviation of a random variable is the square root of the variance.

Why do we care about the variance of a random variable? The spread of the **sampling distribution** is measured by the variance of the associated random variable. If the sampling distribution is very spread out, then our estimates could differ quite a bit from sample to sample. This means that there is a lot of uncertainty in our estimate. For example, we may have estimated a slope of 3 from our linear regression model, but if we have high variance, that slope could easily have been 7 or -2 depending on the particular sample we draw. *The variance of an estimated regression slope quantifies how much our estimate could vary from sample to sample, which will allow us to give reasonable margins of error on our estimates.*

Wait! These ideas are fine in theory... but how do we actually use them?

It turns out that many random variables that we've worked with so far are approximated well by **named probability distributions** that allow for theoretical calculations of expected value and variance. This theory means that we don't have to go out and collect random sample after sample to get a sense of sampling variability. Statistical theory gives us a way to understand/approximate

how our estimates vary from sample to sample using mathematics. This process is based on **theoretical probability** rather than on bootstrapping which was giving us the same type of information but through **empirical probability** that we calculated using computational tools.

5.5 Bernoulli/Binomial Model

There are situations in which your response of interest only has two possible outcomes: success or failure, heart disease or no heart disease, o-ring failure or no o-ring failure, etc. In the past, we used logistic regression to create a model to predict a binary response variable based on explanatory variables. For a moment, let's just consider the response itself as a random process.

If a random process satisfies the following three conditions, then we can use the **Bernoulli Model** to understand its long-run behavior:

1. Two possible outcomes (success or failure)
2. Independent "trials" (the outcome for one unit does not depend on the outcome for any other unit)
3. P(success) $= p$ is constant (the relative frequency of success in the population that we are drawing from is constant)

If a random variable X follows a Bernoulli model, X denotes the number of successes (0 or 1) in the single trial that is conducted. The probability mass function is $P(X = 1) = p$ and $P(X = 0) = 1 - p$. In the long run, the expected number of successes will be p, the variance will be $p(1 - p)$, and the standard deviation will be $\sqrt{p(1-p)}$. If you have looked through the [Appendix], you can show this mathematically using definitions.

Let's think back to the disease testing example. Say we have a very large population where 1 in every 1000 has a particular disease ($p = 1/1000$). We could model the disease outcome from randomly drawing an individual from the population using a Bernoulli Model where $P(X = 1) = 0.001$. If we just randomly drew 1 person, we'd expect 0.001 of a person to have the disease and the spread in outcomes is $\sqrt{0.001 * 0.999} = 0.031$. These values don't make a lot of sense because we often sample more than one person.

Imagine we had a sample of n individuals from the population. Then we are considering n independent Bernoulli "trials". If we let the random variable X be the count of the number of successes in a sample of size n, then we can use the **Binomial Model.** In this case we say that "X follows a binomial distribution".

The probability mass function for the Binomial Model is

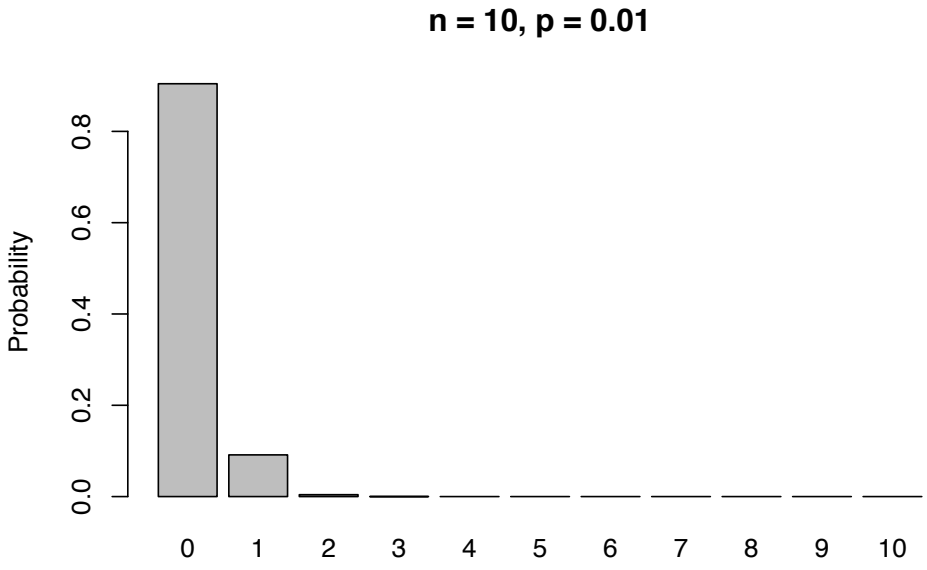
$$p(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

where x is the number of successes in n trials. In the long run, the expected

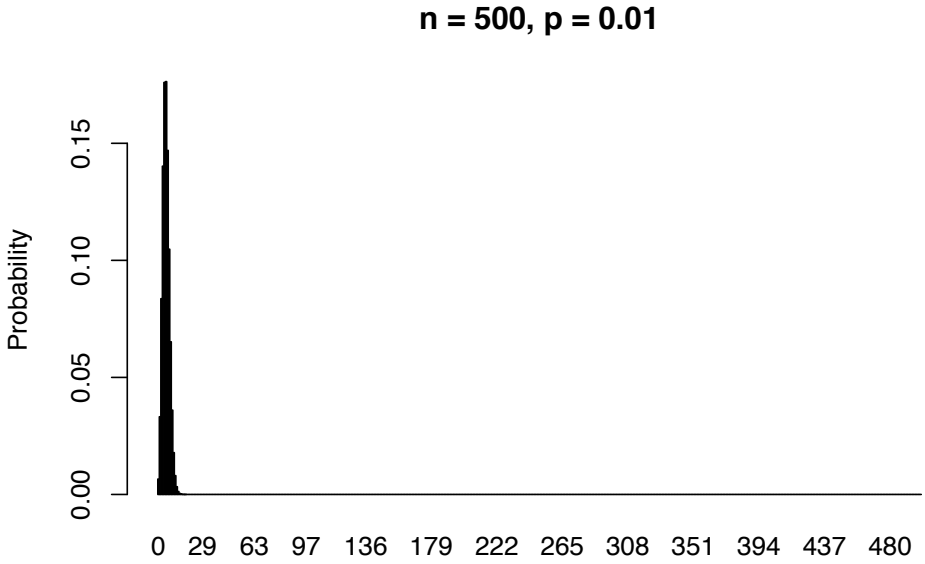
number of successes will be np, the variance will be $np(1-p)$, and the standard deviation will be $\sqrt{np(1-p)}$. In the long run, the expected relative frequency of successes in n trials will be p, the variance will be $\frac{p(1-p)}{n}$ and the standard deviation will be $\sqrt{\frac{p(1-p)}{n}}$. If you are working through the [Appendix], you can show these mathematically using known properties.

If we randomly draw 5000 people, we'd expect $0.001 * 5000 = 5$ people to have the disease and the spread in the count is $\sqrt{5000 * 0.001 * 0.999} = 2.23$. Let's adjust this to relative frequencies. We'd expect 0.1% ($5/5000 = 0.001$) of the people to have the disease and a measure of the variability in relative frequency is $\sqrt{0.001 * 0.999/5000} = 0.0004$.

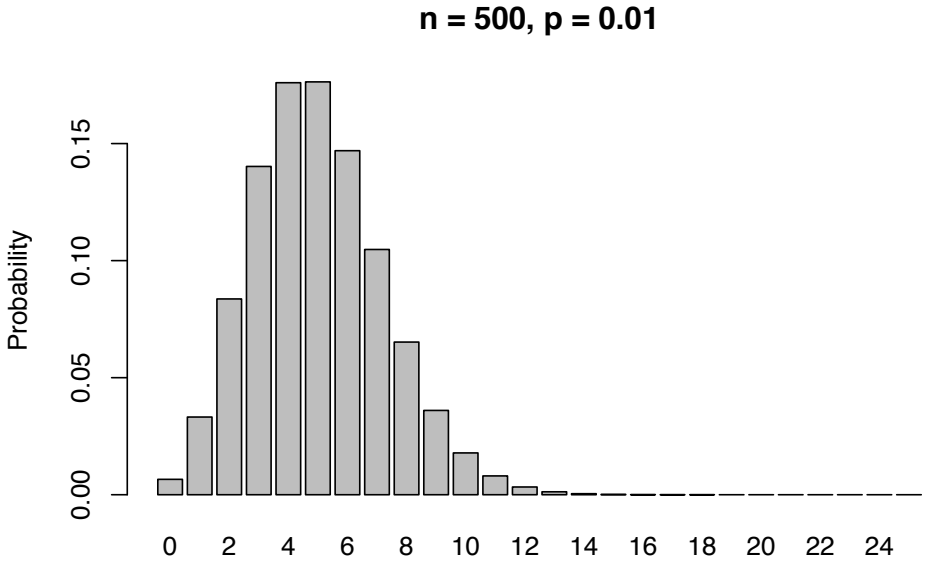
Let's look at the probability mass function for the Binomial Model for $n = 10, p = 0.01$.



If we increase the sample size to $n = 500$, then the probability mass function looks like this.



Let's zoom in on the left hand side of this plot.



What does this look like? A Normal Model!!!

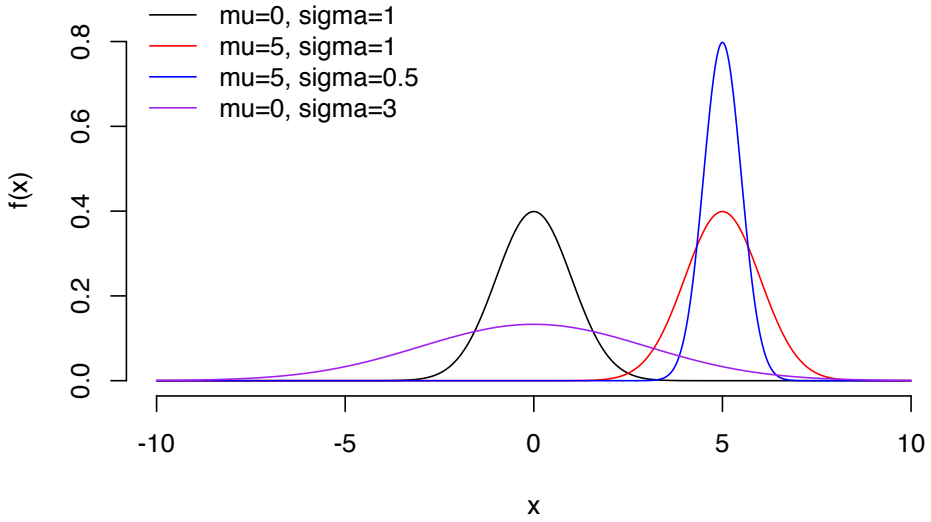
5.6 Normal Model

We've been introduced to the Normal model already as a smooth version of a unimodal, symmetric histogram. For a quantitative random variable X (whose value can be any real number), if the expected value is μ and the variance is σ^2,

a Normal random variable has a probability density function of

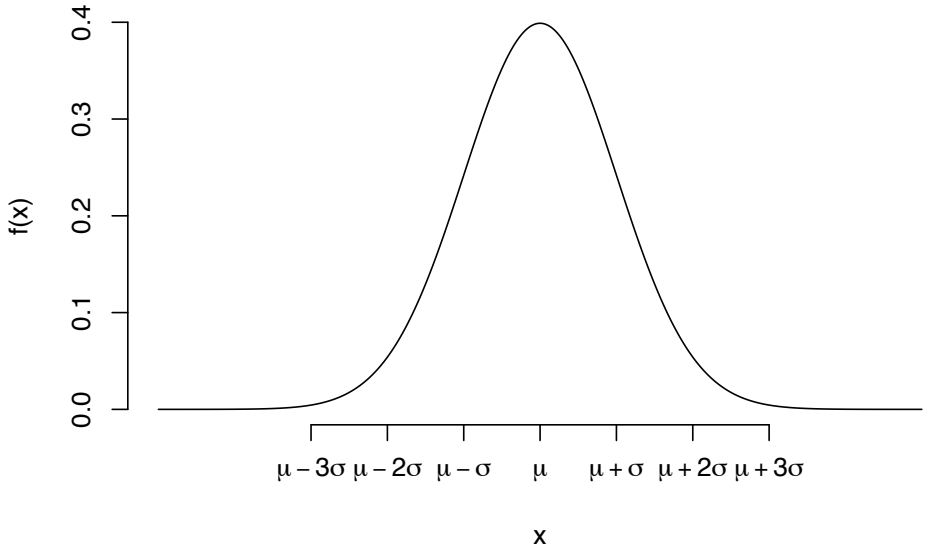
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For every potential value of μ and σ, there is a different function/curve. Some examples are shown below.

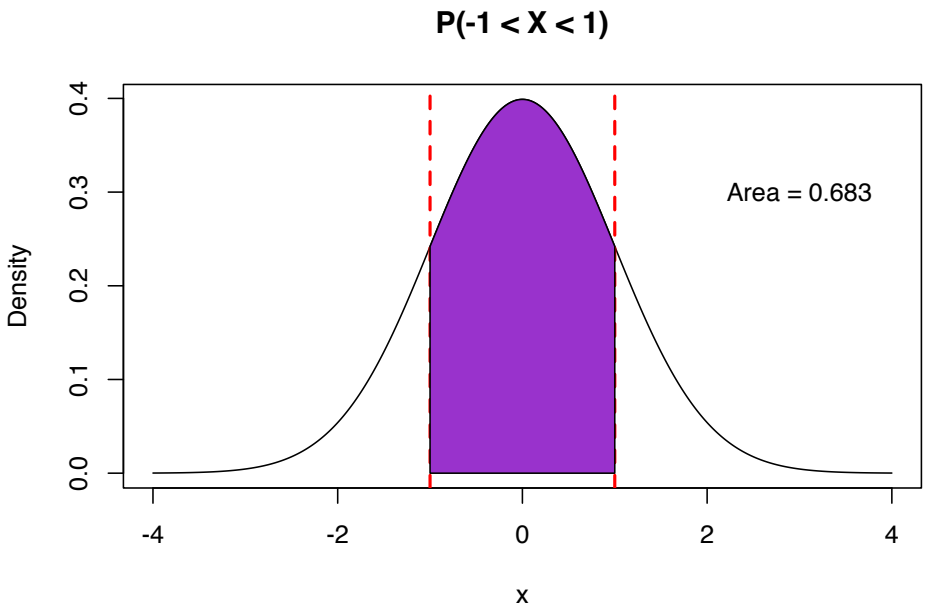


If a random variable X is modeled with a Normal model, we also say that "X follows a normal distribution" or that "X is normally-distributed".

