Introduction to Statistics

Dr. Lauren Perry

Contents

Welcome to Statistics!

For the Student

There are a lot of ways to approach an introductory statistics class. Historically, the topics found in these course notes have been taught in a way that emphasizes hand calculations and the use of tables full of numbers.

My philosophy is a little different. This class is designed for students who will need to read statistical results and may need to produce basic statistics using a computer. If you go on to be a scientist and need more statistical know how, this course will give you enough background knowledge to take the inevitable next course in statistics. There is plenty of math here, but none of these situations require the ability to do that math by hand.

In many sections, the math is provided and explained but not emphasized. This is intentional. Instead, we focus on the "why"... Why do we care about this topic? Why is this concept important? Why do I run this test when I have that kind of data? ...and we focus on the interpretation. What does this number tell us about an experiment? What can we conclude based on these statistical results?

We see statistics all the time in the media - in the form of graphs, tables, averages, predictions about elections or sports, you name it! Hopefully, by learning the whys and the interpretation, you will finish this text feeling like you can read and understand statistical results when you run into them in the real world.

R Programming

This text is designed to teach you introductory statistics with the option to learn some R (a statistical programming language) along the way. As a result, some sections have some introductory material on R. R is an incredibly powerful tool, but we're going to keep it relatively simple. Using R will save us the headache of doing a lot of calculations by hand.

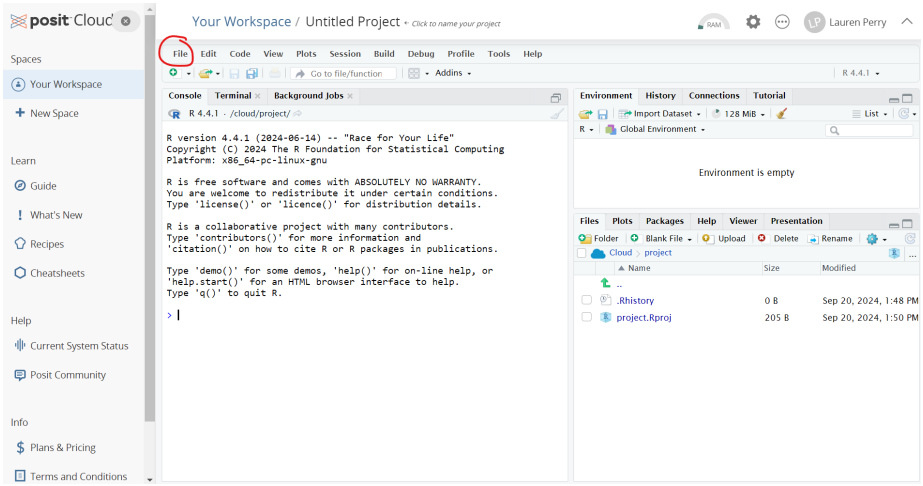
For right now, you can run R right here in the course notes! This is exactly what you will see on the rdrr.io website. Type in your command and click the green "Run" button. Try running the command `print("Welcome to Statistics!")`.

If it prints out "Sorry, something went wrong. All I know is:", just press the "Run" button again.

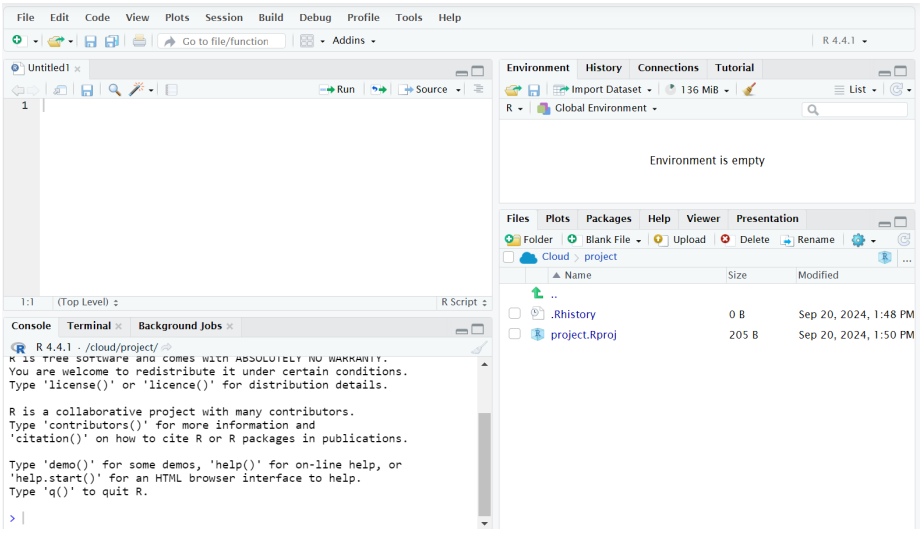
As we progress, we will run it completely online at the website posit.cloud. You do need to create an account to use this website. This will allow you to save your work as you go. I strongly recommend you bookmark this website and it to run the code alongside the R labs.

If you prefer, you can also download R and RStudio to your computer. There are basic instructions here.

Let's take a look at Posit Cloud. When you first log in, you will see a landing page that says "Your Content". On the upper right part of the screen you should see a button that says "New Project". Click on the button and select "New RStudio Project" from the dropdown menu.



Click on "File" (circled in red above) then "New File" and "R Script" to get the following four "panels":



We will take a look at each panel in turn.

1. *Bottom left*: This is the R command console, where you will type your R commands to be run immediately. It's possible to do everything discussed in these course notes without ever leaving this panel! Whenever you finish typing an R command, just hit "Enter" on your keyboard to run it. If you made a mistake before pressing enter, you can press the up arrow on your keyboard to get the line of code back so that you can edit it.
2. *Top left*: This is a place where you can write (and edit) multiple lines of R commands at once. To run one of these lines, press the "Run" button at the top right of the panel. This is also a good place to write code if you want to save it for later. Selecting the save icon will allow you to save whatever code you've written here directly to your posit.cloud account (or to your computer, if you are working locally in RStudio).
3. *Bottom right*: This is where any plots/data visualizations will appear.
4. *Top right*: This is where stuff stored in the R environment will appear.

For the Instructor

Thanks for checking out my Introduction to Statistics course! Sections are designed to be short, easy-to-read introductions to each concept. Some of the more conceptual sections do not have section exercises, but I am working on adding exercises wherever it seems appropriate. The topics and course ordering reflect the department syllabus for the 3-unit Introduction to Statistics at Sacramento State. I am sure there are topics we've left out, but there are only so many things one can cover in 15 weeks.

Each module after the first is designed to take approximately two weeks of class time. In an ideal world, I would cover at least the first nine in a 15 week semester.

However, with assessment, activities, student questions, holidays, etc., I usually get through the first eight. Despite the time constraints, I am slowly working on including additional topics.

These notes are a work in progress and gets updated each semester that I each Introduction to Statistics (which is very nearly every semester) and sometimes during winter and summer breaks.

Currently, I am focused on creating more in-class worksheets and activities. Solutions to these, and the text's exercise problems, are provided to students through the learning management system.

Slides for many of the sections are available on my website: lgpperry.github.io/teaching/stat1/, although these days I present mostly by writing on the board.

Please feel free to reach out to me with any questions, comments, or concerns by emailing me at perry@csus.edu

Course Learning Outcomes

The CLOs for Stat 1: Introduction to Statistics at Sacramento State are as follows.

Students will be able to:

1. Organize, summarize, and interpret data in tabular, graphical, and pictorial formats.
2. Organize and interpret bivariate data and learn simple linear regression and correlation.
3. Understand the basic rules of probability.
4. Use the binomial distribution as a model for discrete variables.
5. Use the normal distribution as a model for continuous variables.
6. Apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation.
7. Apply techniques of testing various statistical hypotheses concerning population parameters.

Stat 1 is also a General Education Area B4 course. The Area B4 learning outcomes are as follows.

Students will be able to:

A. Solve problems by thinking logically, making conjectures, and constructing valid mathematical arguments.
B. Make valid inferences from numerical, graphical and symbolic information.
C. Apply mathematical reasoning to both abstract and applied problems, and to both scientific and non-scientific problems.

Each module also has module-specific learning outcomes and their corresponding CLOs. (The Area B4 outcomes are not included, as each module addresses all

three outcomes.)

Chapter 1

Introduction to Data

What is statistics? There are two ways to think about this:

1. Facts and data, organized or summarized in such a way that they provide useful information about something.
2. The science of analyzing, organizing, and summarizing data.

As a field, Statistics provides tools for scientists, practitioners, and laypeople to better understand data. You may find yourself using knowledge from this course in a research lab, while reading a research report, or even while watching the news!

Module Learning Objectives/Outcomes

After completing Module 1, you will be able to:

1. Understand basic statistical terminology.
2. Describe sampling and experimental design techniques.
3. Organize and visualize data using techniques for exploratory data analysis.
4. Identify the shape of a data set.
5. Understand and interpret graphical displays.

R objectives

1. Manually enter data.
2. Generate random numbers.
3. Create histograms.

This module's outcomes correspond to course outcomes (1) organize, summarize, and interpret data in tabular, graphical, and pictorial formats and (2) organize and interpret bivariate data and learn simple linear regression and correlation.

1.1 Statistics Terminology

There are two ways to think about statistics:

1. **Descriptive statistics** are methods for *describing* information.

For example, 66% of eligible voters voted in the 2020 presidential election (the highest turnout since 1900!).

2. **Inferential statistics** are methods for *drawing inference* (making decisions about something we are uncertain about).

For example, a poll suggests that 75% of voters will select Candidate A. People haven't voted yet, so we don't know what will happen, but we could reasonably conclude that Candidate A will win the election.

The first three modules of this text are dedicated to methods for descriptive statistics. Modules 4 and 5 build up some background information to help with inferential statistics, and then Modules 6 and beyond deal with inferential statistics.

Data is factual information. We collect data from a **population**, the collection of all individuals or items a researcher is interested in.

- Collecting data from an entire population is called a **census**.
  - This is complicated and expensive! There's a reason the United States only does a census every 10 years.
- We can also take a **sample**, a subset of the population we get data from.
  - If you think of the population as a pie, the sample is a small slice. Whether it's a pumpkin pie, a cherry pie, or a savory pie, the small slice will tell you that. We don't need to eat the entire pie to learn a lot about it!

Data are often organized in what we call a **data matrix**. If you've ever seen data in a spreadsheet, that's a data matrix!

	Age	**Gender**	**Smoker**	**Marital Status**
Person 1	45	Male	yes	married
Person 2	23	Female	no	single
Person 3	36	Other	no	married
Person 4	29	Female	no	single

Each row (horizontal) represents one **observation** (also called **observational units**, **cases**, or **subjects**). These are the individuals or items in the sample.

Each column (vertical) represents a **variable**, the characteristic or thing being measured. Think of variables as measurements that can *vary* from one observation to the next.

There are two broad types of variable, each of which can be further categorized into two sub-types:

1. **Numeric** or **quantitative** variables take *numeric* values AND it is sensible to do math with those values.
    a. **Discrete numeric** variables take numeric values with jumps. Typically, this means they can only take whole number values. These are often counts of something. For example, the number of pets you have, the number of cars that drive through an intersection during rush hour, or the number of classes students are taking.
    b. **Continuous numeric** variables take values "between the jumps". Typically, this means they can take decimal values. For example, weights of guinea pigs, milliliters of medication administered, or any measurements of time.
2. **Categorical** or **qualitative** variables take values that are *categories*. These could be something like gender, ice cream flavors, or dog breeds.
    a. **Ordinal categorical** variables have categories with some kind of intrinsic ordering, meaning we can rank them in a meaningful way. For example, a survey asking for approval levels might have categories "strongly disapprove, disapprove, neutral, approve, strongly approve"; and letter grades have the standard ordering "A, B, C, D, F".
    b. **Nominal categorical** variables have categories with no intrinsic ordering. Examples include eye color, college major, and the city people live in.

The "Does it make sense"? Test

- Sometimes, categories can be represented by numbers. Ask yourself if it makes sense to do math with those numbers. If it doesn't make sense, it's probably a categorical variable. Some examples: zip codes, phone area codes, or student ID numbers.
- If you're unsure whether a variable is discrete or continuous, pick a number with some decimal places and ask yourself if that value makes sense. If it doesn't, it's probably discrete. For example, number of siblings is discrete (you can't have 2.3 siblings), but age is continuous (a number like 21.3 may not be how we usually share our age, but it is meaningful).

Section Exercises

For exercises 1-10, determine whether the variable is discrete numeric, continuous numeric, ordinal categorical, or nominal categorical.

species

temperature in Celsius

level of education

blood type

grams of flour in a cake recipe

political party

level to which a person agrees with some statement

number of siblings

number of cars that cross a bridge during rush hour

heart rate (beats per minute)

For exercises 11-16, refer to the following table showing part of the data matrix from an Intro Stats course survey. Note that some rows have been removed.

	Age	**Year in college**	**What is your major?**	**Units this semester**
1	19	Sophomore	Health Sciences	15
2	19	Sophomore	Business	15
3	19	Sophomore	Undecided	14
⋮	⋮	⋮	⋮	⋮
29	21	Junior	Business	15

What does each row of the data matrix represent?

What does each column of the data matrix represent?

How many observations are there in the full dataset?

What are the three other terms that mean the same thing as "observation"?

How many variables are there?

Indicate whether each variable in the data matrix is discrete numeric, continuous numeric, ordinal categorical, or nominal categorical.

For exercises 17-21, identify the population and the sample.

A survey of 2084 US households found that 45% have multiple TVs.

A local university wants to impose a new student fee in order to offer a better student rec center. They ask 87 students whether they support this fee.

A medical research company wants to test how well their new prosthetic hand design works for amputees. They recruit 27 amputees and have them wear the hands for a week, tracking their comfort and how well the hands respond to different cues.

A scientist wants to track the life cycles of invasive Burmese pythons in Florida. She spends a month in the field and tags 52 pythons for monitoring.

A college student wants to know if people living in her dorm actually enjoy the food offered at the dorm's food court. He approaches 18 students during meals at the food court and asks about their opinions.

In your own words, explain the differences between a population, a sample, and an observation.

Dig Deeper Read the article, Here's Why an Accurate Census Count Is So Important from the New York Times. (If you can't access the article, try a Google search for "why an accurate census count is important".) Take a moment to write down your thoughts on the relationship between how we collect data (for example - the questions asked in the census) and the power data has over people's lives. As researchers, scientists, and consumers of media, what are some reasons this is important to think about?

1.2 Statistical Sampling

How do we get samples? We want a sample that represents our population. **Representative samples** reflect the relevant characteristics of our population.

In general, we get representative samples by selecting our samples *at random* and with an adequate sample size.

A non-representative sample is said to be **biased**. For example, if we used a sample of chihuahuas to represent all dogs, we probably wouldn't get very good information; that sample would be *biased*.

These can be a result of **convenience sampling**, choosing a sample based on ease.

In our daily lives, common sources of bias are *anecdotal evidence* and *availability bias*. Anecdotal evidence is data based on personal experience or observation. Typically this consists of only one or two observations and is NOT representative of the population. For example, suppose a friend tells you their grandpa smoked a pack of cigarettes a day and lived to be 100. That may be entirely true, but it does not negate the fact that smoking is bad for your health.

Availability bias is your brain's tendency to think that examples of things that come readily to mind are more representative than is actually the case. For example, shark attacks are actually extremely uncommon, but the media tends to report on extreme anecdotes, making us more prone to this kind of bias! Anecdotal evidence is more directly connected to data, but both are important to be mindful of as responsible consumers of information.

1.2.1 Sampling Types

Simple Random Sampling

We avoid bias by taking random samples. One type of random sample is a **simple random sample**. We can think of this as "raffle sampling", like drawing names out of a hat. Each case (or each possible sample) has an equal chance of being selected. Knowing that A is selected doesn't tell us anything about whether B is selected. Instead of literally drawing from a hat, we usually use a **random number generator** from a computer.

Stratified Sampling

In a **stratified sample**, we break the population down into groups called **strata** based on characteristics we think might be relevant to our study. Individuals or items within a strata should be fairly similar to each other with respect to the outcome of interest. We then take a random sample from each strata. This ensures we have representation from each group.

> Example: A local politician believes men and women will vote differently on an upcoming ballot measure. She goes into the community and randomly samples 50 men and 50 women to ask for their thoughts on the ballot measure.

We can also use stratified sampling to make sure that the proportion of items in each group in the population matches the proportions in our sample.

> Example: A local high school is 29% freshmen, 27% sophomores, 24% juniors and 19% seniors. They want to collect a sample of 100 students using class level as strata. Since some class levels have more students than others, they set their strata to match: they select 29 freshmen (29% of their sample), 27 sophomores, 24 juniors, and 19 seniors.

This approach to stratified sampling can also help us ensure that small strata are adequately represented in our study.

> Example: Suppose we are doing some drug development research for a particular disease and know that a very small part of our population develops an especially severe form of the disease. In order to make sure those individuals are represented in our sample, we could treat disease severity as strata. The motivation in this case is that, in a simple random sample, we might miss those individuals entirely! By constructing these strata, we make sure they are accounted for.

Cluster Sampling

In a **cluster sample**, we break the population into **clusters**, where each cluster is similar to the population (and so the clusters are all similar to each other).

We then take a random sample of clusters and measure *all* items or individuals within each of those randomly selected clusters.

> Example: An airline wants to survey people who take its international flights from the United States to Asia. They randomly select 10 of these flights and give the survey to every individual on each of those 10 flights.

> Example: A farmer wants to know something about the plants in their fields. They randomly select 5 of their fields and examine all of the plants in each field.

The potential downside to cluster sampling is that there may be factors that make clusters meaningfully different from one other.

> Example: If our airline randomly samples flights to Asia, they may have failed to take into account that the people going to Vietnam are different from the people going to the Philippines who are different from the people going to India. That is, it's probably not reasonable to assume that every flight to Asia is representative of the entire population of individuals who use that airline to fly to Asia.

Systematic Sampling

In a **systematic sample**, we choose some starting point in our population and then collect every kth observation.

> Example: If I had a list of student ID numbers for every student at Sacramento State, I could generate a sample of students by selecting every 100th number.

> Example: Suppose we want to examine some machine part coming off on an assembly line. We collect a sample by pulling every 20th part off of the assembly line for additional testing.

One potential issue with systematic sampling is that there may be some pattern in the data.

> Example: A machine on an assembly line is oiled after producing 20 components, and its performance degrades steadily after it is oiled. If we select every 20 components, we match this *periodicity* and fail to capture a representative sample of components.

Section Exercises

For exercises 1-5, determine whether the sample is likely to be *biased*. Explain your thought process.

A professor wants to learn something about Sacramento State students, so they take a sample of all of their Stat 1 students.

A psychologist wants to draw conclusions about adults, and they run their study on a sample of randomly selected college students.

A pollster wants to draw conclusions about how people in California will vote on an upcoming ballot measure, so they select a random sample of 1000 likely voters in California.

Your roommate wants to know how many energy drinks other college students drink, so he asks all of his buddies to spend a week keeping track of how many energy drinks they consume.

A researcher wants to know how well some new medical intervention works, so they find volunteers willing to come into their clinic and participate in their study.

Exercises 6-11 each describe a statistical sampling scenario. Determine the statistical sampling type.

A spa wants to know how customers like their treatments, so they ask every tenth customer to fill out a satisfaction survey.

A landscaping company wants to know which suburbs they should advertise their services to. They get a list of all the suburbs in their area and randomly sample 10 households from each suburb to ask about their landscaping needs.

A city council wants to know whether its residents support building a new sports stadium. They get a list of all residents in the city and randomly call 200 of them.

A farmer wants to know which varieties of beans will yield the most product. She randomly selects five of her fields and examines the total bean yields for each of those five fields.

A researcher wants to know if a new medical intervention will help reduce the higher anesthetic requirement for redheads. His company has clinics in a number of US cities, so he randomly selects five of those clinics. Each clinic then asks all of their patients if they are willing to participate in the study.