- In general, the center of the distribution is μ and the standard devation σ, the square root of the variance, determines the spread of the distribution.



- Let's consider the particular Normal model with $\mu = 0$ and $\sigma = 1$. This is called the **standard normal distribution**. We know that $P(-1 \leq X \leq 1) = 0.68$, which is calculated as the area under the curve between -1 and 1.

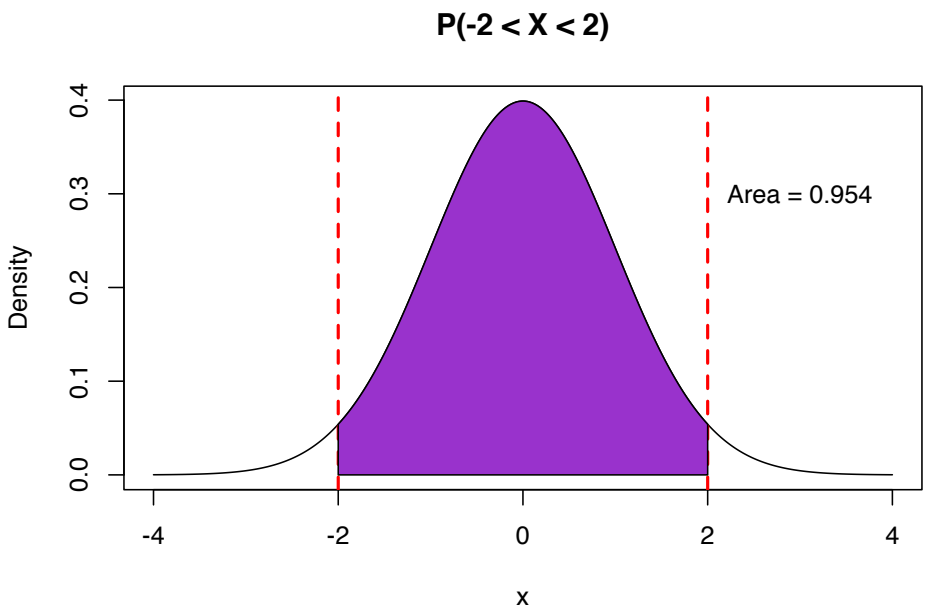



```
pnorm(1) - pnorm(-1)
```

```
## [1] 0.683
```

```
#pnorm(1) gives the area under the curve to the left of 1
#pnorm(-1) gives the area under the curve to the left of -1
```

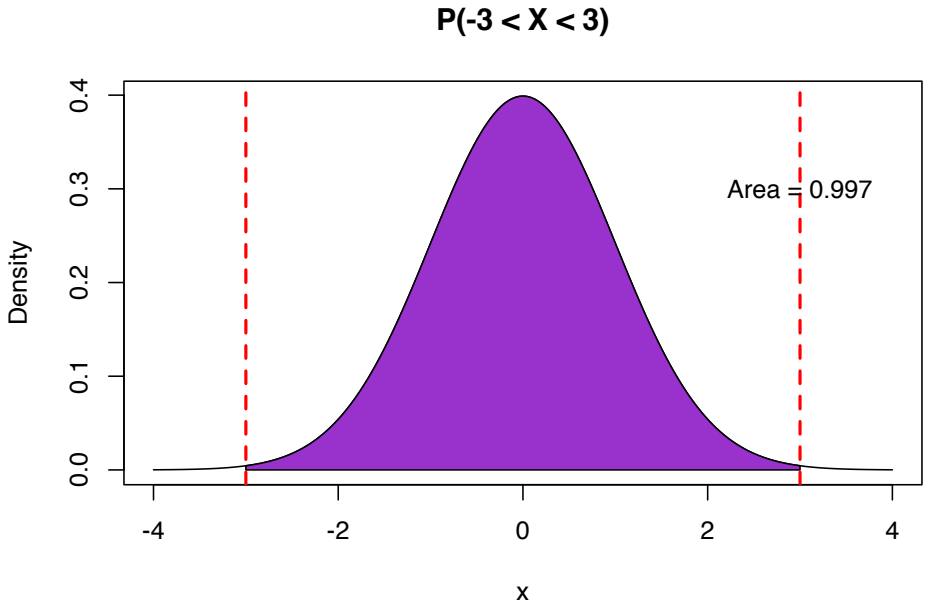
- We know that $P(-2 \leq X \leq 2) = 0.95$, calculated as the area under the curve between -2 and 2.




```
pnorm(2) - pnorm(-2)
```

```
## [1] 0.954
```

- We know that $P(-3 \leq X \leq 3) = 0.997$, calculated as the area under the curve between -3 and 3.



```
pnorm(3) - pnorm(-3)
```

```
## [1] 0.997
```

The standard normal distribution is very convenient to work with. No matter what the long-run average μ and standard deviation σ are for a normally-distributed random variable X, we can standardize the values to obtain **z-scores** by subtracting μ and dividing by σ:

$$\text{z-score} = \frac{X - \mu}{\sigma}$$

We typically denote z-scores with Z. It turns out that Z follows a standard normal distribution. That is $\mu = 0, \sigma = 1$ for Z. This allows us to focus solely on the areas for the standard normal distribution rather than the particular normal distribution with mean μ and standard deviation σ.

Important: If a random variable can be modeled with a Normal model, then we know that:

- About 68% of the time, the values will be within 1 standard deviation of the expected value.
- About 95% of the time, the values will be within 2 standard deviations of the expected value.
- About 99.7% of the time, the values will be within 3 standard deviations of the expected value.

We will call this the **68-95-99.7 rule**.

5.7 Sampling Distribution and CLT

5.7.1 Sampling distributions

As we wade through the formal theory, let's remind ourselves why we need to understand randomness and the tools of formal probability.

We appreciate that our estimates will vary from sample to sample because we have different units in each sample. We care about *how much* they can vary because we want to *quantify* the amount of uncertainty in our estimates. Sure, we may use linear regression modeling to find that the estimate for an interaction coefficient is \$26.85, but how different could that estimate have been if we had obtained a different sample (of the same size)?

The **sampling ditribution** of a statistic (e.g. sampling distribution of the sample mean, of the sample median, of a regression coefficient) tells us exactly how that statistic would vary across all possible samples of a given size. For example, in our New York housing data (1728 homes), the sampling distribution for the interaction coefficient tells us exactly the distribution of interaction

coefficient estimates that result from all possible samples of size 1728 *from the population of New York homes*. This distribution is centered at the true value of the population interaction coefficient (the value we would get from linear modeling if we did indeed use the full population). The spread of the distribution gives us a measure of the precision of our estimate. If the sampling distribution is very wide, then our estimate is imprecise; our estimate would vary widely from sample to sample. If the sampling distribution is very narrow, then our estimate is precise; our estimate would not vary much from sample to sample.

The emphasis on "from the population of New York homes" when defining the sampling distribution earlier is deliberate. We don't have the population, and we never will (usually)! The sampling distribution is a theoretical idea, but we wish to know what it looks like because we care about the precision of our estimate. We have done our best to estimate the center of the sampling distribution: our best guess is the sample estimate itself. For our housing data, we estimate an interaction coefficient of \$26.85. This is our best guess for the center of the sampling distribution. What about the spread?

We have looked at one way to estimate the spread of the sampling distribution: bootstrapping! In bootstrapping, we pretend that our sample is the population and look at lots of different samples from this "fake population". In bootstrapping, we have essentially tried to replicate the idea of getting "all possible samples of a given size from the population".

5.7.2 The Central Limit Theorem

Another powerful way to understand what the sampling distribution looks like is called the **Central Limit Theorem**. This is a famous theorem in Statistics that can be stated as follows.

Central Limit Theorem (CLT): Let $X_1, X_2, \ldots, X_n$ be random variables that denote some value measured on each of n units in a random sample. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ denote the sample mean of those values. Let all of the random variables $X_1, \ldots, X_n$ have the same expected value μ and variance σ^2. Then as the sample size n get bigger and bigger (grows to infinity), the sample mean $\bar{X}_n$ becomes normally-distributed with mean μ (the true expected value) and variance σ^2/n.

What does the CLT tell us? It tells us that with a "sufficiently large" sample size, the sample mean is normally-distributed with a particular mean and variance. This tells us exactly what the sampling distribution is! The center and shape of the sampling distribution are given by a named probability model, the Normal distribution! Let's focus on the implications of these two parameters:

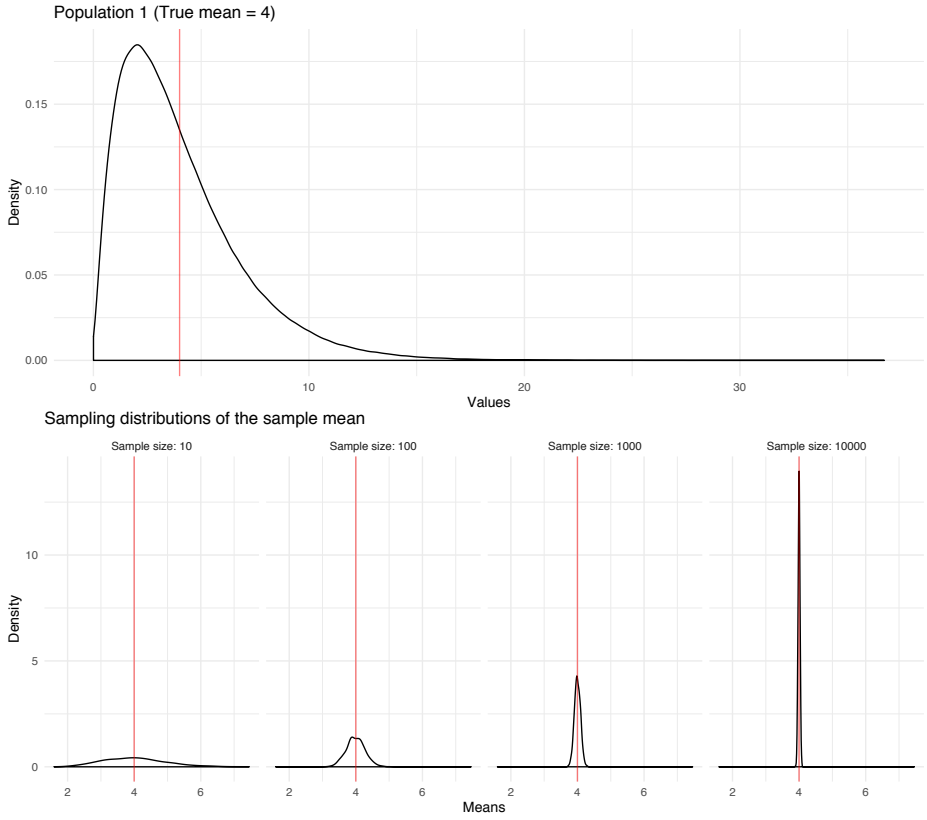
- The mean: The CLT says that for large samples the sample mean is normally distributed with mean μ. μ is the actual expected value of the random variables! This tells us that the sample mean estimates the population mean correctly, on average. For example, let's say that the

X's are house prices, and we want to know the true population mean of house prices. The CLT tells us that, **on average**, the sample mean will be correct! Another way to say this is that the sample estimates are "unbiased": underestimates are balanced out by overestimates.

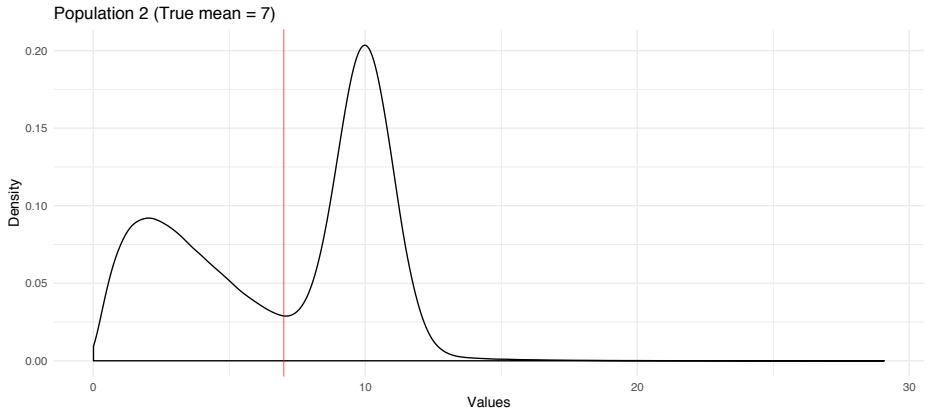
- The variance: The CLT says that for large samples the sample mean is normally distributed with variance σ^2/n. Equivalently, the standard deviation is $\sigma/\sqrt{n}$. This tells us the spread of the sampling distribution! Sample estimates are more likely to be "close to" than "far from" the population parameter μ. We see that the sample size n plays a key role. The larger the sample size, the smaller the spread. This matches our intuition that larger samples should generate more **precise** estimates. In practice, we estimate σ and refer to this estimated standard deviation as the **standard error**.

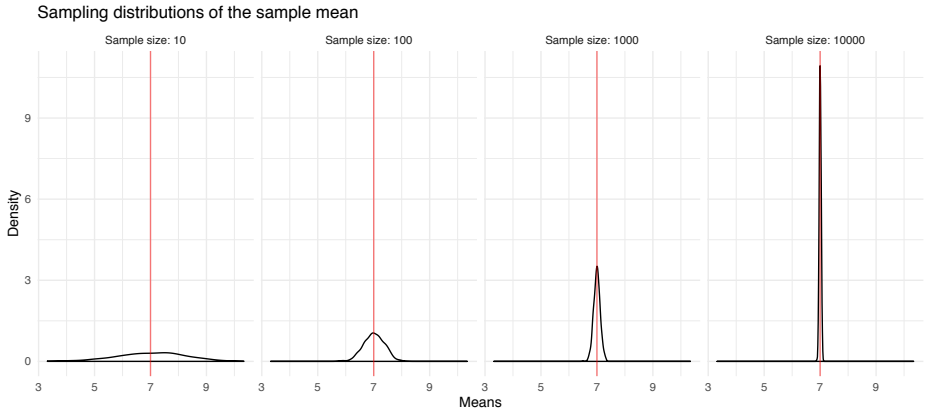
Where does the CLT come up in practice? When we fit linear regression and logistic regression models, there were columns in the summarized output labeled "Std. Error". We have neglected to look at these columns... until NOW. With our understanding of the CLT, we can understand where this column comes from. It turns out that we can apply the CLT to see that our estimated regression coefficients (our betas) are normally-distributed. That is, the theory of the CLT allows us to approximate the sampling distributions of regression coefficients. Behind the scenes, R uses that theory to give us the Standard Error column. This column gives us an estimate of the standard deviation of the sampling distribution for that regression coefficient.

Let's visualize what is happening with the Central Limit Theorem. Below we have simulated (generated ourselves) data for a full population of cases. The distribution of values in the full population is shown below. The true population mean (expected value) is 4. Beneath, we show the sampling distributions of the sample mean for different sample sizes. We see the CLT in action. The center of the sampling distributions is the expected value 4. The sampling distributions look normally distributed, and the spread decreases with sample size. Note that the original population values do not look normally-distributed, and yet, by the CLT, the sample means are normally-distributed.

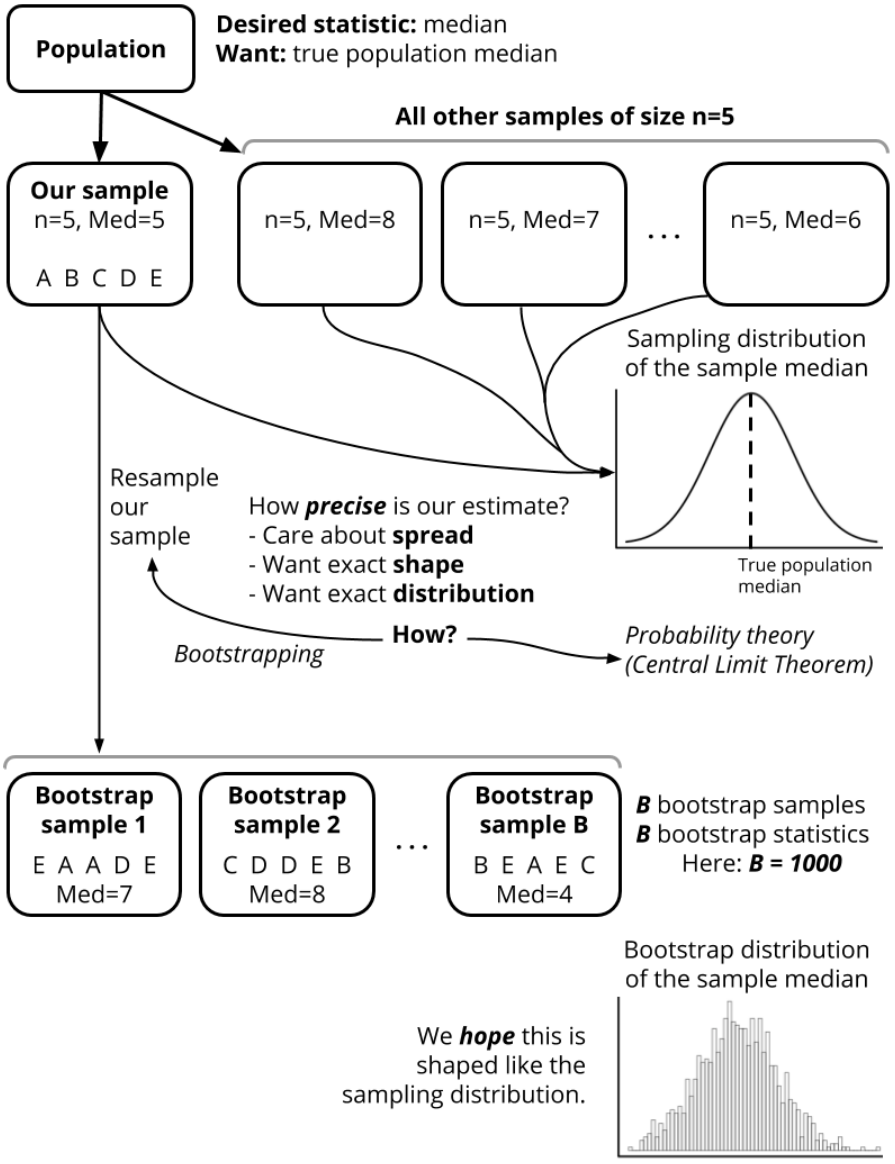


Let's look at one more example. Again, the population distribution is clearly not normally-distributed, and yet, the sample means are normally-distributed. These normal distributions are centered at the true population mean (expected value) and have smaller and smaller spread with increased sample size.





The different ways in which we think about sampling variability in Statistics are shown in the figure below.



5.8 Z-scores and the Student's "t" distribution

Let's stop for a short story time. Back in the early 1900's, William S. Gosset was working as a quality control chemist and mathematician for Guinness Brewery in Ireland.

![Source: Wikipedia(gosset.jpg)

His research goal was to determine how to get the best yielding barley and hops through agricultural experimentation, but the main issue that he had was that he had very small sample sizes (only a handful of farms)!

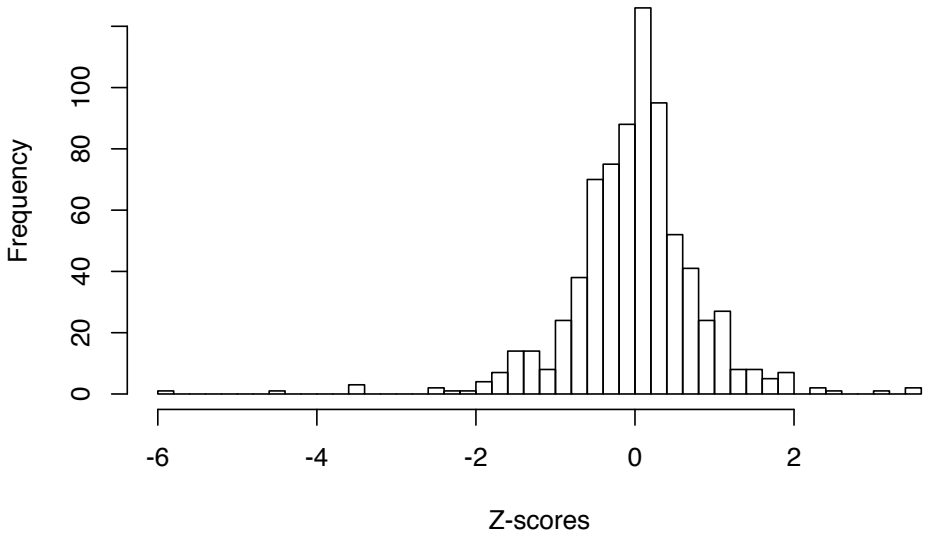
He was used the Normal distribution because the work underlying the Central Limit Theorem had already been worked out, but he noticing that there were discrepancies in his work.

5.8.1 Gosset's Work

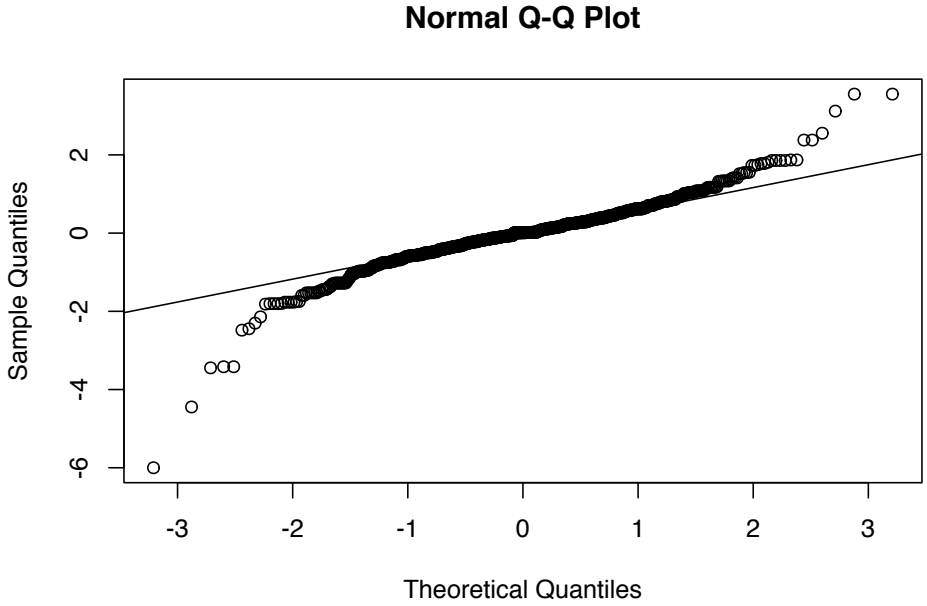
To further investigate the issues that he was noticing, Gosset performed a simulation using data on the height and left middle finger measurements of 3000 criminals (we don't know how he got access to this data...). His simulations went as follows:

- Split the data (3000 criminals) into samples of size 4 ($n = 4$)
- Estimate the mean and standard deviation of the height within each of the 750 samples
- Look at the distribution of the z score heights, using the mean of the 3000 as the true population mean, μ, but using s in the denominator from each sample instead of the population variation, σ, of the 3000:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{4}}$$



- Does this distribution of z-scores look Normal? We can try to see more clearly by making a **Q-Q plot**, shown below.



If the z-scores follow a normal distribution, the points in the Q-Q plot should fall on a straight line. However, we see some deviation at the left and right of the plots. This indicates that the tails of our z-score distribution are thicker (more populous) than expected if they were actually normally-distributed.

Why? With small samples, the sample standard deviation, s, is smaller so we tend to get larger z-scores than we'd expect if they were normally-distributed.

5.8.2 Beer Helps the Field of Statistics

For the sake of making better beer, William Gosset worked out the sampling distribution of a standardized sample mean (with the *sample* standard deviation, s, plugged into the denominator),

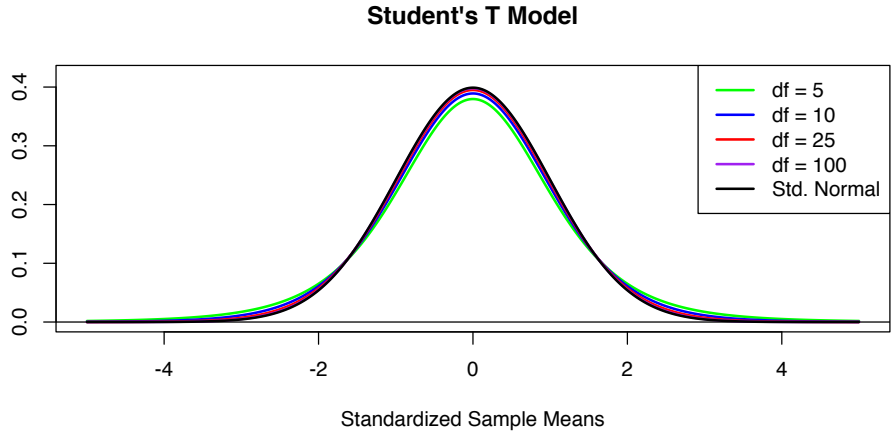
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

if the histogram of the **population is unimodal and roughly symmetric** (approximately Normal). This was crucial when **sample sizes are small.** He had to publish under the pseudonym **Student** (trade secrets, etc.). Thus, the sampling distribution model that he discovered became known as **Student's model.** Later R.A. Fisher (another famous statistician) recoined it as the **Student's T model.**

5.8.3 Student's T Model

- The center is always 0.

- The spread is determined by a parameter called **degrees of freedom** (for standardized means, $df = n - 1$).
- As $n \to \infty$, this model looks more like the Normal model.



Since the Student T distribution is approximately Normal when sample sizes are large, we will typically use the Normal model. However, if sample sizes are small ($n < 30$ or so), the Normal model is not appropriate. So, in those rare cases, we'll need to refer to Student's T.

Chapter 6

Statistical Inference

Let's remember our goal of "turning data into information." Based on a sample data set, we want to be able to say something about the larger population of interest. This endeavor is called **statistical inference**. In statistical inference, we care about using sample data to make statements about "truths" in the larger population.

- To make causal inferences in the sample, we need to account for all possible confounding variables, or we need to randomize the "treatment" and assure there are no other possible reasons for an observed effect.
- To generalize to a larger population, we need the sample to be representative of the larger population. Ideally, that sample would be randomly drawn from the population. If we actually have a census in that we have data on country, state, or county-level, then we can consider the observed data as a "snapshot in time". There are random processes that govern how things behave over time, and we have just observed one period in time.

Let's do some statistical inference based on a simple random sample (SRS) of 100 flights leaving NYC in 2013.

> **Reflect:** What is our population of interest? What population could we generalize to?

```
set.seed(2018)
## This creates a dataset called flights_samp
## that contains the SRS of size 100
flights_samp <- flights %>%
    sample_n(size = 100)
```

We've already been thinking about random variation and how that plays a role in the conclusions we can draw. In this chapter, we will formalize two

techniques that we use to do perform statistican inference: confidence intervals and hypothesis tests.

6.1 Confidence Intervals

A **confidence interval** (also known as an interval estimate) is an interval of plausible values of the unknown population parameter of interest based on randomly sampled data. However, the interval computed from a particular sample does not necessarily include the true value of the parameter. Since the observed data are random samples from the population, the confidence interval obtained from the data is also random.

The **confidence level** represents the proportion of possible random samples and thus confidence intervals that contain the true value of the unknown population parameter. Typically, the confidence level is represented by $(1 - \alpha)$ such that if $\alpha = 0.05$, then the confidence level is 95% or 0.95. What is α? We will define α when we get to hypothesis testing, but for now, we will describe α as an error probability. Because we want an error probability to be low, it makes sense that the confidence level is $(1 - \alpha)$.