To conduct an exit poll for voters at a specific voting location, a pollster stands outside the voting location and interviews every fifth person who exits the building.

For exercises 12-16, come up with a research scenario for the listed sampling type.

simple random sampling

stratified sampling

cluster sampling

systematic sampling

1.3 Experimental Design

When we do research, we have two options: run an experiment or an observational study.

1.3.1 Experiments

In an **experiment**, researchers assign **treatments** (experimental conditions) to cases.

> **Example** A biologist wants to know if different diets impact reproductive behaviors in mice. Of the 50 mice they have in the lab, 17 will be given Diet A, 17 will be given Diet B, and 16 Diet C. The biologist is going to provide each mouse with a specific diet, so this is an *experiment*. That is, they are assigning treatments (diets) to cases (mice).

> **Example** A medical researcher wants to know if a new heartburn medication is as effective as antacids. They bring 150 people into the lab and have them drink something that causes heartburn. After a set period of time, 75 of them are given an antacid and 75 are given the new heartburn medication. The researchers then measure how long it takes for each person's heartburn to subside. This is an *experiment* because the researchers provided each person with a treatment - an antacid or the new medication. That is, the researchers assigned a treatment to each subject.

In an experiment, cases may also be called **experimental units** (items or individuals on which the experiment is performed).

1.3.2 Observational Studies

In an **observational study**, no conditions are assigned. These are often done for ethical reasons, like examining the impacts of smoking cigarettes.

> **Example** A psychology researcher asked 100 people to take a survey on a variety of personality traits. Because the researcher did not assign any treatments to the subjects in the study (everyone took the same survey), this is an *observational study*.

> **Example** A researcher wanted to examine the relationship between cigarette smoking and stomach cancer. They follow 65 people from ages 40-70 and compare the stomach cancer rates of smokers and non-smokers. This is an observational study because the researcher did not *assign* treatments to cases. That is, each subject in the study was free to choose whether to smoke cigarettes. (If the reseachers found a strong relationship between smoking and stomach cancer, they would not be able to say that smoking *causes* stomach cancer, but they would have strong motivation for futher research!)

Importantly, experiments allow us to infer causation. Observational studies do not.

1.3.3 Experimental Design Principles

We have some additional things to think about for experiments, starting with our experimental design principles:

Control: two or more treatments are compared. We want to compare multiple treatments because it helps us be confident that our treatments are causing the effect we are observing. For example, if we wanted to know whether ibuprofen reduces pain from headaches, we would want to compare the use of ibuprofen to, for example, not taking any painkiller. This comparison allows us to confirm that any reduction in pain happened because of the ibuprofen, rather than the pain reduction being something that would have happened over time even without the drug.

Randomization: experimental units are assigned to treatment groups, usually and preferably at random. Essentially, we want each treatment group to look like a mini random sample and, just like with samples, that random assignment helps ensure that each group is representative of the population.

Replication: a large enough sample size is used to test each treatment many times (on many different experimental units). Perhaps the best way to think about why this is important is to think of the scenario where there is only one case in each treatment group. With such a small number in each group, we would have no way of knowing if the treatment is causing some effect or if any changes are happening by random chance. By selecting a larger sample size, we essentially "average out" the things that happen at random so that we can focus on the treatments themselves.

Blocking: if variables other than treatment are likely to have an impact on study outcome, we use blocks. Blocks give us a little bit of additional control over making sure that each treatment is representative of our population. For example, I might separate patients in a medical study into "high risk" and "low risk" blocks. I would randomly assign all of the high risk patients to a treatment and then randomly assign all of the low risk patients to a treatment. This helps ensure an even distribution of high/low risk patients in each treatment group.

An experiment without blocking has a completely randomized design; an experiment with blocking has a randomized block design.

In an experimental setting, we talk about

- **Response variable**: the characteristic of the experimental outcome being measured or observed.
- **Factor:** a variable whose impact on the response variable is of interest in the experiment.
- **Levels**: the possible values of a factor.

- **Treatments**: experimental conditions (based on combinations of factor levels).

  Example: Some entomologists are interested in the number of eggs laid by carrion beetles at different temperatures and different levels of moisture in the environment. They set up various enclosures for the beetles with temperatures either above or below freezing; and humidity levels of 10%, 50%, and 80%. After 24 hours in an enclosure, they check how many eggs were laid by each beetle.

  In this scenario, the entomologists are interested in observing how different things impact number of eggs laid, so this is the response variable. The factors are temperature and humidity, the two variables our researchers think will impact that response variable. For humidity, the levels are 20%, 50%, and 80%; for temperature, the levels are "below freezing" and "above freezing".

  To get at the treatments, we need to consider all possible combinations of factor levels. That is, we need to think about all of the possible ways to combine the different temperatures and humidities:

  1. 20% humidity, below freezing.
  2. 50% humidity, below freezing.
  3. 80% humidity, below freezing.
  4. 20% humidity, above freezing.
  5. 50% humidity, above freezing.
  6. 80% humidity, above freezing.

  There are six different combinations, which make up the six different treatments in this experiment.

In human subjects research, we do a little extra work. If subjects do not know what treatment group they are in, the study is called **blind**. We use a **placebo** (fake treatment) to achieve this. We do this because, psychologically, people's expectations for their outcome (their idea of what is going to happen to them) has a strong impact on how they actually do. This is called the placebo effect. If neither the subjects nor the researchers who interact with them know the treatment group, it is called **double blind.** This helps avoid bias caused by researcher's expectations for outcome. This can happen when, for example, a person does not know what treatment group they are in, but a doctor knows they are getting a fake treatment and acts as if they may have a bad outcome.

Section Exercises

A group of researchers wanted to know if puppies have an effect on heart rate. From a sample of 18 people, they randomly assigned 10 to take a test while in a room with a puppy. The remaining 8 people took the same test in a room with no puppies. During the test, each participant's heart rate was monitored.

Is this an observational study or an experiment? Explain.

Identify the (i) cases and (ii) response variable.

Explain how this study satisfies the principles of (i) control, (ii) randomization, and (iii) replication.

Can this study be used to infer causality (cause and effect)? Why or why not?

A group of researchers wanted to examine the relationship between smoking and type 2 diabetes. In a sample of 200 people, 47 were smokers and 153 were non-smokers. The researchers followed both groups for 10 years and tracked whether they developed diabetes.

Is this an observational study or an experiment? Explain.

Can this study be used to infer causality? Why or why not?

A scientist wanted to examine how different conditions impacted the life cycle of monarch butterflies. Butterfly cocoons were placed individually into enclosures at three different temperatures: 70 degrees, 90 degrees, and 110 degrees; and two different levels of humidity: high (85%) and low (15%). The scientist then measured how long each cocoon took to develop into a butterfly.

Is this an observational study or an experiment? Explain.

Identify the (i) cases and (ii) response variable.

Identify the factors. What are the levels of each factor?

What are the possible treatments in this experiment? (Hint: there should be 6 combinations of factor levels.)

Explain how this study satisfies the principles of (i) control, (ii) randomization, and (iii) replication.

Can this study be used to infer causality? Why or why not?

A drug manufacturer wants to know how well a new blood pressure medication works. They recruit 130 people with high blood pressure and randomly assign 65 of them to the new medication; the other 65 will receive an existing blood pressure medication that is known to work well. Researchers will then monitor each participant's blood pressure over time to determine whether the new medication works at least as well as existing medications.

Is this an observational study or an experiment? Explain.

Identify the (i) cases and (ii) response variable.

Explain how this study satisfies the principles of (i) control, (ii) randomization, and (iii) replication.

What would need to be true for this study to be *blind*?

What would need to be true for this study to be *double blind*?

Why do you think the researchers chose not to use a placebo in this study?

1.4 Frequency Distributions

1.4.1 Qualitative Variables

Frequency (**count**): the number of times a particular value occurs. Suppose we have the following data for the class level of students in a section of Introductory Statistics:

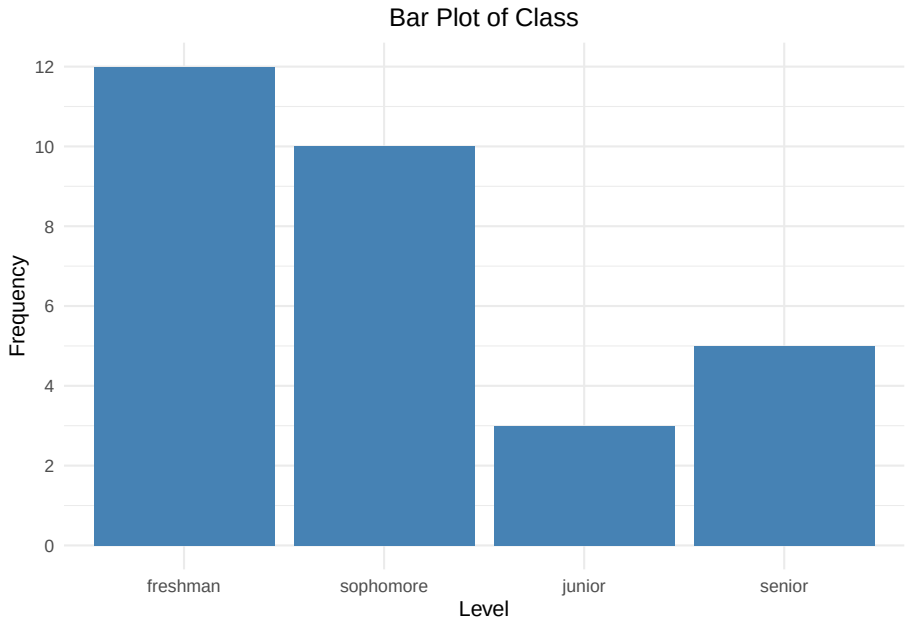
```
sophomore, freshman, freshman, sophomore, sophomore, senior,
sophomore, freshman, senior, sophomore, freshman, junior, freshman,
freshman, senior, sophomore, sophomore, freshman, sophomore,
junior, freshman, sophomore, junior, freshman, senior, freshman,
freshman, senior, freshman, sophomore
```

This is a lot to take in at a glance, so we are going to think about ways to summarize it. A **frequency distribution** lists each distinct value with its frequency.

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

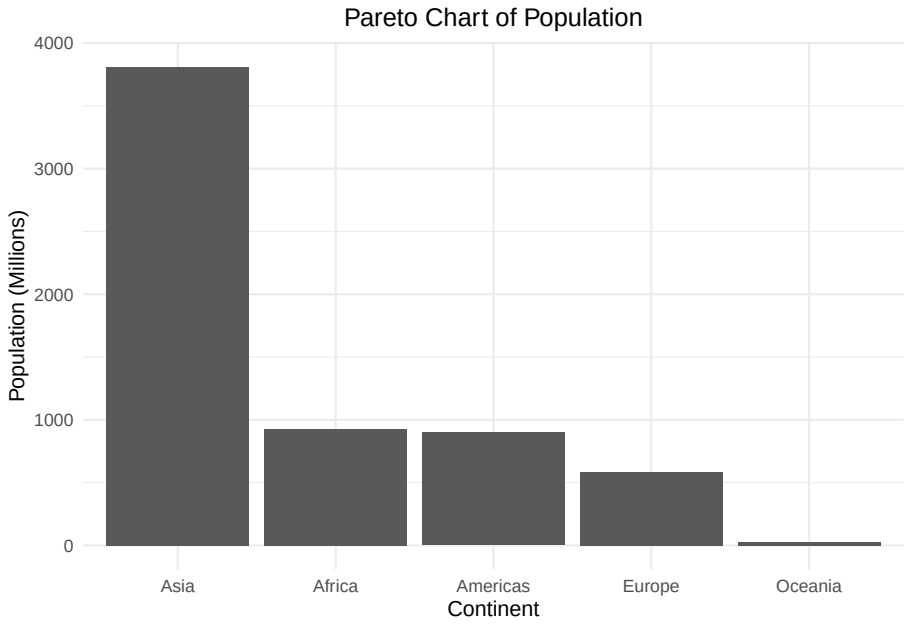
Note that I can also quickly get the total number of students in the class from this frequency distribution; since all students are accounted for in the data, the total number of students is $12 + 10 + 3 + 5 = 30$.

A **bar plot** is a graphical representation of a frequency distribution. Each bar's height is based on the frequency of the corresponding category.



The bar plot above shows the class level breakdown for students in an Introductory Statistics course. Take a moment to notice how the bars match up with the frequency distribution above. Since class level is an ordinal variable, we should match the order of the bars to the order of the categories.

For nominal variables, we can use a **Pareto chart**, which is a bar chart with the bars sorted from highest to lowest frequency.



This allows us to quickly examine which categories appear most and least frequently, as well as how their frequencies compare to each other.

Relative frequency is the ratio of the frequency to the total number of observations.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

This is also called the **proportion**. The **percentage** can be obtained by multiplying the proportion by 100.

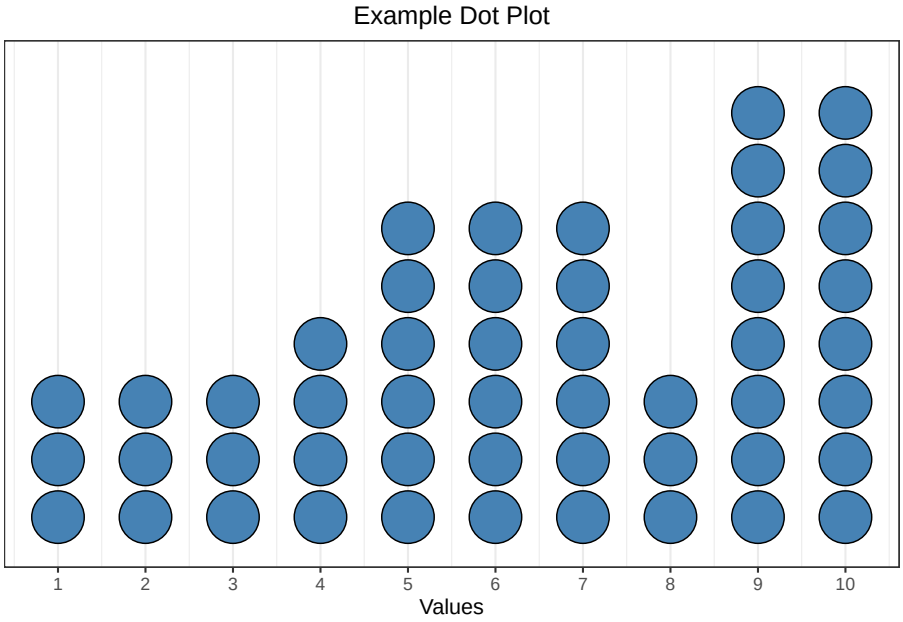
A **relative frequency distribution** lists each distinct value with its relative frequency.

Class	Frequency	Relative Frequency	Percent
freshman	12	$12/30 = 0.4$	40%
sophomore	10	$10/30 \approx 0.3333$	33.33%
junior	3	$3/30 = 0.1$	10%
senior	5	$5/30 \approx 0.1667$	16.67%

Notice that if we add up all of the relative frequencies, we get 1. Equivalently, if we add all of the percents, we get 100%. This total represents 100% of the students in this course.

1.4.2 Quantitative Variables

We can also apply this concept to numeric data. A **dot plot** is one graphical representation of this. A dot plot shows a number line with dots drawn above the line. Each dot represents a single point.



For example, the dot plot above shows a sample where the value 1 appears three times, the value 5 appears six times, etc.

We would also like to be able to visualize larger, more complex data sets. This is hard to do using a dot plot! Instead, we can do this using **bins**, which group numeric data into equal-width consecutive intervals.

Example: A random sample of weights (in lbs) from 12 cats:

6.2 11.6 7.2 17.1 15.1 8.4 7.7 13.9 21.0 5.5 9.1 7.3

The **minimum** (smallest value) is 5.5 and the **maximum** (largest value) is 21. There are lots of ways to break these into "bins", but what about…

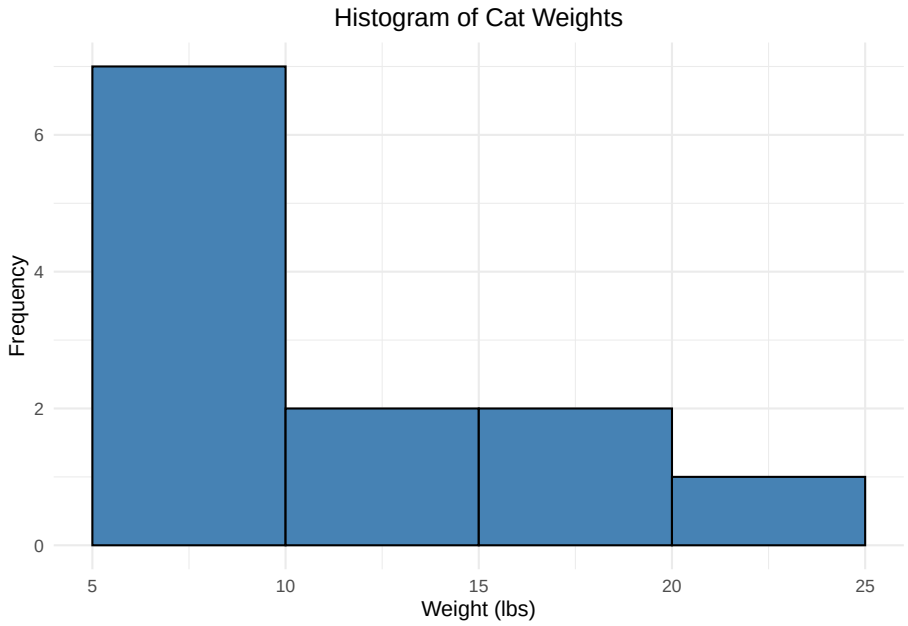
- 5 - 10
- 10 - 15
- 15 - 20
- 20 - 25

Each bin has an equal width of 5, but if we had a cat with a weight of exactly 15 lbs, would we use the second or third bin?? It's unclear. To make this clear, we need there to be no overlap. Instead, we could use:

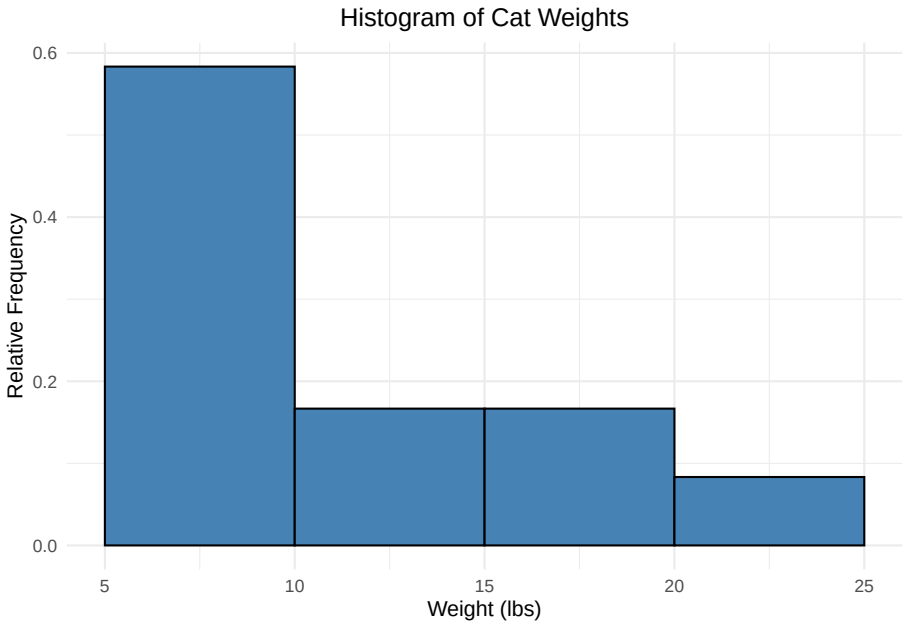
Weight	Count
[5, 10)	7
[10, 15)	2
[15, 20)	2
[20, 25)	1

Now, a cat with a weight of 15.0 lbs would be placed in the third bin (but not the second). (Recall that the interval notation $[5, 10)$ means $5 \leq x < 10$.)