Valid Interpretation: Assuming the sampling distribution model is accurate, I am 95% confident that my confidence interval of (lower, upper) contains the true population parameter (*put in context*), which means that we'd expect 95% of samples to lead to intervals that contain the true population parameter value. We just don't know if our particular interval from our study contains that true population parameter value or not.

6.1.1 Via Classical Theory

If we can use theoretical probability to approximate the sampling distribution, then we can create a confidence interval by taking taking our estimate and adding and subtracting a margin of error:

$$\text{Estimate } \pm \text{ Margin of Error}$$

The margin of error is typically constructed using z-scores from the sampling distribution (such as $z^* = 2$ that corresponds to a 95% confidence interval) and an estimate of the standard deviation of the estimate, called a **standard error**.

Once we have an estimate of the standard deviation (through a formula or R output) and an approximate sampling distribution, we can create the interval estimate:

$$\text{Estimate } \pm z^* * SE(\text{Estimate})$$

The fact that confidence intervals can be created as above is rooted in the Central Limit Theorem (CLT). If you would like to see how the form above is derived, see the Math Box below.

Math Box: (Optional) Deriving confidence intervals from the CLT
The CLT originally expresses that

$$\frac{\text{sample mean} - \text{true mean}}{\text{true std. error of sample mean}} \sim \text{Normal}(0, 1)$$

It turns out that the CLT also applies to regression coefficients:

$$\frac{\hat{\beta} - \beta}{\text{ESTIMATED std. error of } \hat{\beta}} \sim \text{Normal}(0, 1)$$

From there we can write a probability statement using the 68-95-99.7 rule of the normal distribution and rearrange the expression using algebra:

$$P(-2 < \frac{\hat{\beta} - \beta}{\text{ESTIMATED std. error of } \hat{\beta}} < 2) = 0.95$$

$$P(-2SE < \hat{\beta} - \beta < 2SE) = 0.95$$

$$P(-2SE - \hat{\beta} < -\beta < 2SE - \hat{\beta}) = 0.95$$

$$P(2SE + \hat{\beta} > \beta > -2SE + \hat{\beta}) = 0.95$$

You've seen the Student t distribution introduced in the previous chapter. We used the Normal distribution in this derivation, but it turns out that the Student t distribution is more accurate for linear regression coefficients (especially if sample size is small). The normal distribution is appropriate for logistic regression coefficients.

6.1.2 Via Bootstrapping

In order to gauge the sampling variability, we can treat our sample as our "fake population" and generate repeated samples from this "population" using the technique of bootstrapping.

Once we have a distribution of sample statistics based on the generated data sets, we'll create a confidence interval by finding the $\alpha/2$th percentile and the $(1 - \alpha/2)$th percentile for our lower and upper bounds. For example, for a 99% bootstrap confidence interval, $\alpha = 0.01$ and you would find the values that are

the 0.5th and 99.5th percentiles.

6.2 Confidence Interval Examples

6.2.1 Proportion Outcome

Let's return to the flight data and estimate the proportion of afternoon flights based on a sample of 100 flights from NYC.

First, the classical 95% confidence interval can be constructed using the theory of the Binomial Model (Do the 3 conditions hold? Is n large enough for it to look Normal?)

```
flights_samp %>%
  summarize(prop = count(day_hour)/100) %>%
  mutate(SE = sqrt(prop*(1-prop)/100)) %>%
  mutate(lb = prop - 2*SE, ub = prop + 2*SE)
```

```
## # A tibble: 1 x 4
##    prop     SE    lb    ub
##   <dbl>  <dbl> <dbl> <dbl>
## 1  0.53 0.0499 0.430 0.630
```

Or we could bootstrap and get our confidence interval that way.

```
alpha <- 0.05

boot_data <- mosaic::do(1000)*(
    flights_samp %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      summarize(prop = count(day_hour)/100) # Calculate statistics
)

boot_data %>%
  summarize(lower = quantile(prop, alpha/2),
    upper = quantile(prop, 1-alpha/2))
```

```
##   lower upper
## 1  0.44  0.63
```

Our confidence interval gives a sense of the true proportion of flights departed NYC in the afternoon, keeping in mind that this sample could be one of the unlucky samples (the 5%) that have intervals that don't contain the true value.

6.2.2 Mean and then Median

Perhaps you really care about the arrival delay time because you have somewhere important you need to be when you take flights out of NYC. Let's estimate the mean arrival delay based on a sample of 100 flights from NYC.

First off, let's create a classical confidence interval. Since our sample size is relatively large, we can use the Normal model (instead of William Gosset's work). We use the sample standard deviation and plug into the SE formula.

```
flights_samp %>%
  summarize(mean = mean(arr_delay), s = sd(arr_delay)) %>%
  mutate(SE = s/sqrt(100)) %>%
  mutate(lb = mean - 2*SE, ub = mean + 2*SE)
```

```
## # A tibble: 1 x 5
##    mean     s    SE     lb    ub
##   <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1  6.36  33.9  3.39 -0.422  13.1
```

```
alpha <- 0.05

boot_data <- mosaic::do(1000)*(
    flights_samp %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      summarize(means = mean(arr_delay),medians = median(arr_delay)) # Calculate statistics
)

boot_data %>%
  summarize(lower = quantile(means, alpha/2),
    upper = quantile(means, 1-alpha/2))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 0.230  13.4
```

Our confidence interval gives potential values for of the true mean arrival delay for flights that departed NYC, keeping in mind that this sample could be one of the unlucky samples (the 5%) that have intervals that don't contain the true value. Also remember that the mean is sensitive to outliers. . . Let's consider the median.

We get a slightly different story if we are interested in the middle number versus the average. But notice, we aren't using a classical CI here because the sampling distribution of a median is not necessarily Normal.

```
boot_data %>%
  summarize(lower = quantile(medians, alpha/2),
    upper = quantile(medians, 1-alpha/2))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1  -8.5   5.5
```

6.2.3 Logistic Regression Model

Are the same relative numbers of morning flights in the winter and the summer? Let's fit a logistic regression model and see what our sample says.

```
flights_samp$afternoon = flights_samp$day_hour == 'afternoon'

glm.afternoon <- glm(afternoon ~ season, data = flights_samp, family = 'binomial')
summary(glm.afternoon)
```

```
##
## Call:
## glm(formula = afternoon ~ season, family = "binomial", data = flights_samp)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.33   -1.13    1.03    1.03    1.23
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.123      0.286   -0.43     0.67
## seasonwinter    0.479      0.404    1.19     0.24
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.27  on 99  degrees of freedom
## Residual deviance: 136.85  on 98  degrees of freedom
## AIC: 140.8
##
## Number of Fisher Scoring iterations: 4
```

The output for the model gives standard errors for the slopes, so we can create the classical confidence intervals directly from the output,

```
confint(glm.afternoon)
```

```
##                2.5 % 97.5 %
## (Intercept)  -0.691  0.439
## seasonwinter -0.308  1.279
```

Or with bootstrapping,

```
boot_data <- mosaic::do(1000)*(
    flights_samp %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      glm(afternoon ~ season, data = ., family = 'binomial') # Calculate statistics
)

boot_data %>%
```

```
  summarize(lower = quantile(seasonwinter, alpha/2),
    upper = quantile(seasonwinter, 1-alpha/2))
```

```
##   lower upper
## 1 -0.35  1.37
```

Knowing that the sample is random, the interval estimate for the logistic regression slope is given by the confidence interval. But for logistic regression, we exponentiate the slopes to get an more interpretable value, the odds ratio. Here, we are comparing the odds of having a flight in the afternoon between winter months (numerator) and summer months (denominator). Is 1 in the interval? If so, what does that tell you?

```
exp(confint(glm.afternoon))
```

```
##              2.5 % 97.5 %
## (Intercept)  0.501   1.55
## seasonwinter 0.735   3.59
```

```
boot_data %>%
  summarize(lower = exp(quantile(seasonwinter, alpha/2)),
    upper = exp(quantile(seasonwinter, 1-alpha/2)))
```

```
##   lower upper
## 1 0.704  3.92
```

6.2.4 Linear Regression Model Slope (Categorical Variable)

Everything says that there are longer delays in winter. Is that actually true? Let's fit a linear regression model to test it

```
lm.delay <- lm(arr_delay ~ season, data = flights_samp)
summary(lm.delay)
```

```
##
## Call:
## lm(formula = arr_delay ~ season, data = flights_samp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.69 -21.69  -8.69  11.55 166.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.694      4.869    1.37     0.17
## seasonwinter   -0.655      6.818   -0.10     0.92
##
```

```
## Residual standard error: 34.1 on 98 degrees of freedom
## Multiple R-squared:  9.41e-05,   Adjusted R-squared:  -0.0101
## F-statistic: 0.00922 on 1 and 98 DF,  p-value: 0.924
```

The classical CI for the slope is given with by

```
confint(lm.delay)
```

```
##                2.5 % 97.5 %
## (Intercept)   -2.97   16.4
## seasonwinter -14.18   12.9
```

or with bootstrapping,

```
boot_data <- mosaic::do(1000)*(
    flights_samp %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      lm(arr_delay ~ season, data = .) # Calculate statistics
)

boot_data %>%
  summarize(lower = quantile(seasonwinter, alpha/2),
    upper = quantile(seasonwinter, 1-alpha/2))
```

```
##   lower upper
## 1 -15.1  11.8
```

The 95% confidence interval gives a sense of the difference of mean arrival delays of flights between winter and summer is given by the confidence interval. Is zero in the interval? If so, what does that tell you?

6.2.5 Linear Regression Model Slope (Quantitative Var)

How well can the departure delay predict the arrival delay? What is the effect of departing 1 more minute later? Does that correspond to 1 minute later in arrival on average? Let's look at the estimated slope between departure and arrival delays for the sample of 100 flights from NYC.

```
lm.delay2 <- lm(arr_delay ~ dep_delay, data = flights_samp)
summary(lm.delay2)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = flights_samp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27.40 -12.40  -2.90   9.59  60.53
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.5874     1.8552   -1.39     0.17
## dep_delay     1.0042     0.0617   16.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.7 on 98 degrees of freedom
## Multiple R-squared:  0.73,   Adjusted R-squared:  0.727
## F-statistic:  265 on 1 and 98 DF,  p-value: <2e-16
```

The classical CI for the slope is given with by

```
confint(lm.delay2)
```

```
##              2.5 % 97.5 %
## (Intercept) -6.269   1.09
## dep_delay    0.882   1.13
```

or with bootstrapping,

```
boot_data <- mosaic::do(1000)*(
    flights_samp %>% # Start with the SAMPLE (not the FULL POPULATION)
      sample_frac(replace = TRUE) %>% # Generate by resampling with replacement
      lm(arr_delay ~ dep_delay, data = .) # Calculate statistics
)

boot_data %>%
  summarize(lower = quantile(dep_delay, alpha/2),
    upper = quantile(dep_delay, 1-alpha/2))
```

```
##   lower upper
## 1 0.917  1.18
```

If the flight leaves an additional minute later, then we'd expect the arrival delay to be increased by a value in the interval above, on average.

6.2.6 Confidence Intervals for Prediction

Imagine we are on a plane, we left 15 minutes late, how late will arrive? Since we only have a sample of 100 flights, we are a bit unsure of our prediction.

A classical CI can give us an interval estimate of what the prediction should be (if we had data on all flights).

```
predict(lm.delay2, newdata = data.frame(dep_delay = 15), interval = 'confidence')
```

```
##    fit  lwr  upr
## 1 12.5 8.88 16.1
```

This is taking into account how uncertain we are about our model prediction because our model is based on sample data rather than population data.

6.2.7 Prediction Intervals

We also know that every flight is different (different length, different weather conditions, etc), so the true arrival delay won't be exactly what we predict.

So to get a better prediction for our arrival delay, we can account for the size of errors or residuals by creating a **prediction interval**. This interval will be much wider than the confidence interval because it takes into account how far the true values are from the prediction line.

```
predict(lm.delay2, newdata = data.frame(dep_delay = 15), interval = 'prediction')
```

```
##    fit   lwr  upr
## 1 12.5 -22.9 47.8
```

6.2.8 Probability Theory vs. Bootstrapping

In the modern age, computing power allows us to perform boostrapping easily to create confidence intervals. Before computing was as powerful as it is today, scientists needed mathematical theory to provide simple formulas for confidence intervals.

If certain assumptions hold, the mathematical theory proves to be just as accurate and less computationally-intensive than bootstrapping. Many scientists using statistics right now learned the theory because when they learned statistics, computers were not powerful enough to handle techniques such as bootstrapping.

Why do we teach both the mathematical theory and bootstrapping? You will encounter both types of techniques in your fields, and you'll need to have an understanding of what these techniques are to bridge the gap until statistical inference uses modern computational techniques more widely.

6.3 Hypothesis Testing

Hypothesis testing is another tool that can be used for statistical inference. Let's warm up to the ideas of hypothesis testing by considering two broad types of scientific questions: (1) *Is there* a relationship? (2) *What* is the relationship?

Suppose that we are thinking about the relationship between housing prices and square footage. Accounting for sampling variation...

- ...**is there** a relationship between price and living area?
- ...**what** is the relationship between price and living area?

Whether by the Central Limit Theorem (mathematical theory) or bootstrapping, confidence intervals provide a *range of plausible values* for the true population

parameter and allow us to answer both types of questions:

- **Is there** a relationship between price and living area?
  - Is the no difference/relationship value in the interval?
- **What** is the relationship between price and living area?
  - Look at the estimate and the values in the interval

Hypothesis testing is a general framework for answering questions of the first type. It is a general framework for making decisions between two "theories".

- **Example 1**
  Decide between: true support for a law = 50% vs. true support $\neq$ 50%
- **Example 2**
  In the model Price $= \beta_0 + \beta_1$Area, decide between $\beta_1 = 0$ and $\beta_1 \neq 0$.

In a hypothesis test, we use data to decide between two "hypotheses" labeled as follows:

1. **Null hypothesis** (H_0 = "H naught")
   Hypothesis that is assumed to be true by default.
   A status quo hypothesis: hypothesis of no effect/relationship/difference.
2. **Alternative hypothesis** (H_A or H_1)
   A non-status quo hypothesis.
   Claim being made about the population.