We will visualize this using a **histogram**, which is a lot like a bar plot but for numeric data:

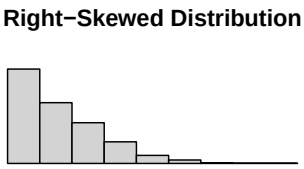


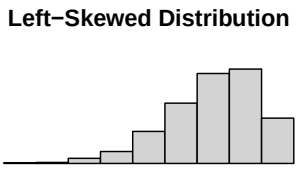
This is what we call a **frequency histogram** because each bar height reflects the frequency of that bin. We can also create a **relative frequency histogram** which displays the relative frequency instead of the frequency:

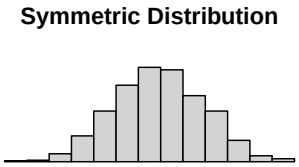


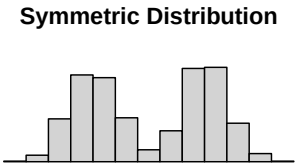
Notice that these last two histograms look the same *except for the numbers on the vertical axis*! This gives us insight into the shape of the data **distribution**, literally how the values are distributed across the bins. The part of the distribution that "trails off" to one or both sides is called a **tail** of the distribution.

When a histogram trails off to one side, we say it is **skewed** (right-skewed if it trails off to the right, left-skewed if it trails off to the left). Data sets with roughly equal tails are **symmetric**.

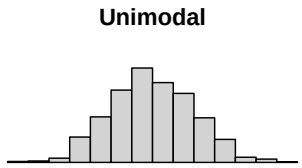


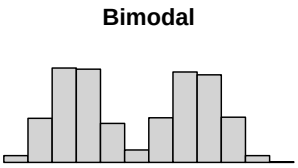


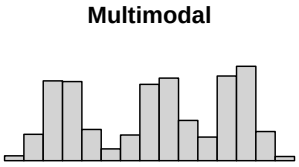


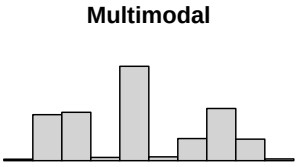


We can also use a histogram to identify **modes**. For numeric data, especially continuous variables, we think of modes as *prominent peaks*.





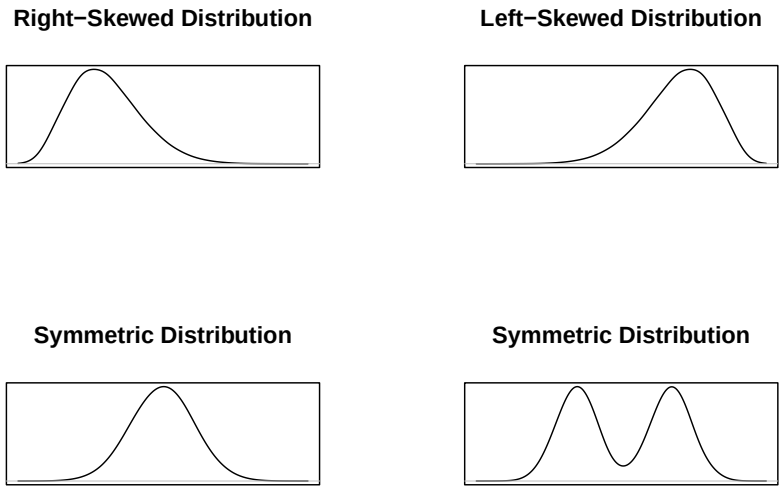




- **Unimodal**: one prominent peak.
- **Bimodal**: two prominent peaks.

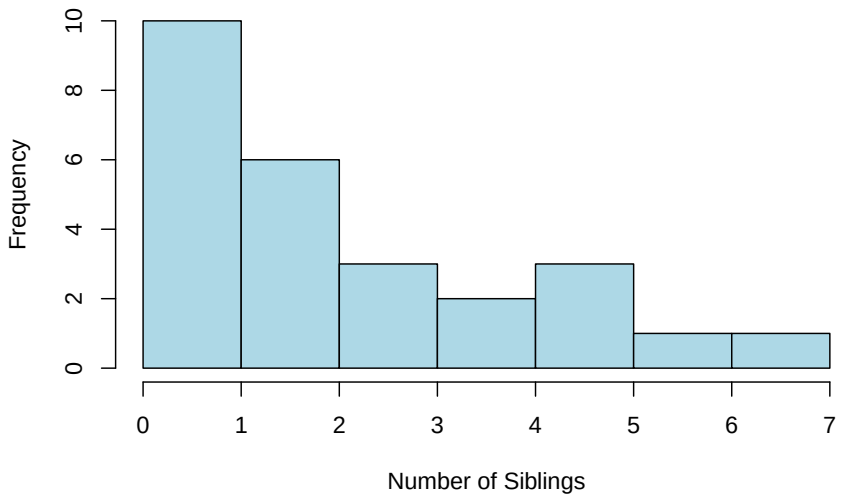
- **Multimodal**: three or more prominent peaks.

Finally, we can also "smooth out" these histograms and use a smooth curve to examine the shape of the distribution. Below are the smooth curve versions of the distributions shown in the four histograms used to demonstrate skew and symmetry.

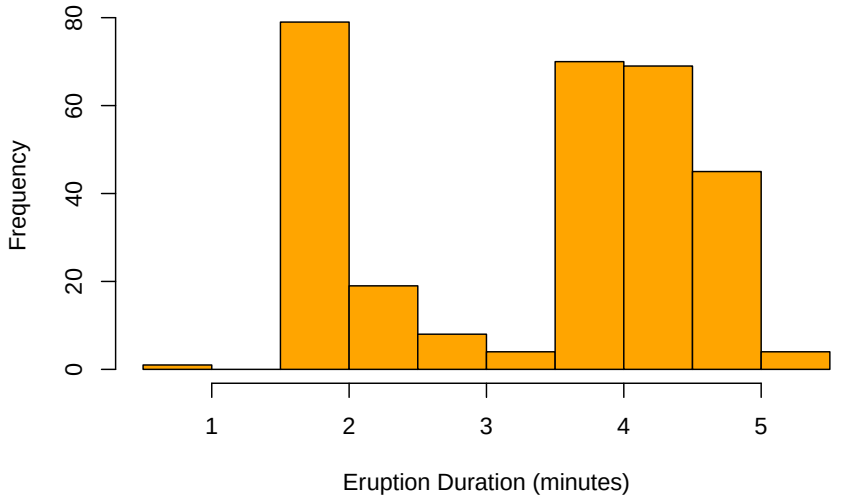


Section Exercises

Twenty-five Stat 1 students were asked how tall they were. A histogram of their responses is shown below. Describe the shape (modality and skew/symmetry) of this distribution.

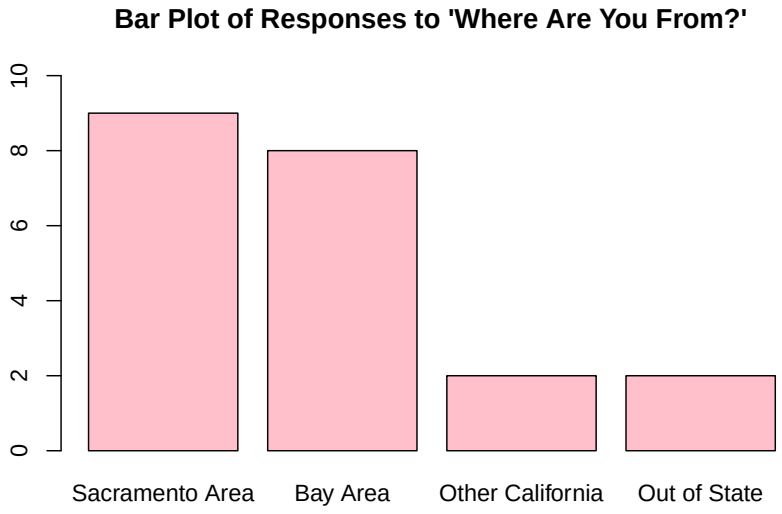


The following histogram shows the duration (in minutes) for 299 eruptions of Old Faithful Geyser in Yellowstone National Park. Describe the shape of this distribution



For exercises 3-5, twenty-one Stat 1 students were asked where they were from.

A bar chart of their responses is shown below.



How many people in this class are from the Bay Area?

Use the bar plot to create the frequency distribution for these data.

Use your frequency distribution from (a) to find the relative frequency distribution for these data.

For exercises 6-8, use the following data on student eye color:

`brown, brown, blue, green, blue, brown, brown, grey, brown, blue, brown, green, brown, blue, brown, brown, brown, blue, blue, brown`

Construct a frequency distribution for these data.

Construct a relative frequency distribution.

Create a Pareto chart for these data.

R Lab: Data Basics and Graphs

R as a Calculator

Let's start simple. Type in `2+2` and click the "Run" button in the top left panel. The answer to $2 + 2$ should appear in the bottom left panel under the line of code you just ran. It will look something like

$> 2+2$

```
[1] 4
```

Basically, using R as a calculator works the same way as the scientific calculator you may have used in math classes. That is, R follows the traditional order of operations. However, some of the operators may be a little different from what you're used to.

- Addition and subtraction are as you would expect: `3+5` will give the solution to $3 + 5$ and `6-4` the solution to $6 - 4$.
- We use an asterisk for multiplication: `3*4` will give the solution to 3×4.
- For division, we use a forward slash: `6/2` gives the solution to $6 \div 2$.
- Finally, for exponents, we use a caret: `7^3` gives the solution to 7^3.
- For the square root, we have a special command: `sqrt(9)` gives the solution to $\sqrt{9}$.

By default, R will always produce either the whole number result or a decimal. That's what we want in this class!

Try entering each of the commands given above in R, pressing the green "Run" button after each one. Notice that you can either delete everything in the box and then do a new calculation, or you can put your new calculation on the next line:

```
6-4
3*4
7^3
```

Try copy and pasting the three lines above into the top left panel and then take a moment to notice what the output looks like and how it matches up with the lines of code you entered.

Your Turn

1. For each of the following mathematical expressions, provide an R expression you could write to find the solution.
   a. 7^{11}
   b. 17×9
   c. $\sqrt{49}$

We can also do much more extensive calculations in R, but we need to be very careful with our order of operations! If in doubt, break your equation up and do it piece by piece. For example, consider the expression

$$\frac{7-4}{5/\sqrt{10}}$$

I can put this entire thing into R as `(7-4)/(5/sqrt(10))` but that requires a bunch of parenthesis to get the order of operations right!

Another option is to break this down. I start with `7-4` to get a value of 3 in the numerator. Then, I can find `5/sqrt(10)` separately, which is 1.581. Finally, I

would enter `3/1.581` to get my final answer of 1.897. (I would do the same thing with a scientific calculator if I weren't 100% comfortable with my parentheses!)

Your Turn

2. For each of the following mathematical expressions, provide an R expression you could write to find the solution.
   a. $4 \times 7 - 3$
   b. $3^5 + 2 \times 2$
   c. $\frac{4.5-2.3}{1.75}$

Random Number Generation

To generate a random whole number using R, we can use the `sample` command. We use the `sample` command like `sample(minimum:maximum, size = n)`, replacing `minimum` with the minimum value (often the number 1), `maximum` with the maximum value, and `n` with the sample size.

The following command takes a random sample of size 1 from the values 1 through 10 (the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10):

```
sample(1:10, size = 1)
```

which results in the output

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

Your Turn

3. Provide an R expression you could use to generate a random sample of size $n = 10$ with minimum value 1 and maximum value 100.
4. Provide an R expression you could use to generate a random number ($n = 1$) between 8 and 20.

Entering Data

We can work with data in R by reading it in from a file or by entering it manually. To enter numeric data manually, we use the `c` command.

The following line of code saves the `ages` data from the data matrix example above:

```
ages = c(45, 23, 36, 29)
```

Notice that we set `ages` equal to `c()` with the numbers in the parentheses, separated by commas. Also notice that the numbers are in the same order as in the data. If I want to use the `ages` variable later, I can refer to it directly in R and it will print out the values in that variable:

```
ages = c(45, 23, 36, 29)
```

Your Turn

5. Provide an R expression for entering the following data.
   a. The variable `pets` has values $1, 0, 2, 1, 1, 0, 2, 3, 4$.
   b. The variable `height` has values $58.2, 69.1, 74.5, 66.0, 62.4, 64.8, 71.5$
6. Provide an R expression for entering the following data. You will need to decide on appropriate variable names.
   a. The number of days per week students go to school resulted in the data $3, 5, 4, 5, 5, 3, 3, 4$.

To enter categorical data in R, we do the same thing, but with the addition of quotation marks:

```
gender = c("Male", "Female", "Other", "Female")
```

Notice that for every variable I entered data for in R, I gave it a *single word name*. That's important! R will not recognize spaces. However, R does recognize upper versus lowercase letters! In R, `age` is different from `Age`.

Your Turn

7. Provide an R expression for entering the following data.
   a. The variable `smoker` has values yes, no, yes, yes, no.
8. Provide an R expression for entering the following data. You will need to decide on an appropriate variable name.
   a. The level of a sample of college students resulted in the data freshman, freshman, sophomore, senior, freshman.

Usually may want to use data without entering it by hand in R. In this class, we will do this in two ways. The first is by using datasets that are built into R. One such dataset is called

To access and this data in R, we enter the following command:

```
data(Loblolly)
```

The final way is what we use most often in practice. When we do data analysis in the real world, often our data is stored in an Excel, csv, or similar spreadsheet-type file. For this class, when we use external data, we will only use data stored in csv files. To read in a csv file, we use the command `read.csv`. For example

```
read.csv("C:\Users\perry\Documents\Courses\STAT 1\dataset.csv")
```

Reads in a file located on my computer. The stuff inside the quotation marks is a *filepath*, which tells R both which file I want (`dataset.csv`) and where the file is located (`C:\Users\perry\Documents\Courses\STAT 1\`). We can do something similar with csv files stored online, which is how we will use external

datafiles in this course. Further, for this course, I will always provide you with the line of code you need to read in any external files.

Histograms

There is a built-in dataset in R called `Loblolly`, which contains the variables `height` and `age` of some Loblolly pine trees. I can refer to this data by typing in `Loblolly` directly. To view just the first few observations (out of the 84 total in the data), I can use the `head` command:

```
head(Loblolly)
```

```
##   height age
## 1   4.51   3
## 2  10.89   5
## 3  28.72  10
## 4  41.74  15
## 5  52.70  20
## 6  60.92  25
```

The information that appears next to each ## is what R prints out for us.

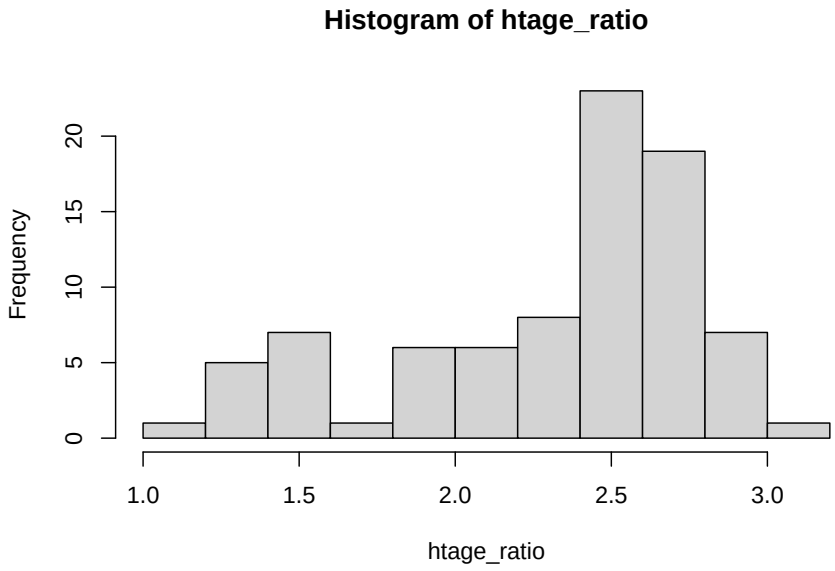
In order to refer to the variables in `Loblolly` directly, I will need to use the `attach` command. This tells R that when I say `age` I mean the age variable from the `Loblolly` dataset (and not from some other dataset).

I want to create a histogram to visualize the ratio of tree height to age. First, I need to find this ratio for each observation. I can do this easily in R by dividing `height` by `age`. I will save this as a new variable called `htage_ratio`.

```
htage_ratio = height/age
```

Then to create a histogram of the height to age ratio, we will use the command `hist` on the variable `htage_ratio`:

```
hist(htage_ratio)
```

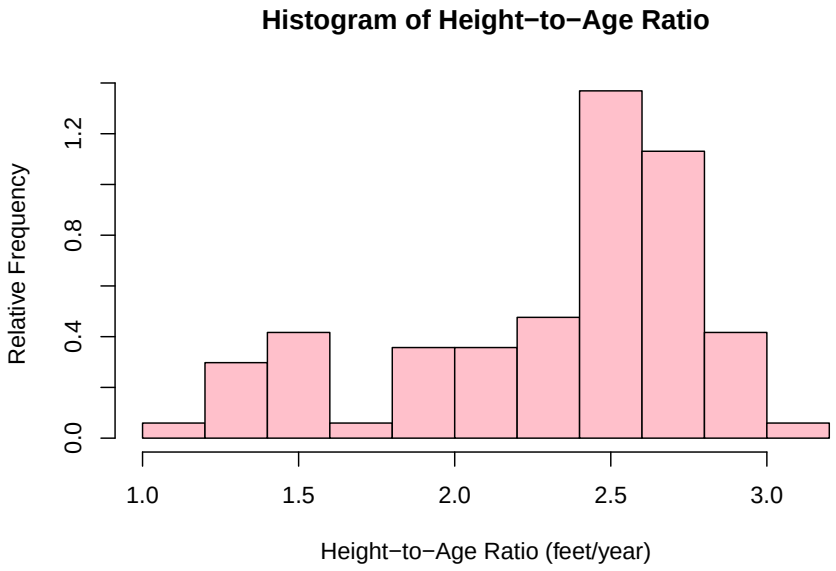


I can clean up this graph by taking advantage of additional *arguments* in the `hist` command:

- `main` is where I can give the plot a new title. (Make sure to put the title in quotes!)
- `xlab` is the x-axis (horizontal axis) title.
- `ylab` is the y-axis (vertical axis) title.
- `freq` allows us to create either frequency or *relative frequency* histograms.
  - If we set it equal to `TRUE` it will produce a frequency histogram. (This is the default if we don't give R any instructions.)
  - If we set it equal to `FALSE` it will produce a relative frequency histogram.
- `col` allows us to give R a specific color for the bars.

Notice that each argument is separated by a comma.

```
hist(htage_ratio,
     main = "Histogram of Height-to-Age Ratio",
     xlab = "Height-to-Age Ratio (feet/year)",
     ylab = "Relative Frequency",
     freq = FALSE,
     col = 'pink')
```



When I am done, I will use the `detatch` command to tell R that I am not working with the `Loblolly` data anymore.

```
detach(Loblolly)
```

Your Turn

9. Provide an R expression to create a histogram of the `height` variable in the `Loblolly` pine tree data. Copy and paste this histogram into your lab solutions.
10. Provide an R expression to create a histogram of the `age` variable in the `Loblolly` pine tree data. Give your histogram an appropriate main title and vertical axis title, and choose a color for the bars. Copy and paste this histogram into your lab solutions.