6.3.1 Test statistics

Let's consider the question: Is there a relationship between house price and living area? We can try to answer that with the linear regression model below:

$$\text{Price} = \beta_0 + \beta_1 \text{Area}$$

We would phrase our null and alternative hypotheses as follows:

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0$$

The null hypothesis H_0 describes the situation of "no relationship" because it hypothesizes that the true slope β_1 is 0. The alternative hypothesis posits a relationship: the true slope β_1 is not 0. That is, there is not no relationship. (Double negatives!)

To gather evidence, we collect data and fit a model. From the model, we can compute a **test statistic**, which tells us how far the observed slope is from the null hypothesis value of 0 (called the **null value**). The test statistic is a *discrepancy measure* where large values indicate higher discrepancy with H_0.

The test statistic below is a reasonable proposal:

$$\text{Test statistic} = \frac{\text{estimate} - \text{null value}}{\text{std. error of estimate}}$$

It looks like a z-score. It expresses: how far away is our estimate from the null value in units of standard error? With large values (in magnitude) of the test statistic, our data (our estimate) is discrepant with what the null hypothesis proposes because our estimate is quite far away from the null value in standard error units.

6.3.2 Logic of hypothesis testing

How large in magnitude must the test statistic be in order to make a decision between H_0 and H_A? We will use another metric called a **p-value**.

Assuming H_0 is true, we ask: What is the chance of observing a test statistic which is "as or even more extreme" than the one we just saw? This probability is called a **p-value**.

If our test statistic is large, then our estimate is quite far away from the null value (in standard error units), and then the chance of observing someone this large or larger (assuming H_0 is true) would be very small. **A large test statistic leads to a small p-value.**

If our test statistic is small, then our estimate is quite close to the null value (in standard error units), and then the chance of observing someone this large or larger (assuming H_0 is true) would be very large. **A small test statistic leads to a large p-value.**

6.3.2.1 Making Decisions

If the p-value is "small", then we reject H_0 in favor of H_A. Why? A small p-value (by definition) says that if the null hypotheses were indeed true, we are unlikely to have seen such an extreme discrepancy measure (test statistic). We made an assumption that the null is true, and operating under that assumption, we observed something odd and unusual. This makes us reconsider our null hypothesis.

How small is small enough for a p-value? We will set a threshold α. P-values less than this threshold will be "small enough". When we talk about error rates of the decisions associated with rejecting or not rejecting the null hypothesis, the meaning of α will become more clear.

6.3.3 Summary of procedure

1. State hypotheses H_0 and H_A.
2. Select α, a threshold for what is considered to be a small enough p-value.
3. Calculate a test statistic
4. Calculate the corresponding p-value

5. Make a decision:
   - If p-value $< \alpha$, reject H_0 and accept H_A.
   - Otherwise, we fail to reject H_0 for lack of evidence.
     (Jurors' decisions are "guilty" and "not guilty". Not "guilty" and "innocent".)

6.3.4 Testing single model coefficients

A big emphasis of our course is regression models. It turns out that many scientific questions of interest can be framed with regression models.

In the `summary()` output, R performs the following hypothesis test by default (for any regression coefficient β):

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_A : \beta \neq 0$$

$$\text{Test statistic} = \frac{\text{estimate} - \text{null value}}{\text{std. error of estimate}} = \frac{\text{estimate} - 0}{\text{std. error of estimate}}$$

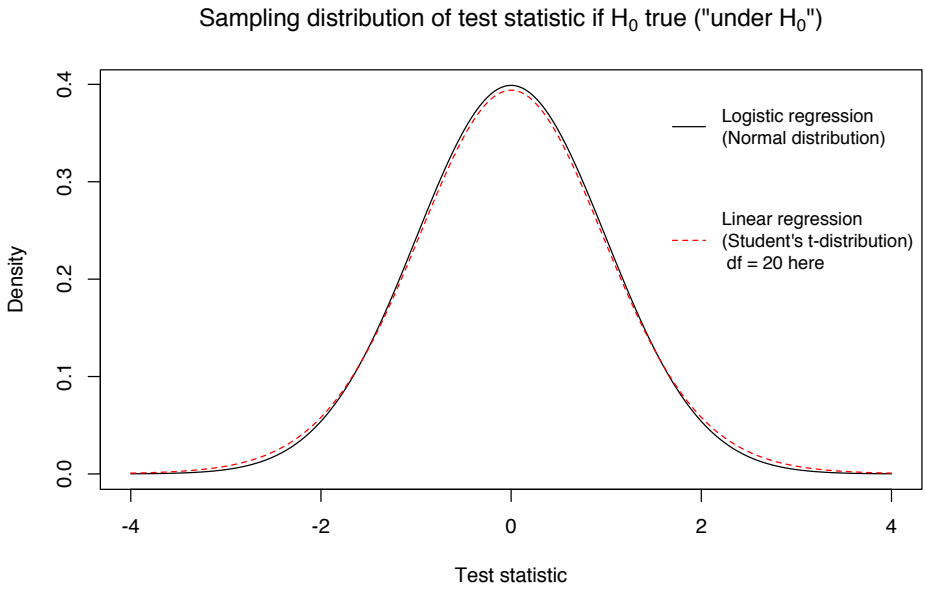
Note that test statistics are random variables! Why? Because they are based on our random sample of data. Thus it will be helpful to understand the distributions of test statistics in terms of probability density functions.

> **Reflect:** If H_0 were true, where would the probability density function of the test statistic be centered?

6.3.5 Distributions of test statistics

What test statistics are we likely to get if H_0 is true? The probability density function of the test statistic "under H_0" (that is, if H_0 is true) is shown below. Note that it is centered at 0. This distribution shows that if indeed the null is true, there is variation in the test statistics we might obtain from random samples, but most test statistics are around zero.

It would be very unlikely for us to get a pretty large (extreme) test statistic if indeed H_0 were true. Why? The density drops rapidly at more extreme values.



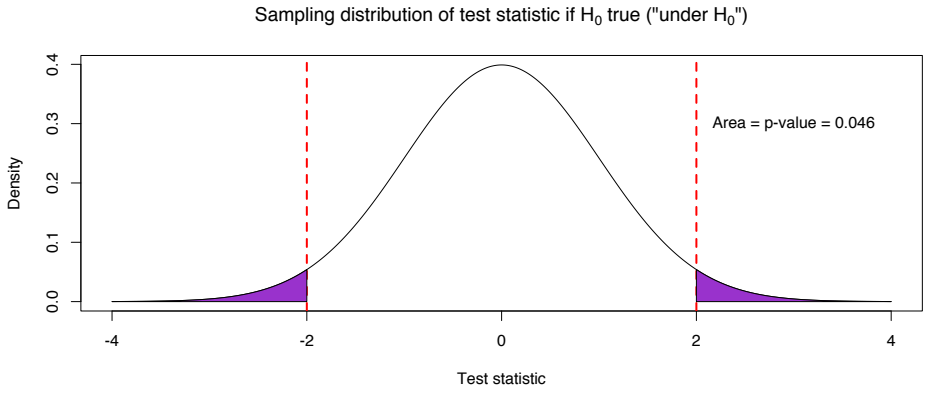
6.3.6 Graphical description of p-values

Suppose that our observed test statistic is 2.

What test statistics are "as or more extreme"?

- Absolute value of test statistic is at least 2: $|\text{Test statistic}| \geq 2$
- In other words: Test statistic ≥ 2 and Test statistic ≤ -2

The p-value corresponding to our test statistic is the area under the probability density function in those "as or more extreme" regions.



6.3.7 Example: Linear Regression

Below we fit a linear regression model of house price on living area:

```
homes <- read.delim("http://sites.williams.edu/rdeveaux/files/2014/09/Saratoga.txt")
mod_homes <- lm(Price ~ Living.Area, data = homes)
confint(mod_homes) ## 95% confidence interval by default
```

```
##             2.5 % 97.5 %
## (Intercept)  3648  23231
## Living.Area   108    118
```

```
summary(mod_homes)$coefficients
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    13439    4992.35    2.69  7.17e-03
## Living.Area      113       2.68   42.17 9.49e-268
```

The `t value` column is the test statistic, and the `Pr(>|t|)` column is the p-value. Note that the "t" comes from the Student t distribution.

- What are H_0 and H_A?
  We write the model Price $= \beta_0 + \beta_1$Area.
  $H_0 : \beta_1 = 0$ (There is no relationship between price and living area.)
  $H_A : \beta_1 \neq 0$ (There is a relationship between price and living area.)

- For a threshold $\alpha = 0.05$, what is the decision regarding H_0?
  Note that when you see `e` in R output, this means "10 to the power". So `9.486240e-268` means 9.49×10^{-268}. This p-value is less than our threshold, so we reject H_0 and say that we have significant evidence for a relationship between price and living area.
- Is this consistent with the confidence interval?
  This result is consistent with the 95% confidence interval in that the interval does not contain 0.

6.3.8 Example: Logistic Regression

Below we fit a logistic regression model of whether a movie made a profit (response) on whether it is a history film:

```
movies <- read.csv("https://www.dropbox.com/s/73ad25v1epe0vpd/tmdb_movies.csv?dl=1")
mod_movies <- glm(profit==TRUE ~ History, data = movies, family = "binomial")
confint(mod_movies) ## 95% confidence interval by default
```

```
##               2.5 % 97.5 %
## (Intercept)  0.0944  0.210
## HistoryTRUE -0.2648  0.309
```

```
summary(mod_movies)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1523     0.0296   5.152 2.58e-07
## HistoryTRUE   0.0207     0.1460   0.142 8.87e-01
```

The `z value` column is the test statistic, and the `Pr(>|z|)` column is the p-value. Note that the "z" refers to z-score and the Normal distribution.

Try for yourselves!

- What are H_0 and H_A?
- For a threshold $\alpha = 0.05$, what is the decision regarding H_0?
- Is this consistent with the confidence interval?

###Errors

Just as with model predictions, we may make errors when doing hypothesis tests.

We may decide to reject H_0 when it is actually true. We may decide to not reject H_0 when it is actually false.

We give these two types of errors names. **Type 1 Error** is when you reject H_0 when it is actually true. This is a false positive because you are concluding there is a real relationship when there is none. This would happen if one study published that coffee causes cancer in one group of people, but no one else could actually replicate that result since coffee doesn't actually cause cancer. **Type 2 Error** is when you don't reject H_0 when it is actually false. This is a false negative because you would conclude there is no real relationship when there is a real relationship. This happens when our sample size is not large enough to detect the real relationship due to the large amount of noise due to sampling variability.

We care about both of these types of errors. Sometimes we prioritize one over the other. Based on the framework presented, we control the chance of a Type 1 error through the confidence level/p-value threshold we used. In fact, the chance of a Type 1 Error is α,

$$P(\text{ Type 1 Error }) = P(\text{ Reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

Let $\alpha = 0.05$ for a moment. If the Null Hypothesis (H_0) is actually true, then about 5% of the time, we'd get unusual test statistics just by chance. With those samples, we would incorrectly conclude that there was a real relationship.

The chance of a Type 2 Error is often notated as β (but this is not the same value as the slope),

$$P(\text{ Type 2 Error }) = P(\text{ Fail to Reject } H_0 \mid H_0 \text{ is false}) = \beta$$

In order to calculate this probability, we'd need to know the value (or at least a good idea) of the true effect.

6.4 Statistical Significance v. Practical Significance

The common underlying question that we ask as Statisticians is "Is there a real relationship in the population?"

We can use confidence intervals or hypothesis testing to help us answer this question.

If we note that the no relationship value is NOT in the confidence interval or the p-value is less then α, we can say that there is significant evidence to suggest that there is a real relationship. We can conclude there is a **statistically significant** relationship because the relationship we observed it is unlikely be due only to sampling variabliliy.

But as we discussed in class, there are two ways you can control the width of a confidence interval. If we increase the sample size n, the standard error decreases and thus decreasing the width of the interval. If we decrease our confidence level (increase α), then we decrease the width of the interval.

A relationship is **practically significant** if the estimated effect is large enough to impact real life decisions. For example, an Internet company may run a study on website design. Since data on observed clicks is fairly cheap to obtain, their sample size is 1 million people (!). With large data sets, we will conclude almost every relationship is statistically significant because the variability will be incredibly small. That doesn't mean we should always change the website design. How large of an impact did the size of the font make on user behavior? That depends on the business model. On the other hand, in-person human studies are expensive to run and sample sizes tended to be in the 100's. There may be a true relationship but we can't distinguish the "signal" from the "noise" due to the higher levels of sampling variability. While we may not always have statistical significance, the estimated effect is important to consider when designing the next study.

Hypothesis tests are useful in determining statistical significance (Answering: "Is there a relationship?").

Confidence intervals are more useful in determining practical significance (Answering: "What is the relationship?")

6.5 Model Selection

In practice, you need to decide which variables should be included in an model. This is referred to as **model selection** or **variable selection**. We have many tools that can help us make these decisions.

- Exploratory visualizations give us an indication for which variables have the strongest relationship with our response of interest (also if it is not

linear)
- R^2 can tell us the percent of variation in our response that is explained by the model – We want HIGH R^2
- The standard deviation of the residuals, s_e (residual standard error), tells us the average magnitude of our residuals (prediction errors for our data set) – We want LOW s_e

With the addition of statistical inference in our tool set, we now have many other ways to help guide our decision making process.

- Confidence intervals or tests for individual coefficients or slopes ($H_0 : \beta_k = 0$, population slope for kth variable is 0 meaning no relationship)
  - See `summary(lm1)`
- Nested tests for a subset of coefficients or slopes ($H_0 : \beta_k = 0$, population slope for kth variable is 0 meaning no relationship)
  - `anova(lm1, lm2)` for comparing two linear regression models (one larger model and one model with some variables removed)
  - `anova(glm1, glm2, test = 'LRT')` for comparing two logistic regression models (one larger model and one model with some variables removed)

In general, we want a simple model that works well. We want to follow Occam's Razor, a philosophy that suggests that it is a good principle to explain the phenomena by the simplest hypothesis possible. In our case, that mean the fewest variables in our models.

So there are model selection criteria that we can use that penalizes you for having too many variables. Here are some below.