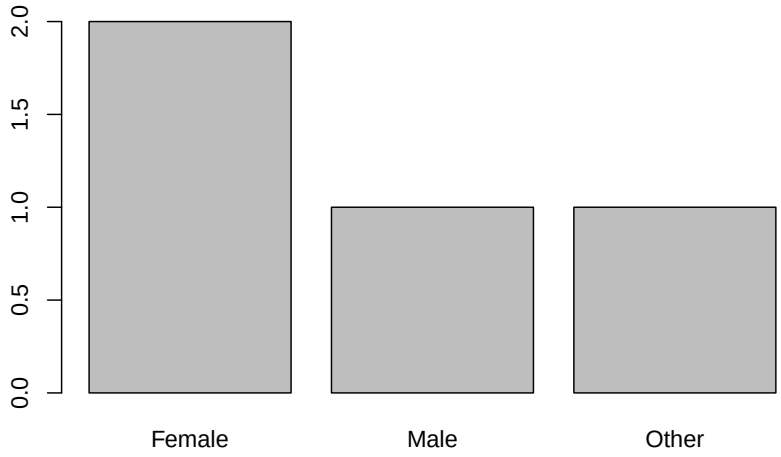
Bar Plots

To create a bar plot, we will begin by asking R to generate a frequency table. We do this using the `table` command. By default, this command shows the categories in alphabetical order. That is fine.

Earlier, we created this `gender` variable. Let's use it to create a frequency table:

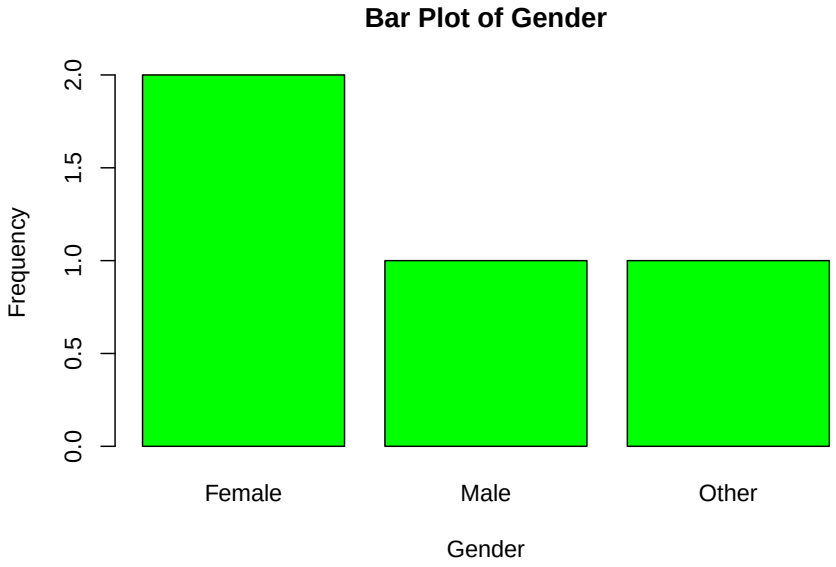
```
## gender
## Female   Male  Other
##      2      1      1
```

To make a bar plot, we need to put that table command into the barplot command. Here's what that looks like:



To do this with a different variable, I would change out `gender` for the other variable. Everything else stays the same!

The customization for the bar plot is essentially the same as for histograms:


Bar Plot of Gender
Frequency
0.0
0.5
1.0
1.5
2.0
Female
Male
Other
Gender

Chapter 2

Descriptive Measures

In the previous module, we thought about descriptive statistics using tables and graphs. Next, we summarize data by computing numbers. Some of these numbers you may already be familiar with, such as averages and percentiles. Numbers used to describe data are called *descriptive measures*.

Module Learning Objectives/Outcomes

After completing Module 2, you will be able to:

1. Calculate and interpret measures of center.
2. Calculate and interpret measures of variation.
3. Find and interpret measures of position.
4. Summarize data using box plots.

R objectives

1. Generate measures of center.
2. Generate measures of variability.
3. Generate measures of position.
4. Create box plots.

This module's outcomes correspond to course outcomes (1) organize, summarize, and interpret data in tabular, graphical, and pictorial formats, (2) organize and interpret bivariate data and learn simple linear regression and correlation, and (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation.

2.1 Measures of Central Tendency

One research question we might ask is: what values are most common or most likely?

Mode: the most commonly occurring value. We can use this for numeric variables, but typically we use the mode when talking about categorical data.

Mean: this is what we usually think of as the "average". Denoted $\bar{x}$. Add up all of the values and divide by the number of observations (n):

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

where x_i denotes the ith observation and $\sum_{i=1}^{n}$ is the sum of all observations from 1 through n. This is called *summation notation.*

Median: the middle number when the data are ordered from smallest to largest.

- If there are an odd number of observations, this will be the number in the middle:
  {1, 3, **7**, 9, 9} has median 7
- If there are an even number of observations, there will be two numbers in the middle. The median will be their average.
  {1, 2, **4, 7**, 9, 9} has median $\frac{4+7}{2} = 5.5$

The mean is sensitive to extreme values and skew. The median is not!

x: 1, 3, 7, 9, 9

y: 1, 3, 7, 9, 45

Median

median = 7

median = 7

Mean

$\bar{x} = \frac{29}{5} = 5.8$

$\bar{y} = \frac{65}{5} = 13$

Notice how changing that 9 out for a 45 changes the *mean* a lot! But the *median* is 7 for both x and y.

If the mean and median are roughly equal, it is reasonable to assume the distribution is roughly symmetric. Since calculating the mean involves adding all the values, the mean will get pulled toward any extremes (and away from the median).

Because the median is not affected by extreme observations or skew, we say it is a **resistant measure** or that it is **robust**.

Which measure should we use?

- Mean: symmetric, numeric data
- Median: skewed, numeric data

- Mode: categorical data

2.1.1 Weighted Means

Sometimes we have reason to calculate a *weighted mean*. For example, course grades are often calculated using a grading scheme that weights each category.

$$\bar{x}_w = w_1 x_1 + w_2 x_2 + ... w_n x_n$$

In this case, each p represents the proportion attributed to that category. In general, we require that all of the w values sum to 1.

Example Consider the following grade distribution:

- Assignments: 15%
- Quizzes: 20%
- Exam 1: 15%
- Exam 2: 15%
- Project: 15%
- Final Exam: 20%

We can use this to calculate an overall grade. Suppose some student has the following score in each category

- Assignments: 92%
- Quizzes: 76%
- Exam 1: 56%
- Exam 2: 69%
- Project: 89%
- Final Exam: 70%

We can calculate their overall grade in the class using the weighted average formula.

$$\begin{aligned}\text{grade} &= 92(0.15) + 76(0.20) + 56(0.15) + 69(0.15) + 89(0.15) + 70(0.20) \\ &= 13.8 + 15.2 + 8.4 + 10.35 + 13.35 + 14 \\ &= 75.1\end{aligned}$$

So this student would get a 75.1% in the class. (Notice that we changed the grade distribution values from percents to proportions!)

This is how learning management systems like Canvas automatically calculate grades.

Example: Now suppose a student has the following scores

- Assignments: 83%
- Quizzes: 71%

- Exam 1: 61%
- Exam 2: 68%
- Project: 91%

and has not taken the final exam yet. He really wants to pass the class with at least a C-, but is not sure what kind of final exam grade would allow him to do that.

If he wants to pass, he needs a minimum overall grade of 70%. So he needs his weighted average to be 70% or higher. We know everything except his final exam score, so we'll make that F in our formula:

$$70 = 83(0.15) + 71(0.20) + 61(0.15) + 68(0.15) + 91(0.15) + F(0.20)$$

To figure out what he needs to get on the final, we need to solve for F.

$$\begin{aligned} 70 &= 83(0.15) + 71(0.20) + 61(0.15) + 68(0.15) + 91(0.15) + F(0.20) \\ 70 &= 12.45 + 14.2 + 9.15 + 10.2 + 13.65 + 0.2 \\ 70 &= 59.65 + 0.2F \\ 10.35 &= 0.2F \\ F &= 51.75 \end{aligned}$$

So he needs to get at least a 51.75% on the final exam in order to pass the class.

Section Exercises

For exercises 1-5, calculate the mean and median. Then, determine which measure of center should be used to describe the data. Explain your thought process.

$9, 2, 7, 3, 5$

$56, 87, 21, 95, 236$

$1.3, 2.4, 1.5, 2.1, 3.9, 3.2, 0.9$

$4, 6, 2, 4, 5, 5, 7, 2, 5$

$345, 654, 234, 123, 432, 152$

Consider the following frequency distribution for students in a class.

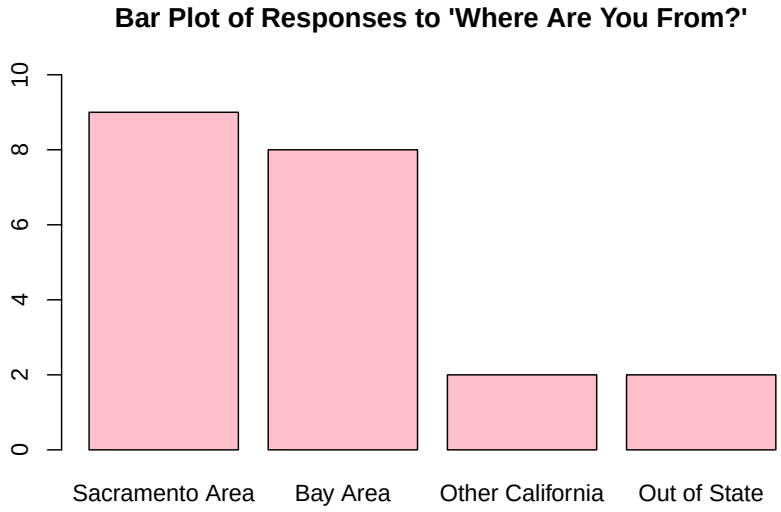
Class	Frequency
freshman	8

Class	Frequency
sophomore	17
junior	21
senior	3

How many students are in this class?

What is the mode of these data?

A class of intro stats students were asked where they were from. A bar chart of their responses is shown below.



How many students are in this class?

What is the mode of these data?

Twenty-five Stat 1 students were asked how many siblings they have. The following data shows their responses:

$$5, 1, 7, 3, 2, 0, 1, 4, 3, 2, 0, 2, 2, 0, 0, 5, 1, 1, 1, 5, 2, 3, 2, 4, 6$$

Calculate the mean number of siblings.

Determine the median number of siblings.

What do the mean and median tell you about the skew/symmetry of this distribution? Based on this assessment, which measure of center would be the best choice for describing these data?

A professor is teaching four sections of the same class and wants to know the overall average course grade of her students.

	Class 1	Class 2	Class 3	Class 4
Mean grade	78	81	76	80
Number of students	27	31	33	30

We can weight these grades using the number of students in the class. Convert the number of students to proportions by dividing each number by the total number of students.

Using your proportions from (a) as the weights, calculate the weighted mean grade for the professor's four classes.

2.2 Measures of Variability

How much do the data vary?

Should we care? Yes! The more variable the data, the harder it is to be confident in our measures of center!

If you live in a place with extremely variable weather, it is going to be much harder to be confident in how to dress for tomorrow's weather… but if you live in a place where the weather is always the same, it's much easier to be confident in what you plan to wear.

We want to think about how far observations are from the measure of center.

One easy way to think about variability is the **range** of the data:

$$\text{range} = \text{maximum} - \text{minimum}$$

This is quick and convenient, but it is *extremely* sensitive to outliers! It also takes into account only two of the observations - we would prefer a measure of variability that takes into account *all* the observations.

2.2.1 Standard Deviation

Deviation is the distance of an observation from the mean: $x - \bar{x}$. If we want to think about how far - on average - a typical observation is from the center, our intuition might be to take the average deviance… but it turns out that summing up the deviances will *always* result in 0! Conceptually, this is because the stuff below the mean (negative numbers) and the stuff above the mean (positive numbers) end up canceling each other out until we end up at 0. (If you are interested, *Appendix A: Average Deviance* has a mathematical proof of this using some relatively straightforward algebra.)

One way to deal with this is to make all of the numbers positive, which we accomplish by squaring the deviance.

	Deviance	Squared Deviance
x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-1.2	1.44
5	1.8	3.24
3	-0.2	0.04
4	0.8	0.64
2	-1.2	1.44
$\bar{x} = 3.2$	Total = 0	Total = 6.8

Variance (denoted s^2) is the average squared distance from the mean:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where n is the sample size. Notice that we divide by $n-1$ and NOT by n. There are some mathematical reasons why we do this, but the short version is that it'll be a better estimate when we talk about inference.

Finally, we come to **standard deviation** (denoted s).

$$s = \sqrt{s^2}$$

The standard deviation is the square root of the variance. We say that a "typical" observation is within about one standard deviation of the mean (between $\bar{x} - s$ and $\bar{x} + s$).

In practice (including in this class), we will use a computer to calculate the variance and standard deviation.

2.2.2 The Interquartile Range

The **interquartile range** (**IQR**) represents the middle 50% of the data.

Recall that the *median* cut the data in half: 50% of the data is below and 50% is above the median. This is also called the **50th percentile**. The p**th percentile** is the value for which p% of the data is below it.

To get the middle 50%, we will split the data into four parts:

1	2	3	4
25%	25%	25%	25%

The 25th and 75th percentiles, along with the median, divide the data into four parts. We call these three measurements the **quartiles**:

- **Q1**, the first quartile, is the median of the lower 50% of the data.
- **Q2**, the second quartile, is the median.
- **Q3**, the third quartile, is the median of the upper 50% of the data.

Example: Consider {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

- Cutting the data in half: {1, 2, 3, 4, 5 | 6, 7, 8, 9, 10}, the median (Q2) is $\frac{5+6}{2} = 5.5$.
- Q1 is the median of {1, 2, 3, 4, 5}, or 3
- Q3 is the median of {6, 7, 8, 9, 10}, or 8

Note: this is a "quick and dirty" way of finding quartiles. A computer will give a more exact result.

Then the interquartile range is

$$\text{IQR} = \text{Q3} - \text{Q1}$$

This is another measure of variability and is resistant to extreme values.

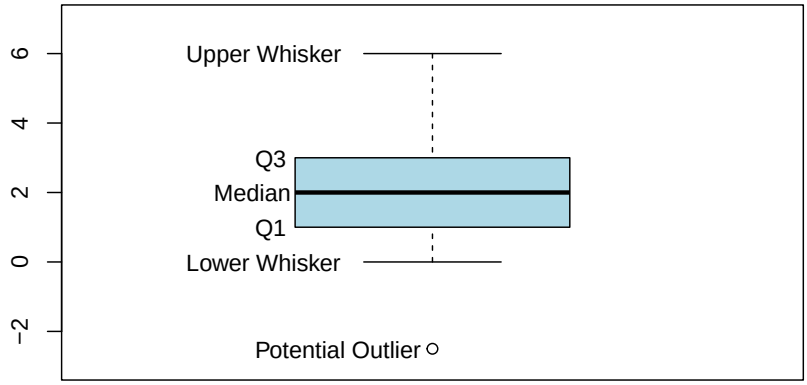
Which measure should we use?

- Mean and standard deviation: symmetric, numeric data
- Median and IQR: skewed, numeric data

Notice that we do not have a measure of variability for categorical data. Instead, if we want to think about variability for categorical data, we can use a frequency table or bar plot to visualize this variability.

2.2.3 Box Plots

Our quartiles are the foundation for constructing what we call a box plot, which summarizes the data with five five statistics plus extreme observations. These statistics are sometimes referred to collectively as the "five number summary": the minimum, Q1, the median (Q2), Q3, and the maximum.



Drawing a box plot:

1. Draw the vertical axis to include all possible values in the data.
2. Draw a horizontal line at the median, at Q1, and at Q3. Use these to form a box.
3. Draw the **whiskers**. The whiskers' upper limit is $Q3 + 1.5 \times IQR$ and the lower limit is $Q1 - 1.5 \times IQR$. The actual whiskers are then drawn *at the next closest data points within the limits.*
4. Any points outside the whisker limits are included as individual points. These are **potential outliers**.

(Potential) outliers can help us…

- examine skew (outliers in the negative direction suggest left skew; outliers in the positive direction suggest right skew).
- identify issues with data collection or entry, especially if the value of an outlier doesn't make sense.

As with most things in this text, we won't draw a lot of boxplots by hand. However, understanding how they are drawn will help us understand how to interpret them!

Section Exercises

Twenty-five Stat 1 students were asked how many siblings they have. The following data shows their responses:

$$5, 1, 7, 3, 2, 0, 1, 4, 3, 2, 0, 2, 2, 0, 0, 5, 1, 1, 1, 5, 2, 3, 2, 4, 6$$

Use a computer to calculate the variance of the number of siblings.

Calculate the standard deviation of number of siblings.

What do the mean and standard deviation tell you about how many siblings a "typical" Stat 1 student has?

Twenty-five Stat 1 students were asked how many siblings they have. The following data shows their responses:

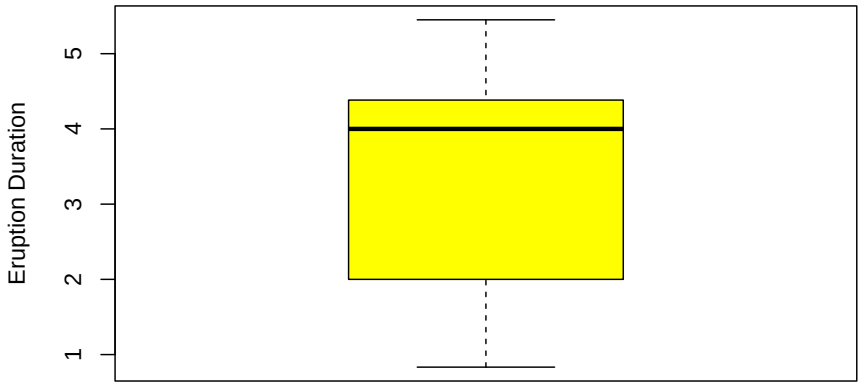
$$5, 1, 7, 3, 2, 0, 1, 4, 3, 2, 0, 2, 2, 0, 0, 5, 1, 1, 1, 5, 2, 3, 2, 4, 6$$

Give the five number summary for these data.

What is the IQR?

Refer to the histogram of these data in Section 1.4, Exercise 1. Is the IQR or the standard deviation a better measure of spread for these data? Explain your reasoning.

Draw or print out the following boxplot.



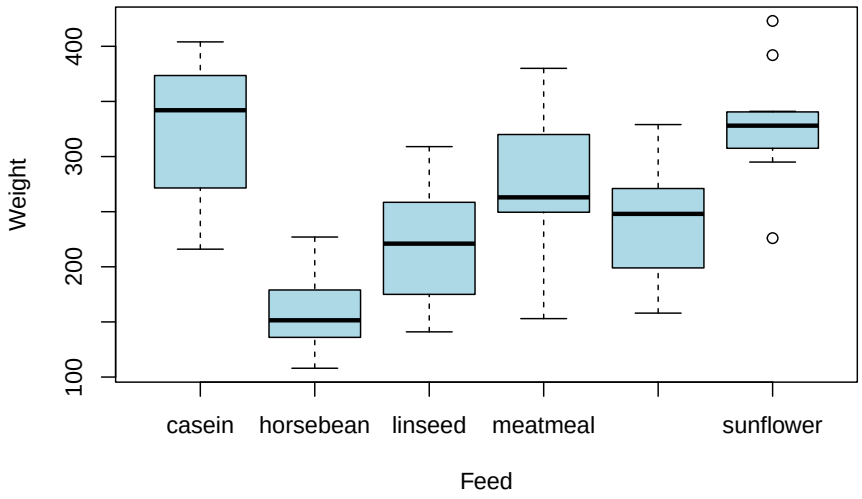
Old Faithful Geyser

Label the median, Q1, and Q3 on the boxplot.

Label the whiskers and, if there are any, potential outliers.

Use the boxplot to find the (approximate) interquartile range for eruption duration.