- Choose a model with a LOWER Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) calculated as

$$AIC = n\log(SSE) - n\log(n) + 2(k+1) = -2\log(L) + 2(k+1)$$

$$BIC = n\log(SSE) - n\log(n) + (k+1)\log(n) = -2\log(L) + (k+1)\log(n)$$

- Calculated using `BIC(lm1)` and `AIC(lm1)`

- Choose a model with a higher adjusted R^2, calculated as,

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SSTO/(n-1)}$$

where k is the number of estimated coefficients in the model.
  - Find the adjusted R squared in the output of `summary(lm1)`

If our goal is prediction, then we will want to choose a model that has the lowest prediction error. BUT, if we fit our model to our data and then calculate the prediction error from that SAME data, we aren't getting an accurate estimate

of the prediction error because we are cheating. We aren't doing predictions for new data values.

Training and Testing

In order to be able to predict for new data, we can randomly split our observed data into two groups, a **training** set and a **testing** set (also known as a validation or hold-out set). - Fit the model to the training set and do prediction on the observations in the testing set. - The prediction Mean Squared Error, $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$ can be calulated based on the predictions from the testing set.

Drawbacks of Validation Testing

- 1) The MSE can be highly variable as it depends on how you randomly split the data.
- 2) We aren't fully utilizing all of our data to fit the model; therefore, we will tend to overestimate the prediction error.

K-Fold Cross-Validation

If we have a small data set and we want to fully use all of the data in our training, we can do **K-Fold Cross-Validation**. The steps are as follows:

- Randomly splitting the set of observations into k groups, or folds, of about equal size.
- The first group is treated as the test set and the method or model is fit on the remaining $k-1$ groups. The MSE is calculated on the observations in the test set.
- Repeat k times; each group is treated as the test set once.
- The k-fold CV estimate of MSE is an average of these values,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

where MSE_i is the MSE based on the ith group as the test group. In practice, one performs k-fold CV with $k = 5$ or $k = 10$ as it reduces the computational time and it also is more accurate.

```
require(boot)
cv.err <- cv.glm(data,glm1, K = 5)
sqrt(cv.err$delta[1]) #out of sample average prediction error
```

Chapter 7

Appendix A - Theoretical Probability

For those of you who want a more thorough, mathematical description of the basics of theoretical probability, please read on.

7.1 Probability Rules

In theoretical probability, we need to define a few terms and set some rules (known as axioms).

The **sample space,** S, is the set of all possible outcomes of a random process.

- Example: If you flip two coins (one side Heads and one side Tails), then the sample space contains four possible outcomes: Heads and Heads (HH), Heads and Tails (HT), Tails and Heads (TH), and Tails and Tails (TT), $S = \{HH, HT, TH, TT\}$.

A subset of outcomes is called an **event**, A.

- Example: If you flip two coins, an event A could be that exactly one of the coins land Heads, $A = \{HT, TH\}$.

For the rules of probability, we can define them with set notation as well as words. If you aren't familiar with set notation,

- $\cup$ means union (inclusive OR)
- $\cap$ means intersection (AND)
- A^C means complement (NOT)

For events A and B and sample space S, the probability of an event A, notated as $P(A)$, follows the rules below:

- Rule 1: $0 \leq P(A) \leq 1$ (probability has to be between 0 and 1)
- Rule 2: $P(S) = 1$ (one of the outcomes has to happen)
- Rule 3: $P(A^c) = P(\text{not } A) = 1 - P(A)$ (if we know the chance of something happening, we also know that chance it doesn't happen)
- Rule 4: $P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$ if A and B are disjoint events.
  - A and B are **disjoint/mutually exclusive** if A occuring prevents B from occurring (they both can't happen at the same time).
- Rule 4*: $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Rule 5: $P(A \cap B) = P(A \text{ and } B) = P(A) \times P(B)$ if A and B are independent.
  - A and B are **independent** if B occurring doesn't change the probability of A occurring.
- Rule 5*: $P(A \text{ and } B) = P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$.
  - The **conditional probability** of A **given** that event B occurs, $P(A \mid B)$, is equal to the probability of the joint event (A and B) divided by the probability of B.

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

  - Intuition: Given that B happened, we focus on the subset of outcomes in S in which B occurs and then figure out what the chance of A happening within that subset.

7.1.0.1 Example: Blood Types

The American Red Cross estimates that 45% of U.S. population has Type O blood, 40% are Type A, 11% Type B, and 4% AB blood.

Imagine that we have a blood drive in St. Paul. The next donor's blood type can be thought of as a random process. The sample space for this random process includes the 4 blood types: $S = \{O, A, B, AB\}$ (it includes all possible outcomes). Assume the people who donate blood have the same distribution of blood types as the U.S. and that St. Paul has the same distribution as the entire U.S.

Think about how you'd justify your answer to the following questions:

1. What is the probability that the next donor is Type O blood?
2. What is the probability that the next donor is Type A or Type B or Type AB blood?
3. What is the probability the next three donors are all Type O blood?
4. What is the probability the next donor is Type O or Type A or Type B or Type AB?

If the possible outcomes were equally likely, we could calculate probabilities

$$P(A) = \frac{\text{Number of outcomes in } A}{\text{Number of possible outcomes}}$$

But the chances of Type O, A, B, and AB blood are all different because they occur with different frequency in the population.

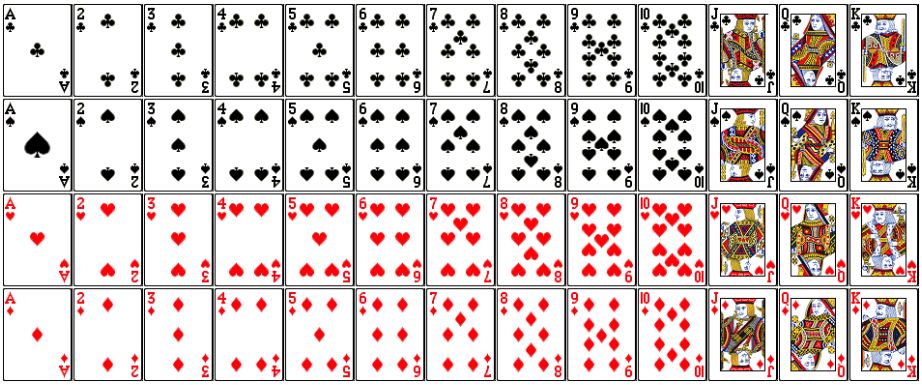
Let's change the sample space to make it easier. Let our sample space, S, be the set of 100 equally likely outcomes (45 are O, 40 are A, 11 are B, and 4 are AB). Now, you can calculate probabilities based on this framework of equally likely outcomes (after we changed the sample space).

1. P(Type O) = 45/100 assuming equally likely outcomes
2. P(Type A or B or AB) = 1 - P(Type O) = 1 - 45/100 by Rule 3
3. P(Type O and then Type O and then Type O) = $(45/100)^3$ by Rule 5 assuming donors are independent, in that the probability of Type O blood stays the same
4. P(Type A or B or AB or O) = P(S) = 1 by Rule 2

7.1.0.2 Example: 52 Cards

Let's consider a perfectly shuffled deck of playing cards. Each card has an icon and a number (or A, J, Q, K) on it. The icon is either a red heart, red diamond, black spade (leaf), or black club (3 leaf clover). The numbers range for 2 to 10 and A is for Ace, J is for Jack, Q is for Queen, K is for King.

The sample space is below.



1. What is the probability of drawing a card with a heart icon on it?
2. What is the probability of drawing a card with a heart or Ace (A) on it?
3. What is the probability of dealing a card with a heart on the table and then another heart card?

Focus on how we'd justify the answer, not just the number.

1. P(heart) = 13/52 by equally likely outcomes
2. P(heart or ace) = 13/52 + 4/52 - 1/52 = 16/52 by Rule 4*
3. P(heart and then heart) = 13/52*12/51 by Rule 5* (draws are not independent here since the probability of hearts changes after you remove a card)

7.1.1 Disjoint/Mutually Exclusive

Think back to the Blood Type example.

Let's say we were interested in the next two donors.

- P(First Type A or Second Type A) = ?

Think about all the ways this could happen.

We will always use an inclusive OR, which means that we care about one or the other or both happening. We just need to make sure we don't double count, which is why we subtract the chance of both.

- P(First Type A or Second Type A) = P(First Type A) + P(Second Type A) - P(both Type A)

So,

- P(First Type A or Second Type A) = 0.40 + 0.40 - 0.40*0.40 = 0.64

Which is the same as if we were to consider the three disjoint options (A: Type A, N: Not Type A),

- P(AN or NA or AA) = 0.4*0.6 + 0.6*0.4 + 0.4*0.4 = 0.64

7.1.2 Independence

Let's stay with the Blood Type example for one moment more.

What if there were only 50 donors in St. Paul? Say 30 of them Type 0 and the other 20 were A or B.

- Would the second donor be independent of the first donor? In other words, would the probability of getting a Type O donors change between donors?

No, they wouldn't be independent. In that case, let's calculate the probability that the first two donors are Type O.

- P(Type O and then Type O) = P(Type O)P(2nd Type O | 1st Type O) = (30/50) * (29/49) = 0.355

7.2 Random Variable

A **Random Variable** (X) is a real-valued function whose outcome we don't know beforehand.

- It is a function of the outcomes from a random process.

I am going to flip a fair coin 3 times (the coin has 2-sides, we'll call one side Heads and the other Tails).

- Assume there are only 2 possible outcomes and P(Heads) = P(Tails) = 0.5 (can't land on its side).
- Below are three possible random variables based on the same random process (flipping a 2-sided coin 3 times):
- Example 1 - X: the number of heads in 3 coin flips
- What are the possible values of X?
- Example 2 - Say I give you 3 dollars for each head
- Y: the amount of money won from 3 coin flips, $Y = 3 * X$
- Example 3 - Z: the number of heads on the last flip of 3 coin flips
- The possible values are 0 or 1.

7.2.1 Probability Models

A **probability model** for random variable X gives the possible values of X and the associated probabilities.

- We have the probability model for X: the number of heads in 3 coin flips.
- What is the probability model for $Y = 3 * X$?
- What about Z?

7.3 Discrete Random Variables

- If there are a finite (more generally, countable) number of possible values, we say that X is a **discrete random variable**.
- We often can write the probability as a function of values, x, and we call this function the **probability mass function (pmf),**

$$p(x) = P(X = x)$$

- and we know that

$$\sum_{all\ x} p(x) = 1$$

7.3.1 Expected Value

The **expected value** (or long-run average) of a discrete random variable is defined as the weighted average of the possible values, weighted by the probability,

$$E(X) = \sum_{all\ x} xp(x)$$

So the expected value is like a mean, but over the long-run.

7.3.2 Variance

The **variance** (or long-run spread) of a discrete random variable is defined as the "average" squared distance of X from its expected value,

$$Var(X) = E[(X - \mu)^2]$$

where $\mu = E(X)$.

- But it's in squared units, so typically we talk about its square root, called the **standard deviation** of a random variable,

$$SD(X) = \sqrt{Var(X)}$$

So the standard deviation of a random variable is like the standard deviation of a set of observed values. They are measures of spread and variability. In one circumstance, we have the data to calculate it and in the other, we are considering a random process and wondering how much a value might vary.

7.3.3 Joint Distributions

The **joint probability mass function** for two random variables is

$$p(x, y) = P(X = x \text{ and } Y = y)$$

We can often calculate this joint probability using our probability rules from above (using multiplication...)

- The expected value for a function of two random variables is

$$E(g(X, Y)) = \sum_{all\ y} \sum_{all\ x} g(x, y)p(x, y)$$

- We could show that the expected value of a sum is the sum of the expected values:

$$E(X + Y) = E(X) + E(Y)$$

- Using this fact, we could show that the variance can be written in this alternative form:

$$Var(X) = E[(X-\mu)^2] = E(X^2) - [E(X)]^2$$

7.3.4 Covariance

When consider two random variables, we may wonder whether they co-vary? In that do they vary together or vary independently? If

- The **covariance** of two random variables is defined as

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)$$

- Note: The covariance of X with itself is just the variance, $Cov(X,X) = Var(X)$

We could use this to show that $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

Two discrete random variables are **independent** if and only if

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$$

for every x and y.

- If two random variables, X and Y are independent, then $Cov(X,Y) = 0$.

7.3.5 Correlation

- The **correlation** of two random variables is

$$Cor(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

7.3.6 A Few Named Probability Models

7.3.6.1 Bernoulli Trials

Three conditions

- Two possible outcomes on each trial (success or failure)
- Independent Trials (result of one trial does not impact probabilities on next trial)
- P(success) $= p$ is constant

$$P(X = x) = p^x(1-p)^{x-1} \text{ for } x \in \{0,1\}$$

$$E(X) = p$$

$$Var(X) = p(1-p)$$

Binomial RV: X is the total number of successes in n trials

For general n and x, the Binomial probability for a particular value of x is given by

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \text{ for } x \in \{0, 1, 2, ..., n\}$$

where $x! = x * (x-1) * (x-2) * \cdots * 2 * 1$ and $0! = 1$, so

$$\frac{n!}{(n-x)!x!} = \frac{n * (n-1) * \cdots * (n-x+1) * (n-x)!}{(n-x)!x!}$$

$$= \frac{n * (n-1) * \cdots * (n-x+1)}{x * (x-1) * \cdots * 2 * 1}$$

If we break this apart, we can see where the pieces came from. Let's consider a simplified example. Let X be the number of Heads in 3 coin flips (but the coin is biased such that $p = 0.2$).

The probability of 2 successes and 1 failure in one particular order (e.g. HHT) is calculated as $p^x(1-p)^{n-x} = 0.2^2(0.8)$ due to Rule 5. However, we could have gotten a different ordering of Heads and Tails (e.g. HTH, THH). To count the number of ways we could get 2 heads and 1 tail in 3 coin flips, we use tools from combinatorics (an area of mathematics). In fact, the first part of the equation does the counting for us,

$$\frac{n!}{(n-x)!x!} = \frac{n * (n-1) * \cdots * (n-x+1) * (n-x)!}{(n-x)!x!}$$

So for our example, there are $\frac{3!}{2!1!} = \frac{3*2!}{2!1} = 3$ orderings of 2 heads and 1 tail.