Use the boxplots below to answer questions a-b. Note that these are called side-by-side boxplots and can be used to compare different categories. In this case, we are comparing weights of chickens who have been given various types of feed.



What is the approximate median weight for chickens given meatmeal?

What is the approximate IQR for chickens fed horsebean?

Based on IQR, which feed has the greatest variability?

Are there any potential outliers? How do you know?

2.3 Descriptive Measures for Populations

So far, we've thought about calculating various descriptive statistics from a sample, but our long-term goal is to estimate descriptive information about a population. At the population level, these values are called **parameters**.

When we find a measure of center, spread, or position, we use a sample to calculate a single value. These single values are called **point estimates** and they are used to *estimate* the corresponding population parameter. For example, we use $\bar{x}$ to estimate the population mean, denoted μ (Greek letter "mu") and s to estimate the population standard deviation, denoted σ (Greek letter "sigma").

Point Estimate	Parameter
sample mean: $\bar{x}$	population mean: μ
sample standard deviation: s	population standard deviation: σ

...and so on and so forth. For each quantity we calculate from a sample (point estimate), there is some corresponding unknown population level value (parameter) that we wish to estimate.

We will discuss this in more detail when we discuss Random Variables and Statistical Inference.

R Lab: Descriptive Statistics and Boxplots

Finding Measures of Center

To find a mean in R, we use the command `mean`. Let's use the `Loblolly` pine data again and find the `mean` tree height.

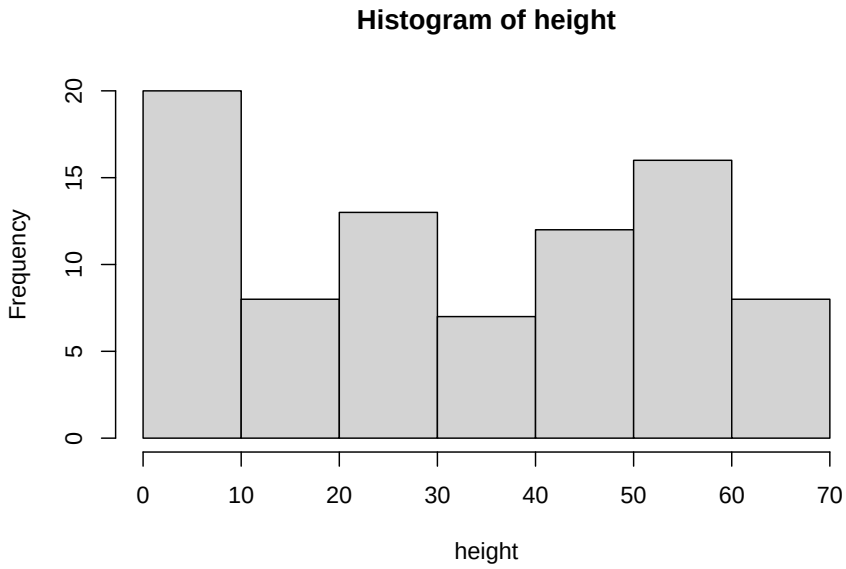
```
attach(Loblolly)
mean(height)
```

```
## [1] 32.3644
```

From R, we can see that the sample mean height of the Loblolly pines is 32.36 feet. (Note that R will often print things out with ## and a number in square brackets. This is just to help us keep track of things if we write a lot of lines of code. You can ignore that number in the brackets!)

Is a mean the appropriate measure of center? We can quickly check the skew with a histogram:

```
hist(height)
```



It might be a little bit hard to tell with this histogram, but it does not look particularly skewed. Let's use our other trick: if the mean and median are approximatly equal, we can say the distribution is approximately symmetric (and therefore the mean is an appropriate measure of center). To calculate the median, we use the command `median`:

```
median(height)
```

```
## [1] 34
```

The sample median of the Loblolly pine heights is 34 feet. Since the mean and median are approximately equal, it would be reasonable to use the mean in this case.

To find a mode, we will use the `table` command. (Recall that this will generate a frequency distribution, from which you can take the mode.)

```
gender = c("Male", "Female", "Other", "Female")
table(gender)
```

```
## gender
## Female   Male  Other
##      2      1      1
```

Your Turn

1. Find the following summary statistics for the `age` variable in the `Loblolly` pine tree data. Include your R code and the output in your answer.

   a. mean
   b. median
2. For the `age` variable in the `Loblolly` pine tree data, which measure of center is most appropriate? Explain.

Finding Measures of Variability

To find a range in R, we get started using the command `range`. Let's keep using the `Loblolly` pine data and find the `mean` tree height.

```
range(height)
```

```
## [1]  3.46 64.10
```

This command gives us the minimum and maximum values. Then to calculate the range, we would find $64.10 - 3.46$. We can also do this in R, because R doubles as a calculator!

```
64.1 - 3.46
```

```
## [1] 60.64
```

The sample range of the Loblolly pine heights is 60.64 feet.

To find the variance and standard deviation, we use the commands `var` and `sd`, respectively:

```
var(height)
```

```
## [1] 427.3979
```

The sample variance of the Loblolly pine heights is 427.4.

```
sd(height)
```

```
## [1] 20.6736
```

The sample standard deviation of the Loblolly pine heights is 20.67 feet.

Your Turn

3. Find the standard deviation of the `age` variable in the `Loblolly` pine tree data. Include your R code and the output in your answer.

Measures of Position

We can get the quartiles quickly using the `summary` command. This will also give us the minimum, mean, and maximum. That's fine!

```
summary(height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.46   10.47   34.00   32.36   51.36   64.10
```

So $Q1 = 10.46$, $Q2 =$ Median $= 34$, and $Q3 = 51.35$. We can also quickly get the interquartile range using the `IQR` command.

```
IQR(height)
```

```
## [1] 40.895
```

Your Turn

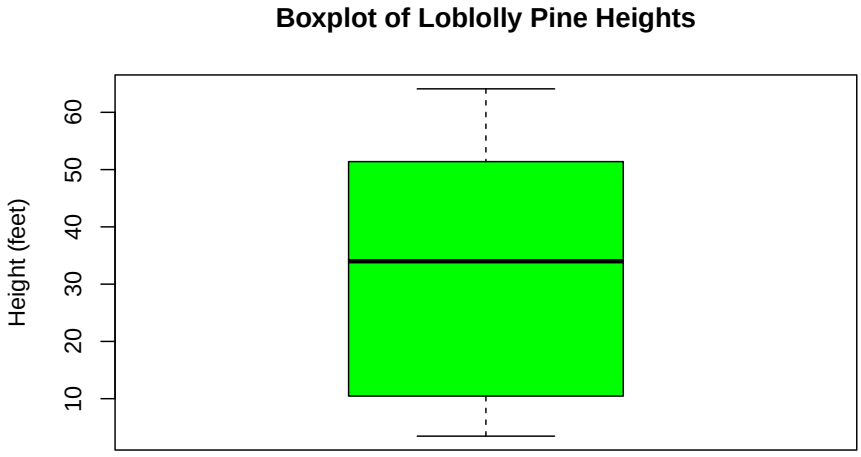
4. Find the interquartile range for the `age` variable in the `Loblolly` pine tree data. Include your R code and the output in your answer.
5. For the `age` variable in the `Loblolly` pine tree data, which measure of spread (variability) is most appropriate? Explain.

Box Plots

To create a boxplot in R, we use the command `boxplot`. We can use some of the same arguments we used with the `hist` command to give it titles and color:

- `main` is where I can give the plot a new title. (Make sure to put the title in quotes!)
- `ylab` is the y-axis (vertical axis) title.
- `col` allows us to give R a specific color for the bars. Notice that I also have to put the color in quotes.

```
boxplot(height,
        main = "Boxplot of Loblolly Pine Heights",
        ylab = "Height (feet)",
        col = "green")
```



R will automatically create the entire box plot, including showing any outliers. (Which we should note that there aren't any in the `height` variable!)

```
detach(Loblolly)
```

Your Turn

6. Create a boxplot of the `age` variable in the `Loblolly` pine tree data. Give your plot an appropriate vertical axis title.

Course Survey Data

To access the course survey data, enter the following command, replacing the `TYY` with the appropriate semester, where `T` is the term (`f` for fall; `s` for spring) and `YY` is the year.

```
source("https://lgpperry.github.io/teaching/stat1/data/data_TYY.R")
```

For example, for spring semester 2024, this command would look like:

```
source("https://lgpperry.github.io/teaching/stat1/data/data_s24.R")
```

```
## [1] "Use the command `str(survey)` to view the variables in the data."
## [1] "The command `attach(survey)` will allow you to access the variables directly."
```

This command references some R code I have on my website that both reads in the data and cleans it up a little bit so that it's ready for you to use. The data saved in R will be called `survey`.

Note: if you happen to try this out before I get this semester's data uploaded, you'll get an error along the lines of "cannot open the connection". In the meantime, you may use the example code above to get an idea for what the data might look like in R.

To figure out exactly what we're working with, we will use the `str` command:

```
str(survey)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ haircolor   : Factor w/ 3 levels "Black","Blonde",..: 1 1 1 1 3 2 3 1 1 3 ...
##  $ numpets     : int  0 0 0 12 2 1 1 0 0 2 ...
##  $ wherefrom   : Factor w/ 4 levels "Bay Area","Other California",..: 4 1 4 1 1 4 4 1 4 1 ...
##  $ height      : num  5.75 5.9 5.58 5 5 ...
##  $ sleep       : int  6 7 6 6 8 7 8 7 6 9 ...
##  $ pets        : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 2 2 1 1 2 ...
##  $ schooldays  : int  5 5 4 5 5 3 5 3 4 3 ...
##  $ schoolperiod: Factor w/ 4 levels "College","Elementary School",..: 1 3 3 4 3 3 2 1 3 2 ...
##  $ siblings    : int  5 1 7 3 2 1 4 3 2 2 ...
##  $ siborder    : Factor w/ 4 levels "Middle child",..: 2 4 1 2 1 2 4 1 1 4 ...
##  $ job         : Factor w/ 2 levels "No","Yes": 2 2 2 1 1 2 2 1 1 2 ...
##  $ exercise    : int  0 5 0 3 0 4 1 0 0 3 ...
##  $ units       : int  8 15 12 12 16 15 12 15 16 12 ...
##  $ major       : Factor w/ 8 levels "Biology","Business",..: 1 2 2 6 6 2 2 5 6 2 ...
```

The stuff shown in the column on the left are all of the variable names. I try to name these to match the survey questions in an intuitive way, but if you're not sure which variable matches which survey question, there will be additional information available on Canvas.

The stuff shown on the right gives some information about the variable *type*. If it says `int` or `num`, those are numeric variables. Variables that say `Factor`, are categorical and the "levels" are the different categories in the data.

Finally, to use these variables directly, use the command `attach(survey)`.

Your Turn

7. Select a numeric variable from this semester's course survey data. Find the following summary statistics for that variable:
   a. mean
   b. median
   c. standard deviation
   d. interquartile range
8. Create a histogram of the numeric variable you chose in (7). Make sure to give it an appropriate title and axis labels.
9. Select a categorical variable from this semester's course survey data. Find the following:
   a. frequency distribution

b. mode

10. Create a bar plot of the categorical variable you chose in (9). Make sure to give it an appropriate title and axis labels.

Chapter 3

Regression and Correlation

In this module, we will extend our conversation on descriptive measures for quantitative variables to include the relationship between two variables.

Module Learning Outcomes/Objectives

1. Interpret a correlation coefficient.
2. Calculate and interpret a regression line.
3. Interpret a coefficient of determination.
4. Understand the relationship between a correlation coefficient and a coefficient of determination.
5. Use a regression line to make predictions.

3.1 Linear Equations

From your previous math classes, you should have a passing familiarity with linear equations like $y = mx + b$. In statistics, we write these as

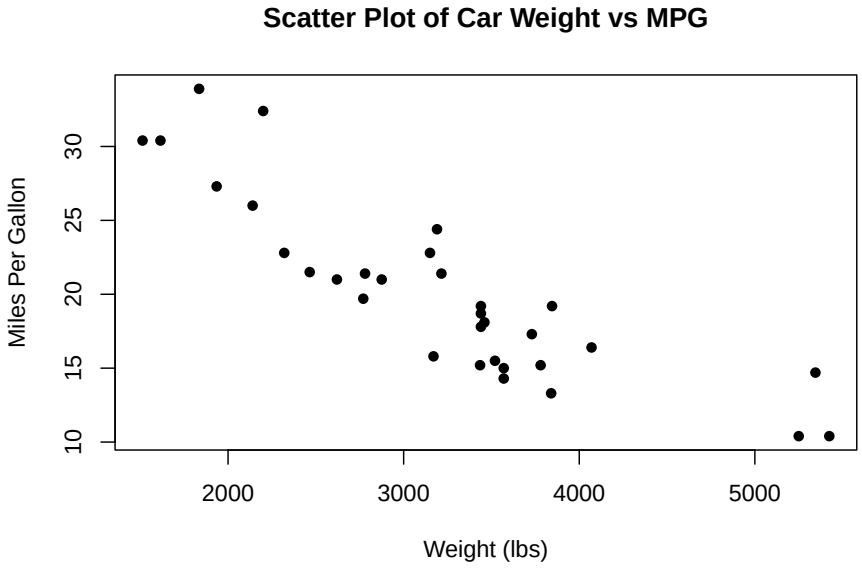
$$y = b_0 + b_1 x$$

where b_0 and b_1 are constants, x is the independent variable, and y is the dependent variable. The graph of a linear function is always a (straight) line.

The **y-intercept** is b_0, the value the dependent variable takes when the independent variable $x = 0$. The **slope** is b_1, the change in y for a one-unit increase in x.

3.1.1 Scatter Plots

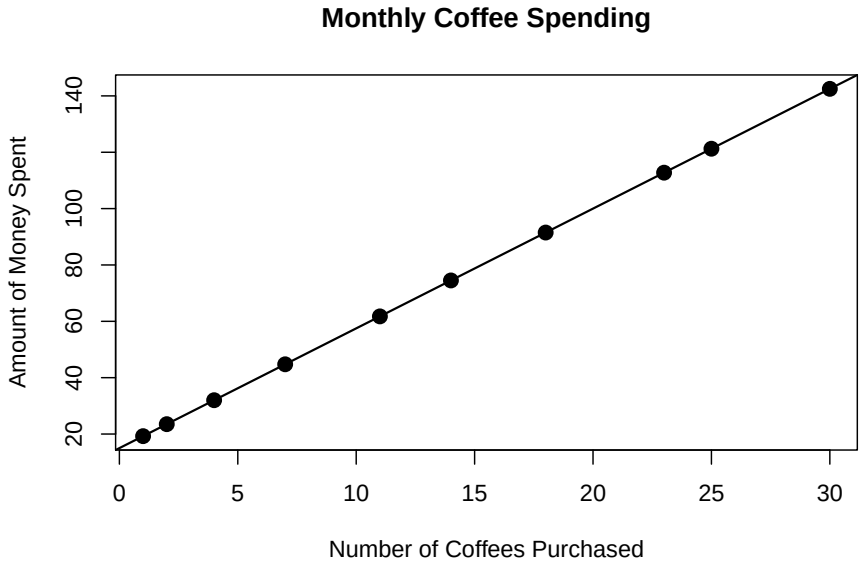
A **scatter plot** shows the relationship between two (numeric) variables.



This is just a bunch of points of the form (x, y) plotted in the xy-plane, where instead of showing the full x- and y-axes, we crop the plot to show only the part where we have data.

At a glance, we can see that (in general) heavier cars have lower MPG. We call this type of data **bivariate data** ("bi" as in "two" and "variate" as in "variable").

Now consider a person who spends \$15 each month buying a bag of coffee for their home coffee machine. Each coffee they purchase costs \$4.25. If we know the numbers of coffees they bought last month, we know *exactly* how much money they spent!

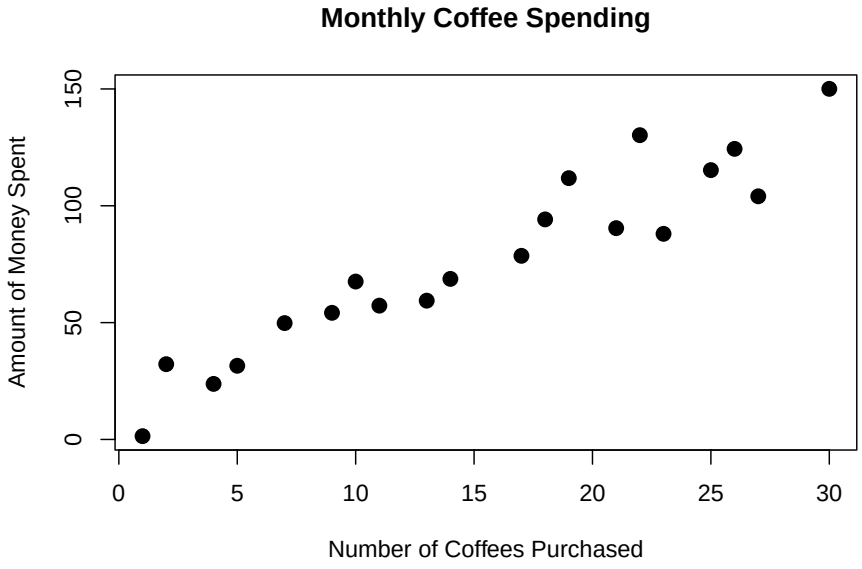


This relationship can be modeled perfectly with a straight line: $y = 15 + 4.25x$. The y-intercept tells us that, in a month when this person buys zero coffees ($x = 0$), they spend \$15. The slope tells us that, for each additional coffee purchased, the amount of money spent in that month goes up by \$4.25

When we can do this – model a relationship perfectly – we know the exact value of y whenever we know the value of x. This is nice (we would love to be able to do this all the time!) but typically real-world data is much more complex than this.

3.1.2 Linear Regression

Linear regression takes the idea of fitting a line to points and allows the relationship to be imperfect. What if that bag of coffee didn't always cost \$15? Or the coffee drinks didn't always cost \$4.25? In this case, you might get a plot that looks something like this:



Without doing any math, you can think about where you might draw a line through these points. Later on, we will formalize how to do that mathematically.

Bear with me for a bit while we go over some mathematical notation. The notation for the linear regression line is

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β is the Greek letter "beta".
- β_0 and β_1, the slope and intercept, are constants.
- ϵ is the Greek letter "epsilon".
- ϵ represents the *error*, the fact that the points don't all line up perfectly.

If you think back to Section 2.3 (Descriptive Measures for Populations), this is the true, unknown (population) version of the line. We estimate β_0 and β_1 using data and denote the estimated line by

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$, "y-hat", is the estimated value of y.
- b_0 is the estimate for β_0.
- b_1 is the estimate for β_1.

That is, b_0 is the point estimate for the population parameter β_0 and b_1 is the point estimate for the population parameter β_1.

We drop the error term ϵ when we estimate the constants for a regression line; we assume that the mean error is 0, so *on average* we can ignore this error.

We use a regression line to make predictions about y using values of x.

- x is the **predictor variable**
  - We use x to *predict* y.
- y is the **response variable**
  - The value of y *responds* to whatever we plug in for x.

Example: Researchers took a variety of measurements on 344 adult penguins (species chinstrap, gentoo, and adelie) near Palmer Station in Antarctica.