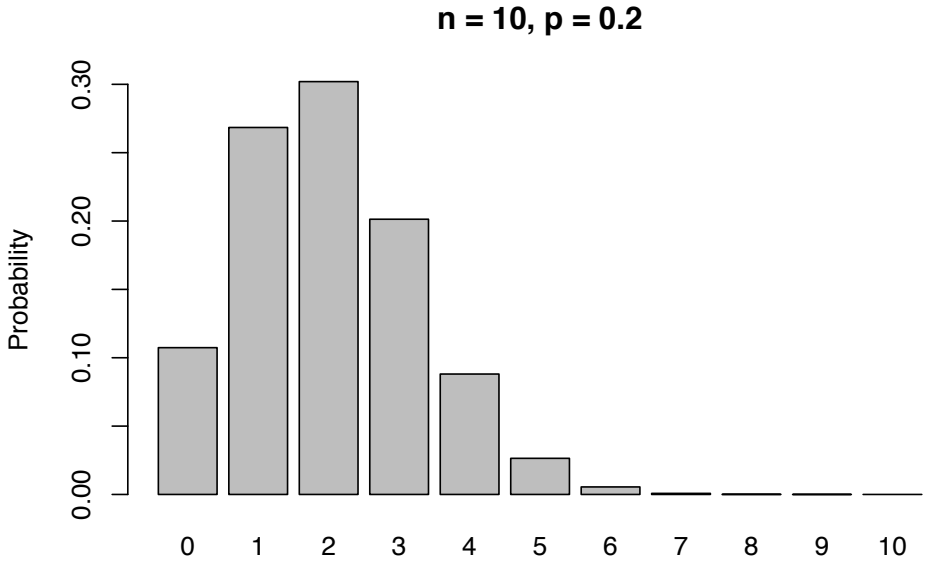
The expected number of successes in the long run is

$$E(X) = np$$

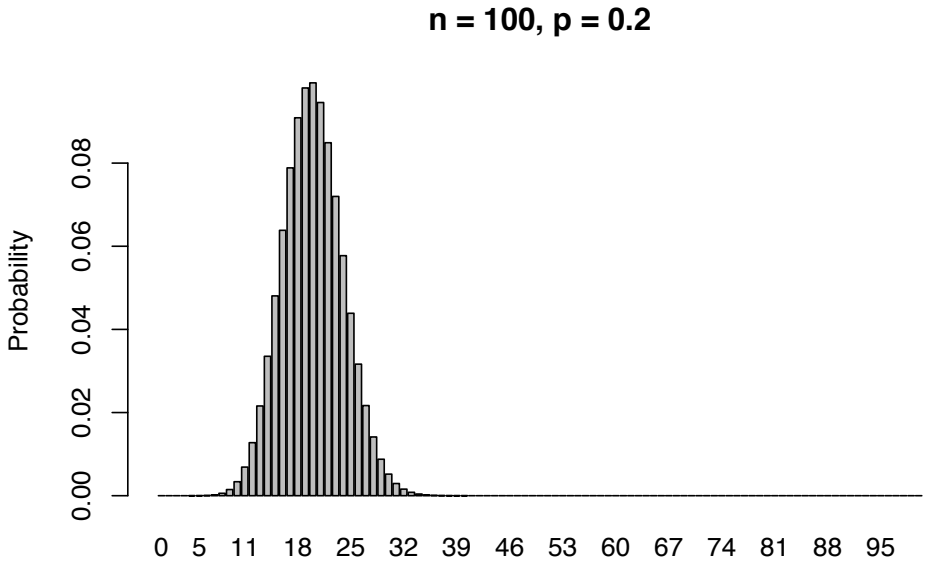
and the variability in the number of successes is given by

$$Var(X) = np(1-p)$$

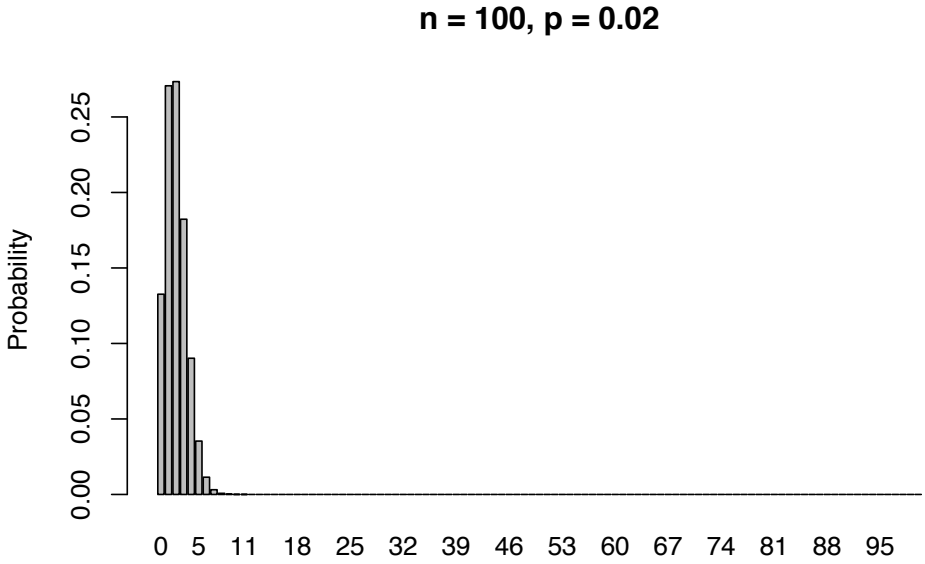
Let's plot the pmf of the Binomial in a bar plot,



If we increase n, but leave p, then



If we increase n, but decrease p proportionally (such that np stays the same), then



We will talk about two ways to approximate the Binomial distribution.

- If n increases while p stays fixed, then we use a Normal approximation.
- If n increases and p decreases, then we use a Poisson approximation (beyond the scope of this course).

7.4 Continuous Random Variables

For continuous random variables X (uncountable, infinite number of values),

- the probability of any one value is 0, $P(X = x) = 0$.
- So we define the probability model using a **culmulative distribution function** (cdf), the probability of having a value less than x,

  $$F(x) = P(X \leq x)$$

  (it is always notated with a capital letter F or G or H).
- and a **probability density function** (pdf), $f(x) \geq 0$ such that the probability is defined by the area under this curve (defined by the pdf). Using calculus, the area under the curve is

  $$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  (it is always notated with a small letter f or g or h) and the total area under the probability density function is 1,

  $$P(S) = P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

- Thus, we can write the cumulative distribution function as, $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$.

7.4.1 Expected Value

Let X be a continuous RV with pdf $f(x)$. The expected value of X is defined as

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Properties of Expected Value

These properties still hold:

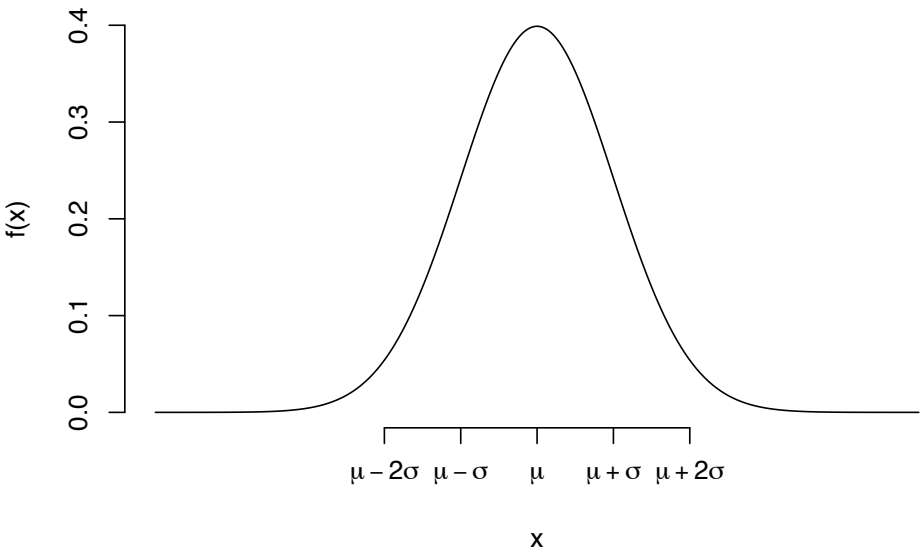
$$E(aX) = aE(X)$$
$$E(X + b) = E(X) + b$$

7.4.2 A Few Named Probability Models

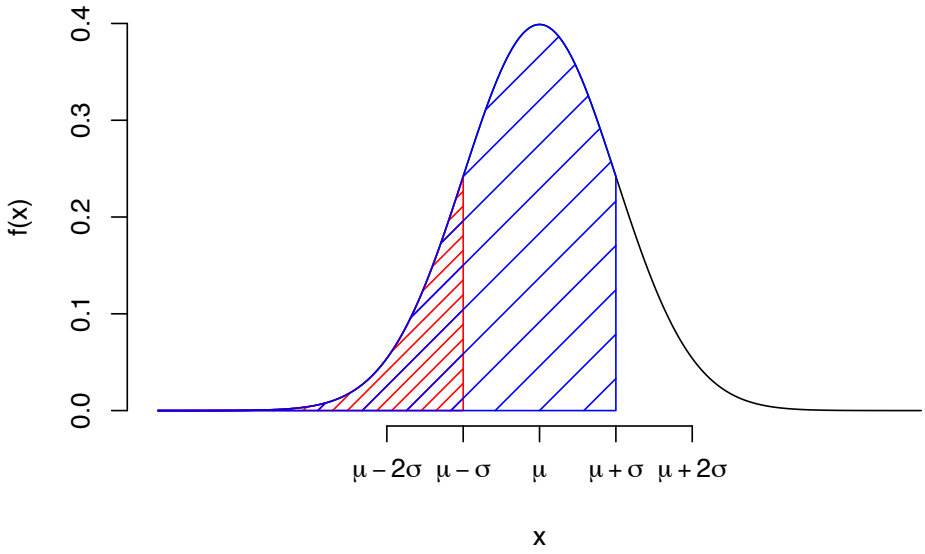
Normal Model

For X such that $E(X) = \mu$ and $SD(X) = \sigma$, a Normal random variable has a pdf of

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



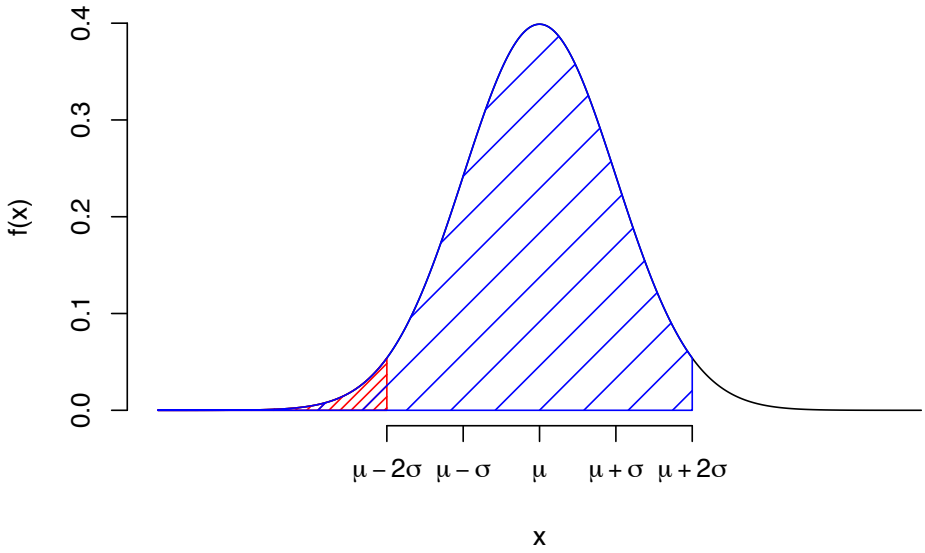
- Let the expected value be 0 and standard deviation be 1, $\mu = 0$ and $\sigma = 1$
- We know that $P(-1 \leq X \leq 1) = F(1) - F(-1) = 0.68$




```
pnorm(1) - pnorm(-1) #pnorm is the cdf
```

```
## [1] 0.683
```

- $P(-2 \leq X \leq 2) = F(2) - F(-2) = 0.95$




```
pnorm(2) - pnorm(-2)
```

```
## [1] 0.954
```

- $P(-3 \leq X \leq 3) = F(3) - F(-3) = 0.997$

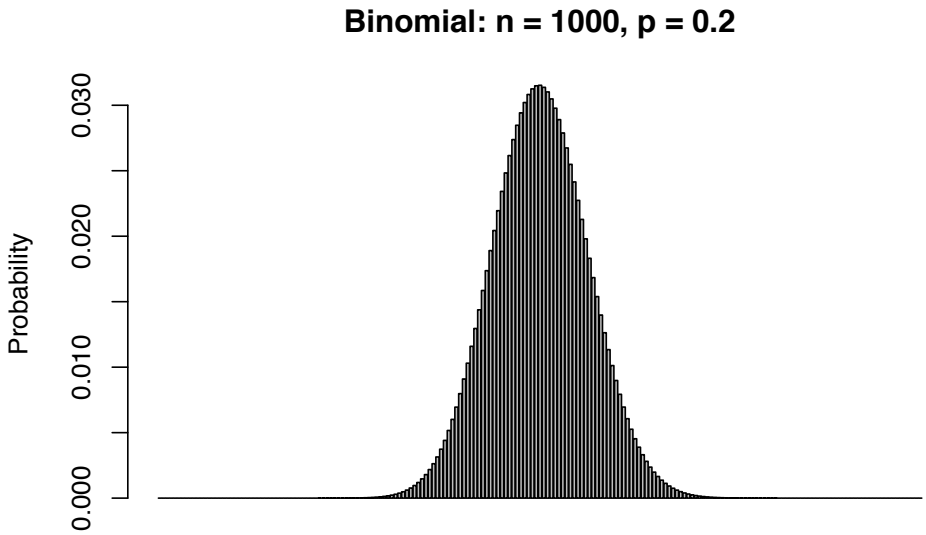
```
pnorm(3) - pnorm(-3)
```

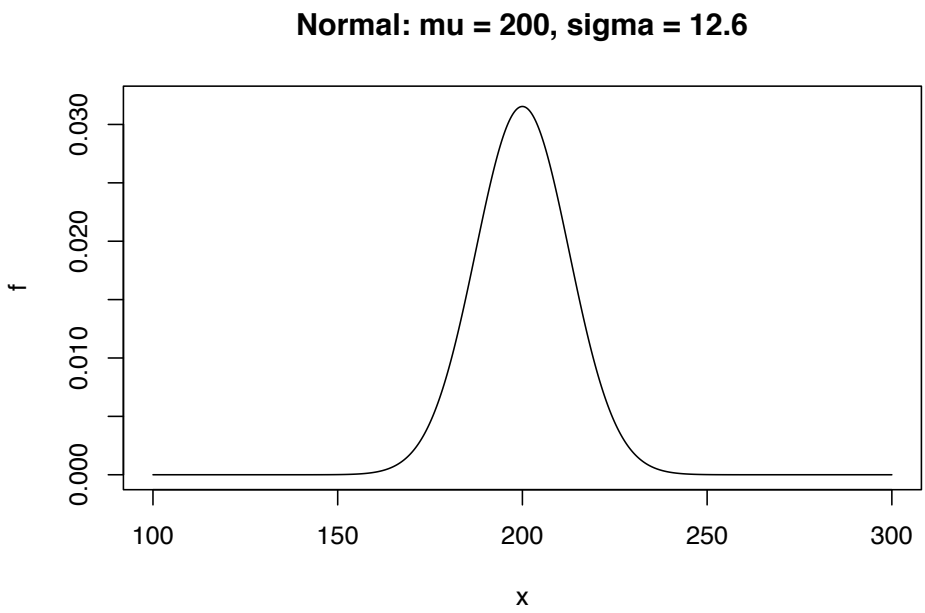
```
## [1] 0.997
```

Let X be a Binomial Random Variable and Y be a Normal Random Variable.