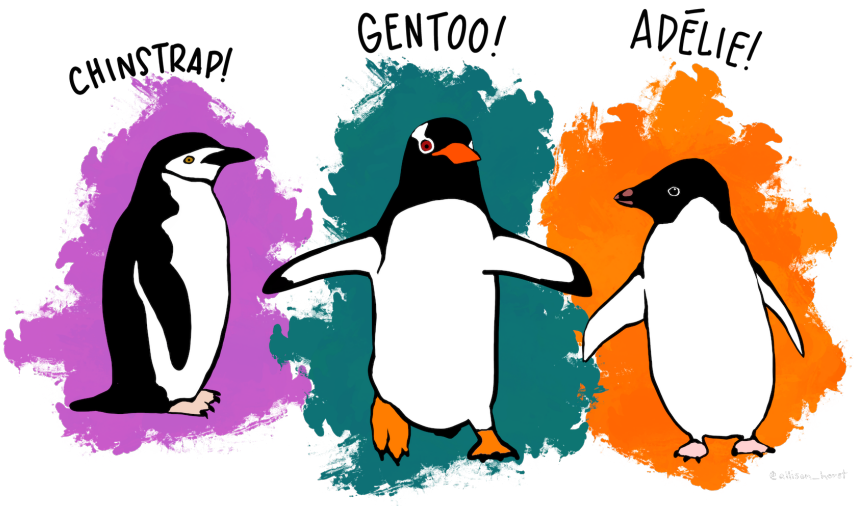
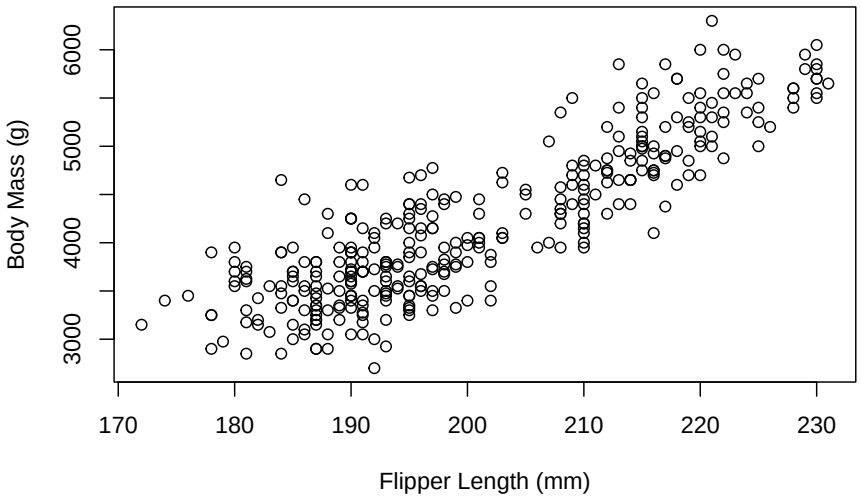


Figure 3.1: Artwork by @allison_horst

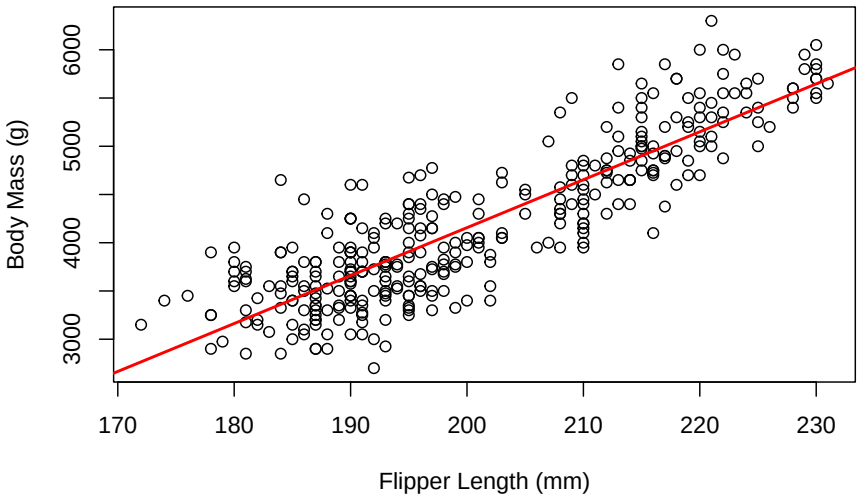
We will consider two measurements for each penguin: their body mass (weight in grams) and flipper length (in millimeters).



Clearly, the relationship isn't perfectly linear, but there does appear to be some kind of linear relationship (as flipper length increases, body mass also increases). We want to try to use flipper length (x) to predict body mass (y).

The regression model for these data is

$$\widehat{y} = -5780.83 + 49.69x$$



We can interpret the slope and intercept by using the definitions given previously and tweaking them to match the context of the data:

- Slope: For a one-mm increase in flipper length, body mass is predicted to increase by 49.69g.
- Intercept: When flipper length is equal to 0mm, body mass is predicted to be -5780.83g.

What happened with our intercept? Sometimes interpretations of intercepts do not make sense and that's ok! It doesn't make sense to have a flipper length of 0, so we shouldn't be concerned that plugging in 0 gives us a nonsense value for body length. That intercept value is still important in the regression line because it helps situate the regression line in the xy plane.

To predict the body mass for a penguin with a flipper length of 180cm, we just need to plug in 180 for flipper length (x):

$$\hat{y} = -5780.83 + 49.69 \times 180 = 3163.37\text{g}.$$

Note: because the regression line is built using the data's original units (mm for flipper length, g for body mass), the regression line will preserve those units. That means that when we plugged in a value in mm, the equation spit out a predicted value in g.

Section Exercises

For exercises 1-5, determine which variable should be the predictor (x) and which should be the response (y).

We want to use annual days of sunshine to predict depression rates in US cities.

We want to know if we can use annual paid vacation days to predict employee satisfation.

A researcher will use mouse hormone levels to predict changes in weight.

A company wants to know if they can determine revenue based on amount of money spent on ads.

I want to know if the amount my dog barks is predictive of how quickly solicitors leave the front door.

Suppose we have a regression line to predict miles per gallon (y) using car weight in pounds (x):

$$\hat{y} = 37.285 - 0.005x$$

Interpret the intercept in the context of the problem.

Interpret the slope in the context of the problem.

Predict the miles per gallon for a car that weighs 3,500 lbs.

Suppose we have a regression line to predict penguin bill length (in millimeters) using bill depth (in millimeters). The regression line is $y = 55.07 - 0.650x$.

Which variable is x and which is y?.

Interpret the intercept in the context of the problem.

Interpret the slope in the context of the problem.

Predict the bill length for a penguin with a bill depth of 17.3 mm.

3.2 Correlation

We've talked about the strength of linear relationships, but it would be nice to formalize this concept. The **correlation** between two variables describes the strength of their linear relationship. It always takes values between -1 and 1. We denote the correlation (or correlation coefficient) by R:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}\right)$$

where s_x and s_y are the respective standard deviations for x and y. The sample size n is the total number of (x, y) pairs.

Example: Suppose $x : 2, 4, 3, 5, 2$ and $y : 1, 8, 6, 10, 4$. Find the correlation between x and y.

Solution: First, $n = 5$ and we should calculate the means and standard deviations for x and y. Using a computer, the means are $\bar{x} = 3.2$ and $\bar{y} = 5.8$ and the standard deviations are $s_x = 1.30$ and $s_y = 3.49$.

Like we did for variance, we'll use a table to help organize our math.

x	$\frac{x-\bar{x}}{s_x}$	y	$\frac{y-\bar{y}}{s_y}$	$\frac{x-\bar{x}}{s_x} \times \frac{y-\bar{y}}{s_y}$
2	$\frac{2-3.2}{1.30} = -0.92$	1	$\frac{1-5.8}{3.49} = -1.37$	$(-0.92)(-1.37) = 1.26$
4	$\frac{4-3.2}{1.30} = 0.61$	8	$\frac{8-5.8}{3.49} = 0.63$	$(0.61)(0.63) = 0.39$
3	$\frac{3-3.2}{1.30} = -0.15$	6	$\frac{6-5.8}{3.49} = 0.06$	$(-0.15)(0.06) = -0.01$
5	$\frac{5-3.2}{1.30} = 1.38$	10	$\frac{10-5.8}{3.49} = 1.20$	$(1.38)(1.20) = 1.66$
2	$\frac{2-3.2}{1.30} = -0.92$	4	$\frac{4-5.8}{3.49} = -0.52$	$(-0.92)(-0.52) = 0.47$
				$\sum_{i=1}^{n} \left(\frac{x_i-\bar{x}}{s_x} \times \frac{y_i-\bar{y}}{s_y} \right) = 3.78$

Then $R = \frac{3.78}{5-1} = 0.94$.

This is a pretty involved formula! In general, we'll let a computer handle this one.

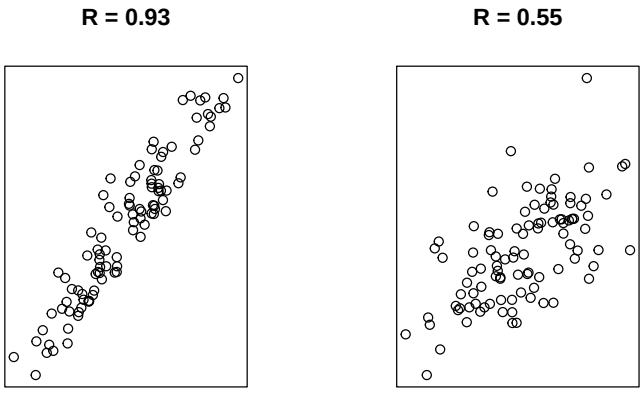
Correlations

- close to -1 suggest strong, negative linear relationships.
- close to $+1$ suggest strong, positive linear relationships.
- close to 0 have little-to-no linear relationship.

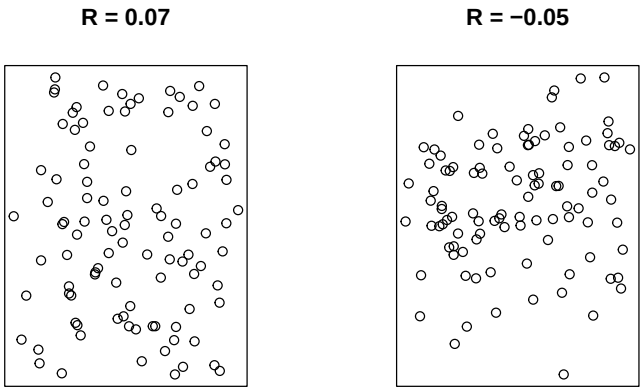
Note: the sign of the correlation will match the sign of the slope!

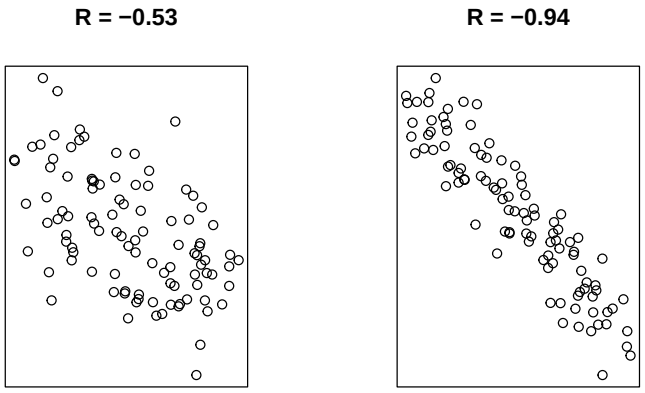
- If $R < 0$, there is a downward trend and $b_1 < 0$.
- If $R > 0$, there is an upward trend and $b_1 > 0$.
- If $R \approx 0$, there is no relationship and $b_1 \approx 0$.

When two variables are highly correlated (R close to -1 or 1), we know there is a strong linear *relationship* between them, but we do not know what *causes* that relationship. For example, when we looked at the Palmer penguins, we noticed that an increase in flipper length related to an increase in body weight… but we have no way of knowing if flipper length *causes* a higher body weight (or vice versa). That is, *correlation does not imply causation.*

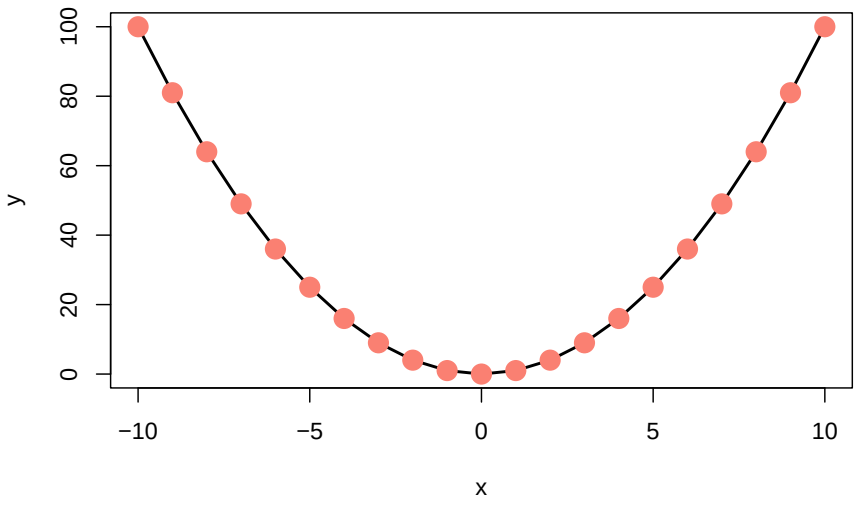


Example Correlations:





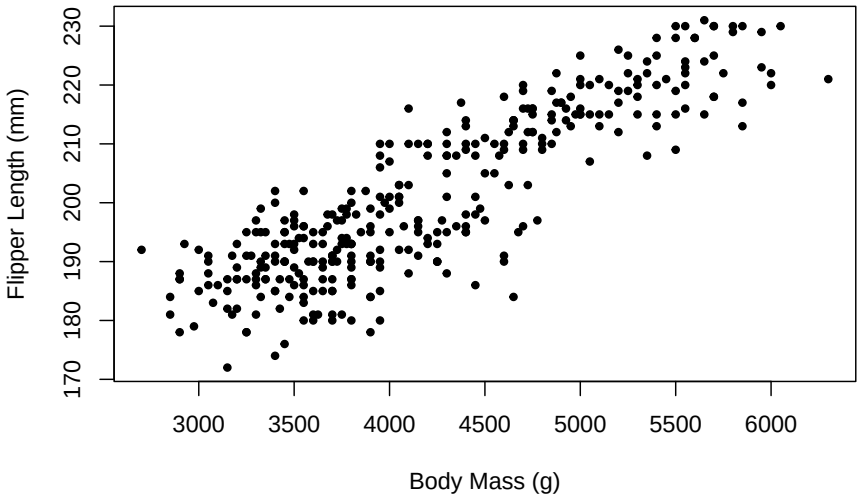
A final note: correlations only represent *linear* relationships. Consider the following scatter plot:



Obviously there's a strong relationship between x and y. In fact, there's a perfect relationship here: $y = x^2$. But the *correlation* between x and y is 0! This is one reason why it's important to examine the data both through visual and numeric measures.

Section Exercises

1. The `penguins` data set (from the `palmerpenguins` package in R) has data on various penguin body measurements. We want to consider the relationship between a penguin's `body mass` from its `flipper length`. The scatter plot below shows the relationship between these two variables.



Based only on the scatter plot, what can you say about the correlation R? Explain.

The actual correlation between these two variables is $R = 0.871$. What does this tell you about the relationship between penguin body mass and flipper length?

3.3 Finding a Regression Line

Residuals are the leftover *stuff* (variation) in the data after accounting for model fit:

$$\text{data} = \text{prediction} + \text{residual}$$

Each observation has its own residual. The residual for an observation (x, y) is the difference between observed (y) and predicted $(\hat{y})$:

$$e = y - \hat{y}$$

We denote the residuals by e and find $\hat{y}$ by plugging x into the regression equation. If an observation lands above the regression line, $e > 0$. If below, $e < 0$.

When we estimate the parameters for the regression, our goal is to get each residual as close to 0 as possible. We might think to try minimizing

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

but that would just give us very large negative residuals. (Remember that this symbol $\sum$ just means I'm adding up all the values of whatever is next to it, so $\sum_{i=1}^{n} e_i$ means we add all of the the residuals.)

As with the variance, we will use squares to shift the focus to magnitude:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.1}$$

$$= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \tag{3.2}$$

This will allow us to shrink the residuals toward 0: the values b_0 and b_1 that minimize this will make up our regression line. This is a calculus-free course, though, so we'll skip the proof of the minimization part.

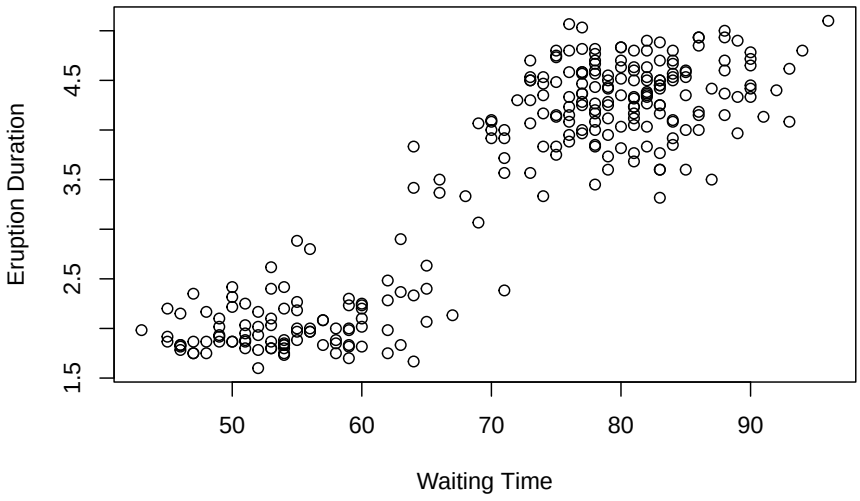
The slope can be estimated as

$$b_1 = \frac{s_y}{s_x} \times R$$

and the intercept as

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

Example: Consider two measurements taken on the Old Faithful Geyser in Yellowstone National Park: `eruptions`, the length of each eruption and `waiting`, the time between eruptions. Each is measured in minutes.



There does appear to be some kind of linear relationship here, so we will see if we can use the wait time to predict the eruption duration. The sample statistics for these data are

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

Since we want to use wait time to predict eruption duration, wait time is x and eruption duration is y. Then

$$b_1 = \frac{1.14}{13.60} \times 0.90 \approx 0.076$$

and

$$b_0 = 3.49 - 0.076 \times 70.90 \approx -1.87$$

so the estimated regression line is

$$\hat{y} = -1.87 + 0.076x$$

To interpret b_1, the slope, we would say that for a one-minute increase in waiting time, we would predict a 0.076 minute increase in eruption duration. The intercept is a little bit trickier. Plugging in 0

for x, we get a predicted eruption duration of -1.87 minutes. There are two issues with this. First, a negative eruption duration doesn't make sense… but it also doesn't make sense to have a waiting time of 0 minutes.

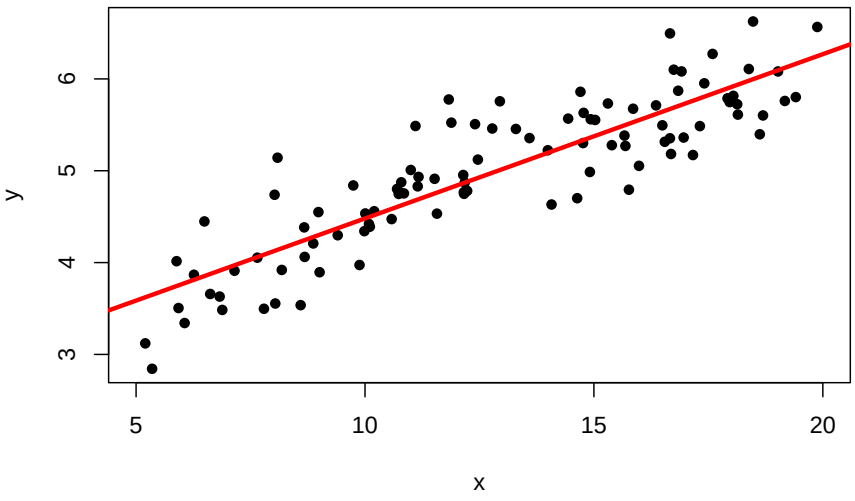
3.3.1 The Coefficient of Determination

With the correlation and regression line in hand, we will add one last piece for considering the fit of a regression line. The **coefficient of determination**, R^2, is the square of the correlation coefficient. This value tells us how much of the variability around the regression line is accounted for by the regression. An easy way to interpret this value is to assign it a letter grade. For example, if $R^2 = 0.84$, the predictive capabilities of the regression line get a B.