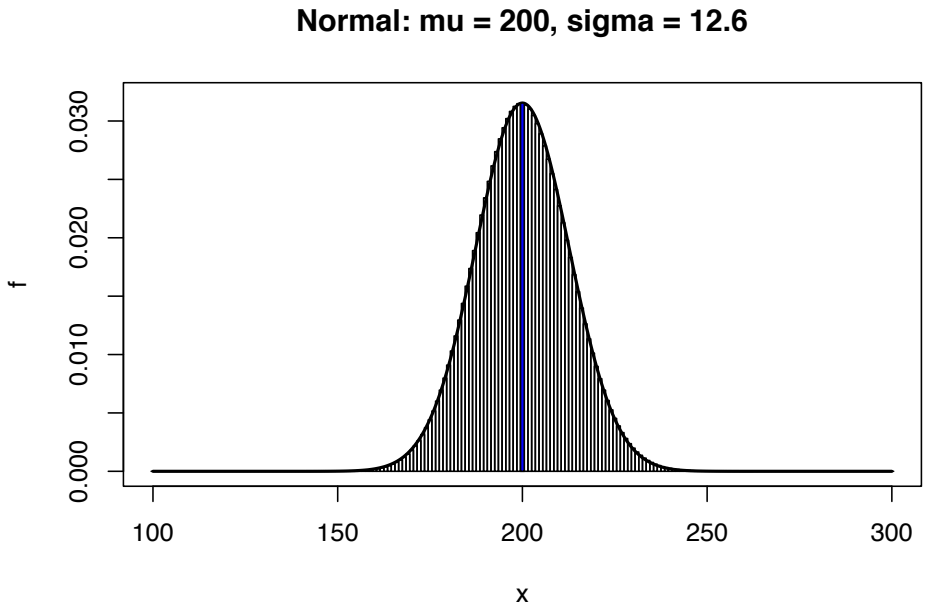
As $n \to \infty$ (p is fixed), the $P(X = x) \approx P(x - 0.5 \leq Y \leq x + 0.5)$.

Note: adding and subtracting 0.5 is the continuity correction





If $n = 1000$ and $p = 0.2$, let's compare $P(X = 200)$ and $P(199.5 \leq Y \leq 200.5)$.



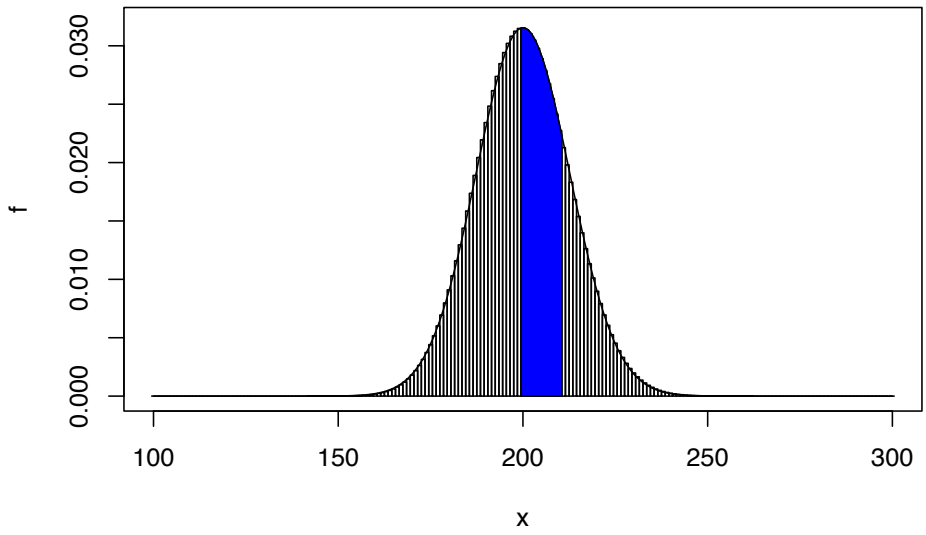

```
dbinom(200,size = n, p = p)
```

```
## [1] 0.0315
```

```
pnorm(200.5,mean = n*p, sd = sqrt(n*p*(1-p))) - pnorm(199.5,mean = n*p, sd = sqrt(n*p*
```

```
## [1] 0.0315
```

If $n = 1000$ and $p = 0.2$, let's compare $P(200 \leq X \leq 210)$ and $P(199.5 \leq Y \leq 210.5)$.




```
sum(dbinom(200:210,size = n, p = p))
```

```
## [1] 0.31
```

```
pnorm(210.5,mean = n*p, sd = sqrt(n*p*(1-p))) - pnorm(199.5,mean = n*p, sd = sqrt(n*p*(1-p)))
```

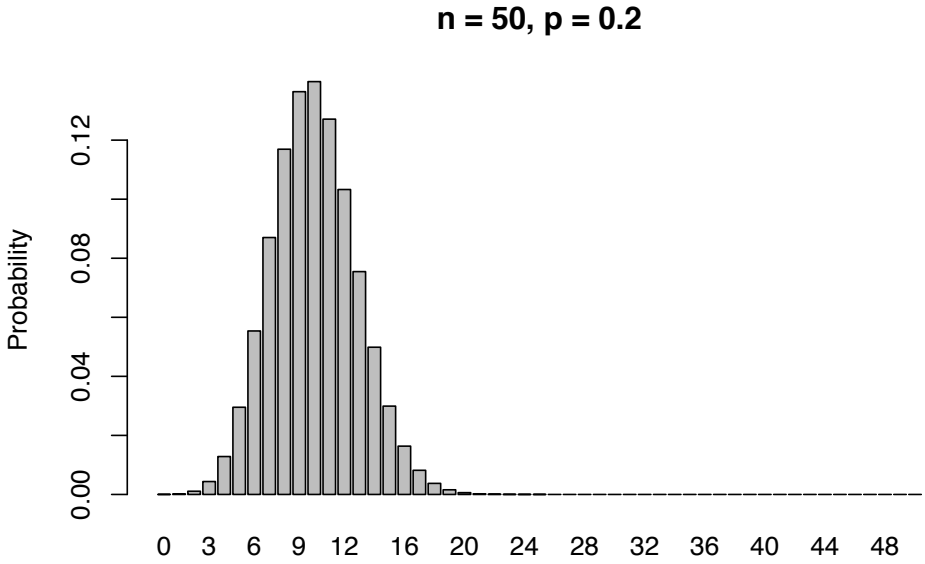
```
## [1] 0.313
```

How big does n have to be for the Normal approximation to be appropriate?

- **Rule of Thumb**: $np \geq 10$ and $n(1-p) \geq 10$ because that makes sure that $E(X) - 0 > 3SD(X)$ (mean is at least 3 SD's from 0).

For $p = 0.2$, that means that $n \geq 50$.

```
n = 50
p = 0.2
barplot(dbinom(0:n,size = n, p = p),names.arg=0:n,ylab='Probability',main='n = 50, p = 0.2')
```



7.5 Random Variation

How has randomness come up in the course so far?

- Random sampling (sampling variation)
- Random assignment of treatment
- Random variation in general (due to biology, measurement error, etc.)

We want to be able to harness the randomness by understanding the random behavior in the long run.

- If we were to repeatedly take samples from the population, how would the estimates (mean, odds ratio, slope etc.) differ?
- If we were to repeat the random assignment many times, how would the estimated effects differ?

Now, based on the theory we know, we could show a few things about means, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Say we have a sequence of independent and identically distributed (iid) random variables, $X_1, ..., X_n$, *(I don't know what their probability model is but the expected value and variance is the same,* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2$*)*

Then we'd expect the mean to be approximately μ,

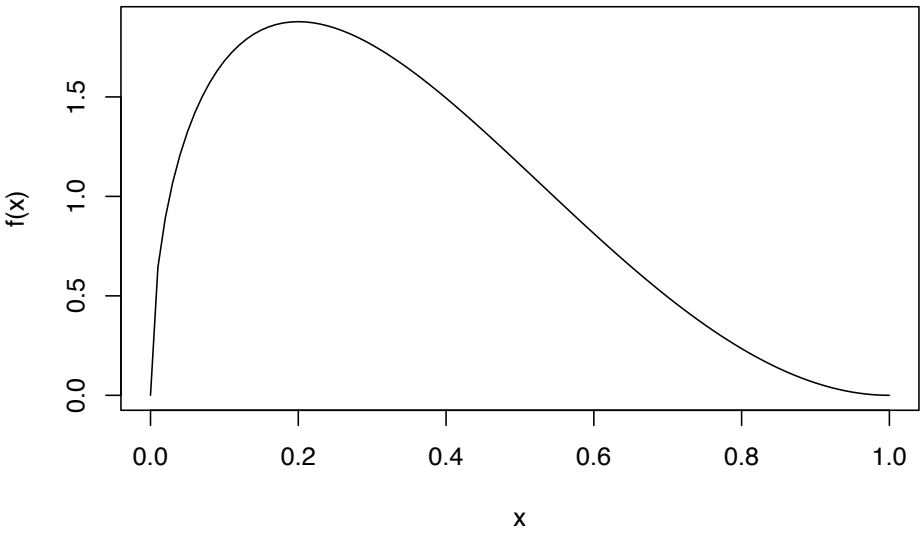
$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mu$$

and the mean would vary, but that variation would decrease with increased sample size n,

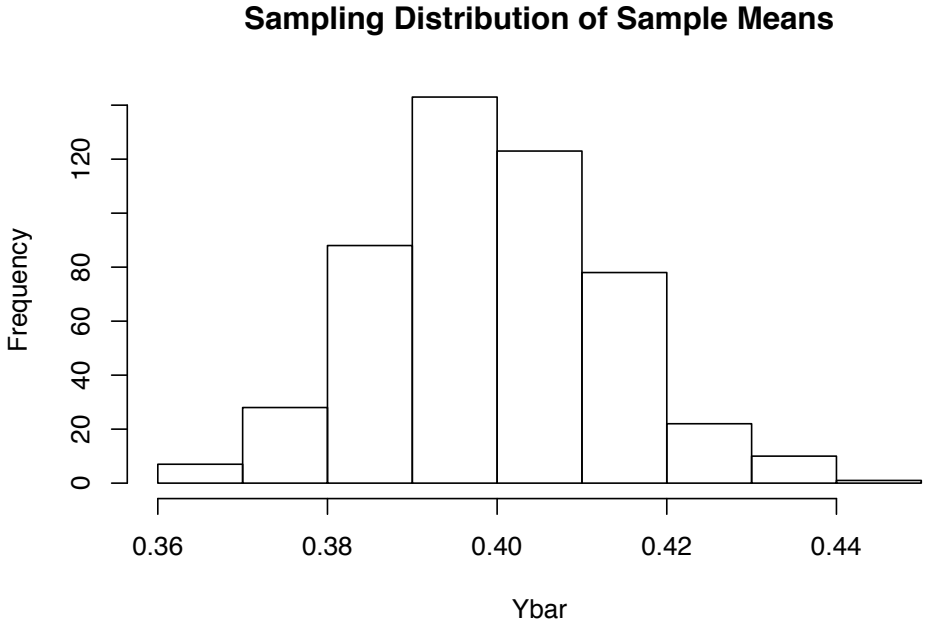
$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$$

But, what is the shape of the distribution (probability model) of the mean?

Let's randomly generate data from a probability model with a skewed pdf.



Let $\bar{y}$ be the mean of those n random values. If we repeat the process multiple times, we get a sense of the **sampling distribution for** $\bar{y}$, the mean of a sample of n random values from the population distribution above.



The **Central Limit Theorem** (CLT) tells us that as the sample size get larger and larger, the shape of the sampling distribution for the sample mean get closer and closer to Normal. That is why it makes sense we've seen unimodal, symmetric distributions pop up when we simulate bootstrapping and random assignments. However, the CLT only applies when we are talking about sample means or proportions.

Let our sample mean be $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ based on a sample size of n.

- Let's subtract the expected value, $E(Y) = \mu$, and scale by $\sqrt{n}$, such that we have a new random variable,

$$C_n = \sqrt{n}\,(Y_n - \mu)$$

- The **Central Limit Theorem** tells us that for any $c \in \mathbb{R}$,

$$\lim_{n\to\infty} P(C_n \leq c) = P(Y \leq c)$$

  where Y is a Normal RV with $E(Y) = 0$ and $Var(Y) = \sigma^2$.