3.3.2 When Prediction Goes Wrong

It's important to stop and think about our predictions. Sometimes, the numbers don't make sense and it's easy to see that there's something wrong with the prediction. Other times, these issues are more insidious. Usually, all of these issues result from what we call *extrapolation*, applying a model estimate for values outside of the data's range for x. Our linear model is only an approximation, and we don't know anything about how the relationship outside of the scope of our data!

Consider the following data with the best fit line drawn on the scatter plot.



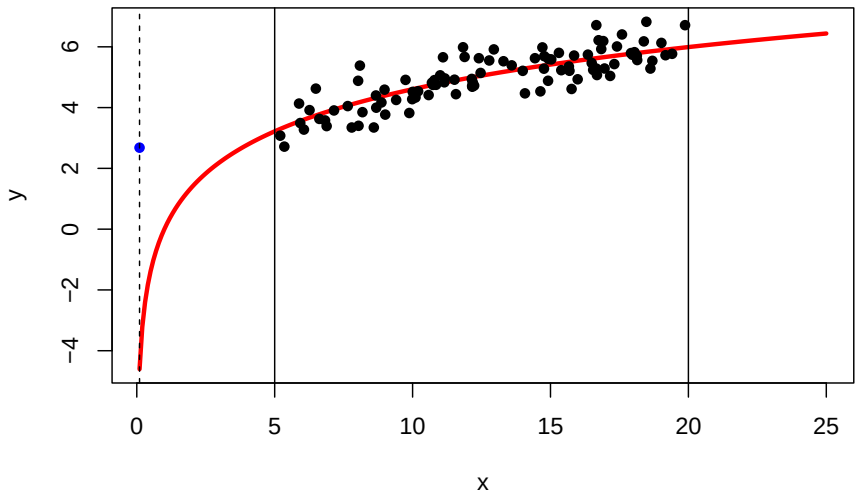
The best fit line is

$$\hat{y} = 2.69 + 0.179x$$

and the correlation is $R = 0.877$. Then the coefficient of determination is $R^2 = 0.767$ (think: a C grade), so the model has decent predictive capabilities. More precisely, the model accounts for 76.7% of the variability about the regression line. Now suppose we wanted to predict the value of y when $x = 0.1$:

$$\hat{y} = 2.66 + 0.181 \times 0.1 = 2.67$$

This seems like a perfectly reasonable number… But what if I told you that I generated the data using the model $y = 2\ln(x) + \text{random error}$? (If you're not familiar with the natural log, ln, don't worry about it! You won't need to use it.) The true (population) best-fit model would look like this:



The vertical lines at $x = 5$ and $x = 20$ show the bounds of our data. The blue dot at $x = 0.1$ is the predicted value $\hat{y}$ based on the linear model. The dashed horizontal line helps demonstrate just how far this estimate is from the true population value! This does *not* mean there's anything inherently wrong with our model. If it works well from $x = 5$ to $x = 20$, great, it's doing its job!

Section Exercises

1. The Loblolly pine tree dataset contains information about `height` (in feet) and `age` (in years) of Loblolly pine trees. We want to use `height` to predict `age`. The sample statistics for these data are

	Height	Age
mean	32.36	13.00
sd	20.67	7.90
		$R = 0.90$

Which variable should be the predictor (x) and which should be the outcome (y)? How do you know?

Calculate the slope.

Calculate the intercept.

Write out your regression line.

Interpret your slope and intercept values in the context of the problem.

2. The `penguins` dataset (from the `PalmerPenguins` package in R) has data on various penguin body measurements. We want to build a model for predicting a penguin's `body mass` from its `flipper length`. The table below contains some summary information about these data. The correlation is $R = 0.871$.

	Body Mass (g)	Flipper Length (mm)
Minimum	2700.0	172.0
Maximum	6300.0	231.0
Mean	4201.75	200.92
Standard Deviation	801.95	14.06

Based on the prompt, which variable should be your outcome (y) and which your predictor (x)?

Calculate the slope. Interpret your value in the context of the problem.

Calculate the intercept. Interpret your value in the context of the problem. Does this interpretation make sense?

Write out the linear regression line using your slope and intercept from parts (b) and (c).

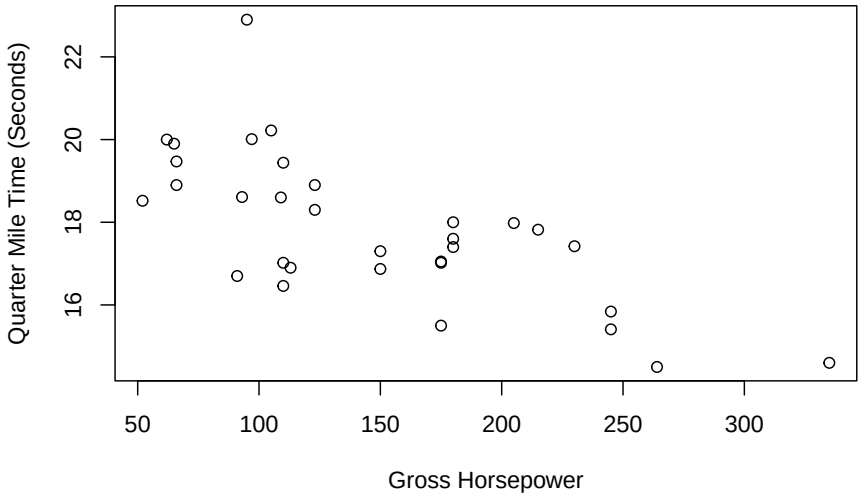
Find and interpret the coefficient of determination, R^2, for this model.

Predict the body mass for a penguin with a flipper length of 200mm. Do you have any concerns with this prediction?

Predict the body mass for a penguin with a flipper length of 150mm. Do you have any concerns with this prediction?

3. The `mtcars` dataset has measurements (from 1974) on various aspects of automobile design for 32 automobiles. Suppose we want to use `horsepower` to predict `quarter mile time`. The sample statistics for these data are

	Horsepower	Quarter Mile Time (sec)
mean	146.69	17.85
sd	68.56	1.79
		$R = -0.708$



Based on the prompt, which variable should be your outcome and which your predictor?

Calculate the slope. Interpret your value in the context of the problem.

Calculate the intercept. Interpret your value in the context of the problem. Does this interpretation make sense?

Write out the linear regression line using your slope and intercept from parts (b) and (c).

What does the correlation tell us about the relationship between horsepower and quarter mile time?

Find and interpret the coefficient of determination, R^2, for this model.

Predict the quarter mile time for a car with horsepower equal to 175. Do you have any concerns with this prediction?

Predict the quarter mile time for a car with horsepower equal to 500. Do you have any concerns with this prediction?

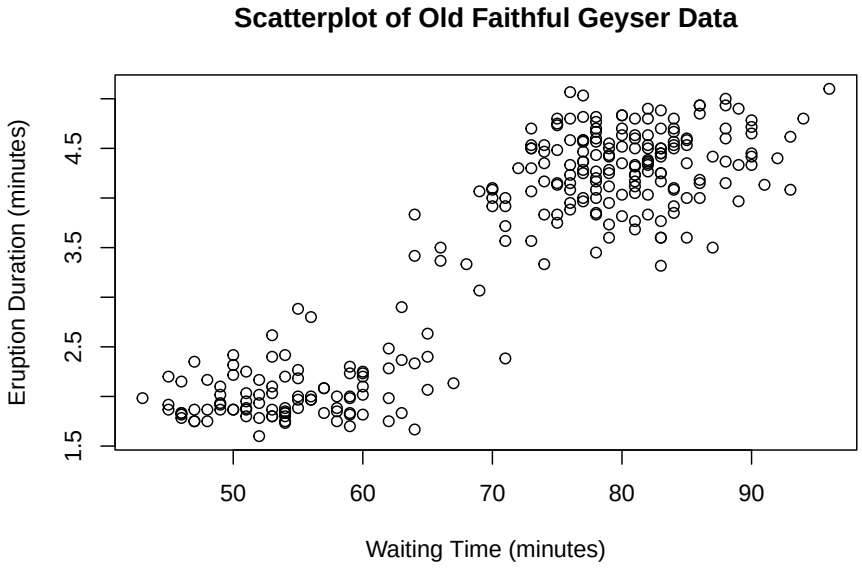
R Lab: Scatterplots and Regression

Scatterplots

Let's create a scatterplot from the `faithful` dataset in R, which contains measurements on waiting time and eruption duration for Old Faithful geyser in Yellowstone National Park. We can use some of the same arguments we used with the `hist` and `boxplot` commands to give it titles:

- `x` is the variable I want represented on the x-axis (horizontal axis).
- `y` is the variable I want represented on the y-axis (vertical axis).
- `main` is where I can give the plot a new title. (Make sure to put the title in quotes!)
- `xlab` is the x-axis title.
- `ylab` is the y-axis title.

```
data("faithful")
attach(faithful)
plot(x = waiting, y = eruptions,
     main = "Scatterplot of Old Faithful Geyser Data",
     xlab = "Waiting Time (minutes)",
     ylab = "Eruption Duration (minutes)")
```



Correlation

To find the correlation between two variables `x` and `y`, we use the command `cor(x,y)`.

```
cor(x = waiting, y = eruptions)
```

```
## [1] 0.9008112
```

The correlation between waiting time and eruption duration for Old Faithful Geyser is -0.645.

Finding a Regression Line

To find a regression line using R, we use the command `lm`, which stands for "linear model". The necessary argument is the `formula`, which takes the form `y ~ x`. For example, to find the regression line for the Old Faithful geyser data with waiting time predicting eruption duration,

$$\text{eruptions} = b_0 + b_1 \times \text{waiting}$$

we would use `formula = eruptions ~ waiting`:

```
lm(formula = eruptions ~ waiting)
```

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

```
## lm(formula = eruptions ~ waiting)
##
## Coefficients:
## (Intercept)      waiting
##    -1.87402      0.07563
```

The `lm` command prints out the intercept, b_0 and the slope (waiting), b_1. So the model is

$$\text{eruptions} = -1.87 + 0.08 \times \text{waiting}$$

To get the coefficient of determination, we can simply find and square the correlation coefficient. I will do this by putting the `cor()` command in parentheses and squaring it by adding `^2` at the end:

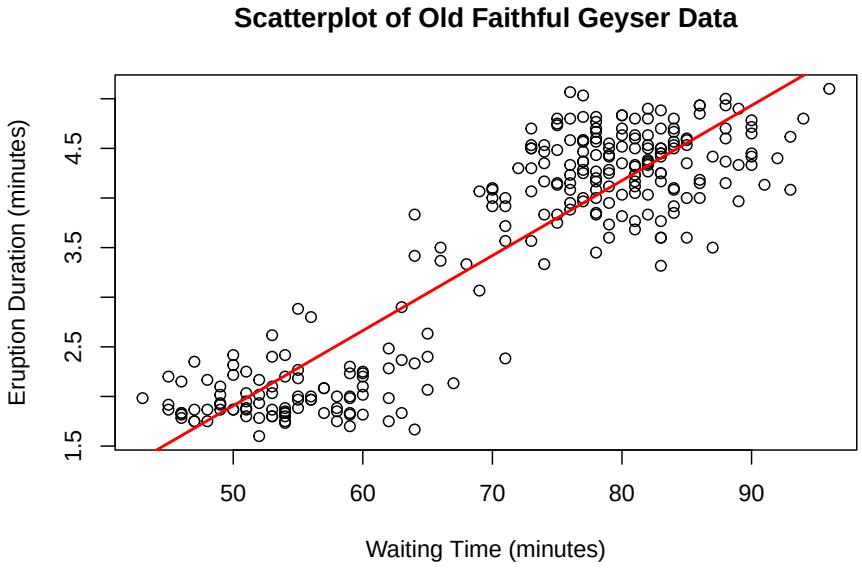
```
(cor(x = waiting, y = eruptions))^2
```

```
## [1] 0.8114608
```

We might also be interested in adding a regression line to a scatterplot. Now that we know how to find both separately, we can put them together. I can add a line to a plot in R by using the command `abline`. This command takes two primary arguments, along with some optional arguments for adjusting line color and thickness:

- `a`: the intercept (b_0)
- `b`: the slope (b_1)
- `col`: the line color
- `lwd`: the line width

```
plot(x = waiting, y = eruptions,
     main = "Scatterplot of Old Faithful Geyser Data",
     xlab = "Waiting Time (minutes)",
     ylab = "Eruption Duration (minutes)")
abline(summary(lm(formula = eruptions ~ waiting))$coef[,1],
       col = "red", lwd = 2)
```


Scatterplot of Old Faithful Geyser Data
Eruption Duration (minutes)
1.5
2.5
3.5
4.5
50
60
70
80
90
Waiting Time (minutes)

Chapter 4

Probability Concepts

In previous modules, we discussed ways to describe variables and the relationships between them. From here, we want to start asking inferential statistics questions like "If my sample mean is 10, how likely is it that the population mean is actually 11?". Probability is going to start us on this path.

Probability theory is the science of uncertainty and it is really interesting! But it can also be pretty challenging. I try to frame probability around things most of us can do at home: flipping a coin, rolling a die, drawing from a deck of cards. You certainly don't need any of these things to get through this module, but you may find it helpful to have a coin/die/deck of cards on hand as you read through the examples.

Take your time running practice problems and going through the examples, using a tactile approach like sorting through your deck of cards whenever it seems helpful.

Module Learning Objectives/Outcomes

1. Find and interpret probabilities for equally likely events.
2. Find and interpret probabilities for events that are not equally likely.
3. Find and interpret joint and marginal probabilities.
4. Find and interpret conditional probabilities.
5. Use the multiplication rule and independence to calculate probabilities.

R Objectives: *none*

This module's outcomes correspond to course outcome (3) understand the basic rules of probability.

4.1 Experiments, Sample Spaces, and Events

Probability is the science of uncertainty. When we run an experiment, we are unsure of what the outcome will be. Because of this uncertainty, we say an experiment is a **random process**.

The probability of an event is the proportion of times it would occur if the experiment were run infinitely many times. For a collection of *equally likely events*, this looks like:

$$\text{probability of event} = \frac{\text{number of ways event can occur}}{\text{number of possible (unique) outcomes}}$$

An **event** is some specified possible outcome (or collection of outcomes) we are interested in observing.

> **Example**: If you want to roll a 6 on a six-sided die, there are six possible outcomes $\{1, 2, 3, 4, 5, 6\}$. In general, we assume that each die face is equally likely to appear on a single roll of the die, that is, that the die is *fair*. So the probability of rolling a 6 is
>
> $$\frac{\text{number of ways to roll a 6}}{\text{number of possible rolls}} = \frac{1}{6}$$

> **Example**: We can extend this to a collection of events, say the probability of rolling a 5 or a 6:
>
> $$\frac{\text{number of ways to roll a 5 or 6}}{\text{number of possible rolls}} = \frac{2}{6}$$

The collection of all possible outcomes is called a **sample space**, denoted S. For the six-sided die, $S = \{1, 2, 3, 4, 5, 6\}$.

To simplify our writing, we use **probability notation**:

- Events are assigned capital letters.
- $P(A)$ denotes the probability of event A.
- Sometimes we will also shorten simple events to just a number. For example, $P(1)$ might represent "the probability of rolling a 1".

We can estimate probabilities from a sample using a frequency distribution.

> **Example**: Consider the following frequency distribution from Module 1

Class	Frequency
freshman	12
sophomore	10
junior	3

Class	Frequency
senior	5

If a student is selected *at random* (meaning each student is equally likely to be selected), the probability of selecting a sophomore is

$$\text{probability of sophomore} = \frac{\text{number of ways to select a sophomore}}{\text{total number of students}} = \frac{10}{30} \approx 0.3333$$

The probability of selecting a *junior or a senior* is

$$\frac{\text{number of ways to select a junior or senior}}{\text{total number of students}} = \frac{3+5}{30} = \frac{8}{30} \approx 0.2667$$

Using probability notation, we might let A be the event we selected a junior and B be the event we selected a senior. Then

$$P(A \text{ or } B) = 0.2667$$

Section Exercises

1. Suppose you're playing a game and need to roll a 17 or higher on a 20-sided die for your next action to be successful.
   a. What is the sample space?
   b. What is the probability of rolling a 17 or higher?
2. Consider again frequency distribution from Module 1

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

For a student selected at random from this course, what is the probability they are a senior?

What is the probability they are *not* a senior?

4.2 Probability Distributions

Two outcomes are **disjoint** or **mutually exclusive** if they cannot both happen (at the same time). Think back to how we developed bins for histograms - the bins need to be nonoverlapping - this is the same idea!

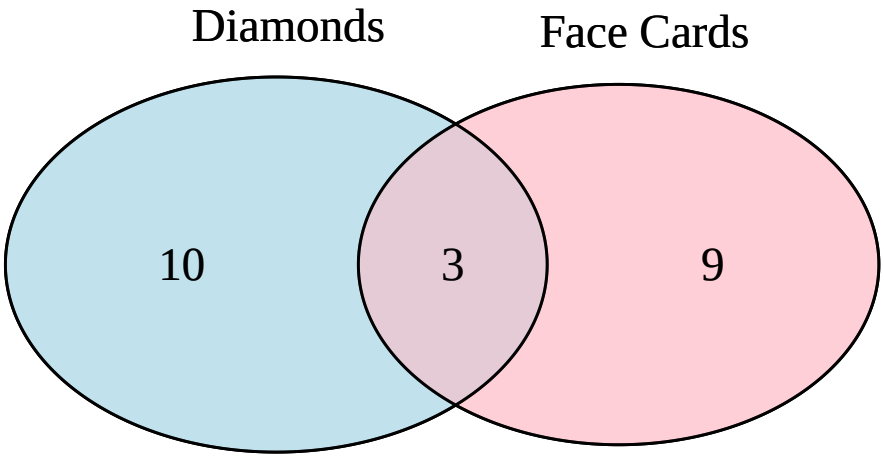
Example: If I roll a six-sided die one time, rolling a 5 and rolling a 6 are disjoint. I can get a 5 *or* a 6, but not both on the same roll.

Example: If I select a student, they can be a freshman *or* a sophomore, but that student cannot be both a freshman and a sophomore at the same time.

The outcome must be one event or the other (it cannot be both at the same time).

4.2.1 Venn Diagrams

Venn Diagrams show events as circles. The circles overlap where events share common outcomes.



When a Venn Diagram has *no overlap* the events are mutually exclusive. This Venn Diagram shows the event "Draw a Diamond" and the event "Draw a Face Card". There are 13 diamonds and 12 face cards in a deck. In this case, the events are *not* mutually exclusive: it's possible to draw both a diamond and a face card at the same time: the Jack of Diamonds, Queen of Diamonds, and King of Diamonds.

The "face cards" are the Jacks, Queens, and Kings. Each row represents a "suit". From top to bottom, the suits are clubs, spades, hearts, and diamonds. Cards can be either red (hearts and diamonds) or black (spades and clubs).

We can also use Venn Diagrams to visualize the relationships between events.

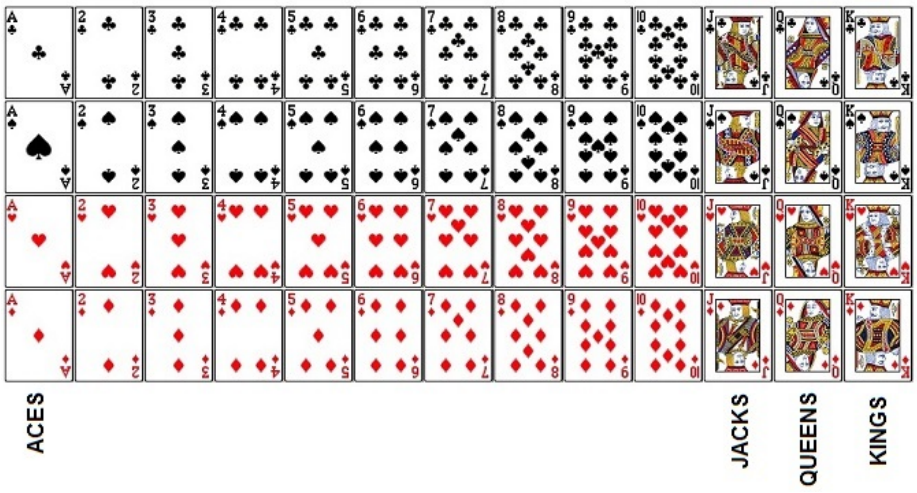
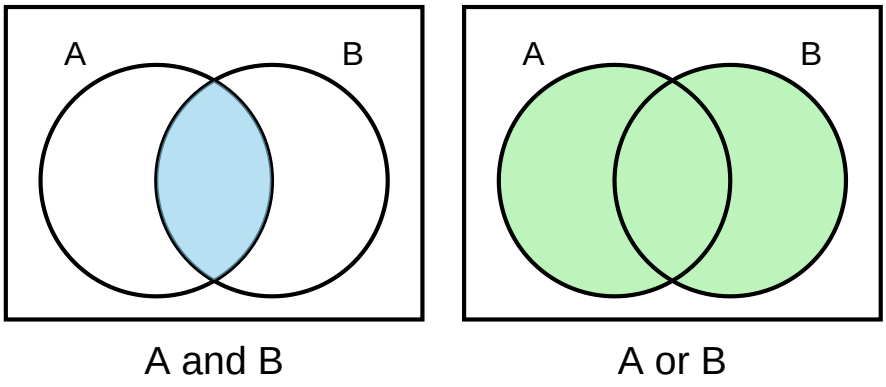
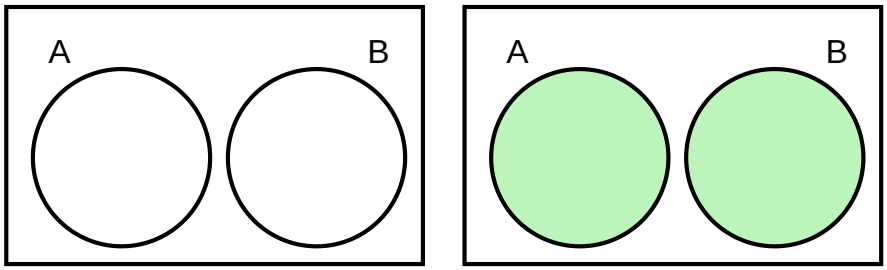


Figure 4.1: Image retrieved from www.cis.upenn.edu/~cis110/



If A and B are disjoint, this looks like



A and B A or B

Notice that in this case there is no shading for A and B because they cannot both happen!

4.2.2 Probability Axioms

A **probability distribution** lists all possible disjoint outcomes (think: all possible values of a variable) and their associated probabilities. This can be in the form of a table

Roll of a six-sided die	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

(note that we could visualize this with a bar plot!) or an equation, which we will discuss in a later module.

The **probability axioms** are requirements for a valid probability distribution. They are:

1. All listed outcomes must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must sum to 1.

Note that #2 is true for ALL probabilities. If you ever calculate a probability and get a negative number or a number greater than 1, you know something went wrong!

Example: Use the probability axioms to check whether the following tables are probability distributions.

A)

X	{1 or 2}	{3 or 4}	{5 or 6}
P(X)	1/3	1/3	1/3

Each axiom is satisfied, so this is a valid probability distribution.

B)

Y	{1 or 2}	{2 or 3}	{3 or 4}	{5 or 6}
P(Y)	1/3	1/3	1/3	-1/3

In this case, the outcomes are not disjoint and one of the probabilities is negative, so this is *not* a valid probability distribution.

Probability distributions look a lot like relative frequency distributions. This isn't a coincidence! In fact, a relative frequency distribution is a good way to use data to approximate a probability distribution.

Section Exercises

1. Consider events A: "Draw a spade", B: "Draw a queen", and C: "Draw a red". Which of these events are mutually exclusive?
2. Use the probability axioms to determine whether each of the following is a valid probability distribution:
   A.

x	0	1	2	3
P(x)	0.1	0.2	0.1	0.3

   B.

x	0 or 1	1 or 2	3 or 4	5 or 6
P(x)	0.1	0.2	0.4	0.3

3. Determine whether the following events are mutually exclusive (disjoint).
   a. Your friend studies in the library. You study at home.
   b. You and your study group all earn As on an exam.
   c. You stay out until 3 am. You go to bed at 9 pm.
4. In a group of 24 people, 13 have cats and 15 have dogs. Four of them have both cats and dogs. Sketch a Venn Diagram for these events.

4.3 Rules of Probability

Consider a six-sided die.

$$P(\text{roll a 1 or 2}) = \frac{2 \text{ ways}}{6 \text{ outcomes}} = \frac{1}{3}.$$

Notice that we get the same result by taking

$$P(\text{roll a 1}) + P(\text{roll a 2}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

It turns out this is widely applicable!

4.3.1 Addition Rules

Addition Rule for Disjoint Outcomes

If A and B are disjoint outcomes, then the probability that one of them occurs is

$$P(A \text{ or } B) = P(A) + P(B).$$

This can also be extended to more than two disjoint outcomes:

$$P(A \text{ or } B \text{ or } C \text{ or } ...) = P(A) + P(B) + P(C) + ...$$

where we add up all of their individual probabilities.

Example:

Class	Frequency
freshman	12
sophomore	10
junior	3
senior	5

In a previous example, we found that

$$P(\text{junior or senior}) = \frac{3+5}{30} = \frac{8}{30}$$

Using the Addition Rule for Disjoint Outcomes, we get

$$P(\text{junior}) + P(\text{senior}) = \frac{3}{30} + \frac{5}{30} = \frac{8}{30}$$

> Essentially, the Addition Rule for Disjoint Outcomes is just breaking up that fraction: $\frac{3+5}{30}$ (3 juniors plus 5 seniors out of 30 students) represents the same thing as $\frac{3}{30} + \frac{5}{30}$ (3 juniors out of 30 students plus 5 seniors out of 30 students).

Now consider a deck of cards. Let A be the event that a card drawn is a diamond and let B be the event it is a face card. (Check back to 3.2 for the Venn Diagram of these events.)

- A: $2\diamondsuit$ $3\diamondsuit$ $4\diamondsuit$ $5\diamondsuit$ $6\diamondsuit$ $7\diamondsuit$ $8\diamondsuit$ $9\diamondsuit$ $10\diamondsuit$ $J\diamondsuit$ $Q\diamondsuit$ $K\diamondsuit$ $A\diamondsuit$
- B: $J\heartsuit$ $Q\heartsuit$ $K\heartsuit$ $J\clubsuit$ $Q\clubsuit$ $K\clubsuit$ $J\diamondsuit$ $Q\diamondsuit$ $K\diamondsuit$ $J\spadesuit$ $Q\spadesuit$ $K\spadesuit$

The collection of cards that are diamonds or face cards (or both) is

$A\diamondsuit$ $2\diamondsuit$ $3\diamondsuit$ $4\diamondsuit$ $5\diamondsuit$ $6\diamondsuit$ $7\diamondsuit$ $8\diamondsuit$ $9\diamondsuit$ $10\diamondsuit$ $J\diamondsuit$ $Q\diamondsuit$ $K\diamondsuit$ $J\clubsuit$ $Q\clubsuit$ $K\clubsuit$ $J\heartsuit$ $Q\heartsuit$ $K\heartsuit$ $J\spadesuit$ $Q\spadesuit$ $K\spadesuit$

Looking at these cards, I can see that there are 22 of them, so

$$P(A \text{ or } B) = \frac{22}{52}$$

However, if I try to apply the addition rule for disjoint outcomes, $P(A) = \frac{13}{52}$ and $P(B) = \frac{12}{52}$ and I would get $\frac{13+15}{52} = \frac{25}{52}$, which isn't what we want!

What happened? When I tried to add these, I *double counted* the Jack of Diamonds, Queen of Diamonds, and King of Diamonds (the cards that are in both A and B). To deal with that, I need to subtract off the double count $\frac{13}{52} + \frac{12}{52} - \frac{3}{52}$.

General Addition Rule

For any two events A and B, the probability that *at least* one will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Because A and B are just placeholders for some events, $P(A \text{ or } B) = P(B \text{ or } A)$.

Notice that when we say "or", we include the situations where A is true, B is true, and the situation where are both A and B are true. This is an *inclusive or*. Basically, if I said "Do you like cats or dogs?" and you said "Yes." because you like cats *and* dogs, that would be a perfectly valid response. I recommend using the inclusive or with your friends any time you want to get out of making a decision.

Also notice that the general addition rule applies to *any* two events, even disjoint events. This is because, for disjoint events, $P(A \text{ and } B) = 0$; it's impossible for both to occur at the same time!

4.3.2 Contingency Tables

A **contingency table** is a way to summarize **bivariate data**, or data from two variables.

Smallpox in Boston (1726)

Inoculated

yes

no

total

Result

lived

238

5136

5374

died

6

844

850

total

244

5980

6224

5136 is the count of people who lived AND were not inoculated.

6224 is the total number of observations.

244 is the total number of people who were inoculated.

5374 is the total number of people who lived.

This is basically a two-variable frequency distribution. And, like a frequency distribution, we can convert to proportions (relative frequencies) by dividing each count (each number) by the total number of observations:

Inoculated

yes

no

total

Result

lived

0.0382

0.8252

0.8634

died

0.0010

0.1356

0.1366

total

0.0392

0.9608

1.0000

0.8252 is the proportion of people who lived AND were not inoculated.

1.000 is the proportion of total number of observations. Think of this as 100% of the observations.

0.0392 is the proportion of people who were inoculated.

0.8634 is the proportion of people who lived.

The row and column totals are **marginal probabilities**. The probability of two events together (A and B) is a **joint probability**.

If we separate out the marginal probabilities, we get the relative frequency distribution for that variable.

Result

Proportion

lived

0.8634

died

0.1366

Example: Find the probability an individual was inoculated or lived.

Let A be the event that inoculated = yes and let B be the event that result = lived. Then $P(A) = 0.0392$ and $P(B) = 0.8632$.

We know these events are *not* disjoint, since there are 238 people who both lived and were inoculated (so clearly it is possible for both to be true at once).

Since they are not disjoint, we use the addition rule for disjoint outcomes.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

We can find $P(A \text{ and } B)$ from the table where the two events overlap: $P(A \text{ and } B) = 0.0382$ So then

$$P(A \text{ or } B) = 0.0392 + 0.8634 - 0.0382 = 0.8644$$

That is, the probability an individual was incolated or lived is 0.8644.

Aside: the history of the 1721 smallpox epidemic in Boston is super interesting! You can read a bit more about it here.

4.3.3 Complements

The **complement** of an event is all of the outcomes in the sample space that are *not* in the event. For an event A, we denote its complement by A^c.

Example: For a single roll of a six-sided die, the sample space is all possible rolls: 1, 2, 3, 4, 5, or 6. If the event A is rolling a 1 or a 2, then the complement of this event, denoted A^c, is rolling a 3, 4, 5, or 6.

We could also write this in probability notation: $S = \{1, 2, 3, 4, 5, 6\}$ and if $A = \{1, 2\}$, then $A^c = \{3, 4, 5, 6\}$.

Property:

$$P(A \text{ or } A^c) = 1$$

Using the addition rule,

$$P(A \text{ or } A^c) = P(A) + P(A^c) = 1.$$

Make sure you can convince yourself that A and A^c are *always* disjoint.

Complement Rule

$$P(A) = 1 - P(A^c).$$

Example: Consider rolling 2 six-sided dice and taking their sum. The event of interest is a sum less than 12. Find

1. A^c
2. $P(A^c)$
3. $P(A)$

If $A =$ (sum less than 12), then $A^c =$ (sum greater than or equal to 12). Take a moment to notice that there is only one way to get a sum greater than or equal to 12: rolling two 6s.

The chart below shows the rolls of Die 1 as columns and the rolls for Die 2 as rows. The numbers in the middle are the sums. Note that there are 36 possible ways to roll 2 dice.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Even without the chart, by noting that there's only one way to get a sum greater than or equal to 12, we can quickly find $P(A^C)$:

$$P(A^c) = \frac{1}{36}$$

But trying to count all of the ways to get A would take a long time! Instead, we can use

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{36} = \frac{35}{36}$$

Section Exercises

1. Consider rolling two four-sided dice and taking their sum.
   a. Find the probability that the sum is 7 or 8.
   b. Find the probability that the sum is less than 8.
   c. Let A be the event that the sum is 8.
      i. Find $P(A)$.

   ii. Describe the event A^c?
   iii. Find $P(A^c)$.
2. Consider a deck of cards (there's an image in Section 4.2.1). If you randomly draw a single card from the deck, what is the probability that it is a Queen or a Heart?
3. Suppose you are playing a game that requires you to roll twenty-sided dice. We are interested in the setting where you roll two of these dice (call them red and blue) and take the *highest* of the two rolls. We want to find the probability that your highest roll is at least a 17.
   a. Let A be the event that red is at least a 17. Then, find $P(A)$.
   b. Let B be the probability that blue is at least 17. What is $P(B)$?
   c. How can we rewrite "the probability that your highest roll is at least a 17" in terms of A and B? Hint: What needs to happen with red and blue for your *highest* roll to be at least a 17?
   d. Find the probability that your highest roll is at least a 17.

4.4 Conditional Probability

Let's return to our data on smallpox in Boston. We had the initial contingency table.

Inoculated

yes

no

total

Result

lived

238

5136

5374

died

6

844

850

total

244

5980

6224

and the relatively frequency version.

Inoculated

yes

no

total

Result

lived

0.0382

0.8252

0.8634

died

0.0010

0.1356

0.1366

total

0.0392

0.9608

1.0000

What can we learn about the result of smallpox if we already know something about inoculation status? For example, given that a person is inoculated, what is the probability of death? To figure this out, we restrict our attention to the 244 inoculated cases. Of these, 6 died. So the probability is 6/244.

This is called **conditional probability**, the probability of some event A if we know that event B occurred (or is true):

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

where the symbol | is read as "given".

Example: For death given inoculation,

$$P(\text{death}|\text{inoculation}) = \frac{P(\text{death and inoculation})}{P(\text{inoculation})} = \frac{0.0010}{0.0392} = 0.0255.$$

Notice that we could also write this as

$$P(\text{death}|\text{inoculation}) = \frac{P(\text{death and inoculation})}{P(\text{inoculation})} = \frac{6/6224}{244/6224} = \frac{6}{244},$$

which is what we found when using the table to restrict our attention to only the inoculated cases.

If knowing whether event B occurs tells us nothing about event A, the events are **independent**. For example, if we know that the first flip of a (fair) coin came up heads, that doesn't tell us anything about what will happen next time we flip that coin.

We can test for independence by checking if $P(A|B) = P(A)$.

4.4.1 Multiplication Rules

Multiplication Rule for Independent Processes

If A and B are independent events, then

$$P(A \text{ and } B) = P(A)P(B).$$

We can extend this to more than two events:

$$P(A \text{ and } B \text{ and } C \text{ and } \dots) = P(A)P(B)P(C)\dots.$$

Note that if $P(A \text{ and } B) \neq P(A)P(B)$, then A and B are *not* independent.

Example: Find the probability of rolling a 6 on your first roll of a die and a 6 on your second roll.

Let $A =$ (rolling a 6 on first roll) and $B =$ (rolling a 6 on second roll). For each roll, the probabiltiy of getting a 6 is 1/6, so $P(A) = \frac{1}{6}$ and $P(B) = \frac{1}{6}$.

Then, because each roll is independent of any other rolls,

$$P(A \text{ and } B) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

General Multiplication Rule

If A and B are any two events, then

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

Because A and B are just placeholders for some events, $P(A \text{ and } B) = P(B \text{ and } A)$.

Notice that this is just the conditional probability formula, rewritten in terms of $P(A \text{ and } B)$!

> **Example**: Suppose we know that 45.5% of US households have dogs and that among those with dogs, 12.1% have cats. Find the probability that a US household has both dogs and cats.
>
> Let $C =$ (household has cats) and $D =$ (household has dogs). Since 45.5% of US households have dogs, if I randomly select a household, 45.5% of the time that household will have dogs. Thus, $P(D) = 0.455$.
>
> The other piece tells us something about the probability of having cats *among those with dogs*. This means that we *know* that these people have dogs. That is, *given* a household has dogs, the probability of cats is 12.1%. In probability notation, $P(C|D) = 0.121$. Then
>
> $$P(C \text{ and } D) = P(C|D)P(D) = 0.121 \times 0.455 = 0.055$$
>
> or the probability that a US household has both cats and dogs is 0.055.

Tests for Independence

We can put our idea of independence together with our multiplication rules to come up with three ways to test for independence:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ and } B) = P(A)P(B)$

These are all mathematically equivalent, so you only need to check one. If the equation holds (they are equal), then A and B are independent. If they are not equal, the events are dependent.

Any time we want to determine whether two events are independent, we need to do one of these tests! That is, we should not rely solely on our intuition to determine if events are independent.

Example: Suppose, in addition to the information from the previous example, that 32.1% of US households have cats. For US households, are having cats and having dogs independent events?

Now we can include the information $P(C) = 0.321$. We know from the previous example that $P(C|D) = 0.121$, so it will be convenient to test

$$P(C|D) \stackrel{?}{=} P(C)$$

(That question mark means we aren't sure yet if those two things are equal!) In this case

$$P(C|D) = 0.121 \neq P(C) = 0.321$$

so they are *dependent*.

4.4.2 Law of Total Probability

We observed in the last section that we can get a marginal probability by adding up all of its associated joint probabilities in a contingency table.

Inoculated

yes

no

total

Result

lived

0.0382

0.8252

0.8634

died

0.0010

0.1356

0.1366

total

0.0392

0.9608

1.0000

That is,

$$P(\text{lived}) = P(\text{lived and inoculated}) + P(\text{lived and not inoculated})$$

We can generalize this for some event A and some variable X whose k possible outcomes can be listed as (mutually exclusive) events $B_1, B_2, \ldots B_k$. Then

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k)$$

Using our General Multiplication Rule, we have $P(A \text{ and } B) = P(A|B)P(B)$, so we can rewrite this as

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)$$

4.4.3 Bayes' Theorem

Sometimes, we get some additional information about a probability and we want to *update* our understanding based on this new information. This the basic idea behind Bayes' Theorem.

Our conditional probability rule states that

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

and our multiplication rule that

$$P(A \text{ and } B) = P(B \text{ and } A) = P(B|A)P(A)$$

Putting these together (by substituting $P(B|A)P(A)$ in for $P(A \text{ and } B)$ in the conditional probability rule) gives us one last probability rule.

Bayes' Theorem

If A and B are any two events such that $P(B) \neq 0$, then

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

Example: Consider a test for some rare disease. If you have the disease, this test accurately identifies it 99% of the time. Suppose 0.5% of the population has this disease and the test results in a positive result 8% of the time. . Given a positive test result, what's the probability a person actually has the disease?

Solution: Let's start by assigning some probabilities to events. Let D be the event a person has the disease and let T be the probability the test is positive. From the prompt, we know that $P(D) = 0.005$ (0.5% of the population has the disease) and $P(T) = 0.08$ (the test gives a positive result 8% of the time).

The other probabilities represent the test being accurate *given* some disease status. That is, given a person has the disease, the test will be positive 99% of the time: $P(T|D) = 0.99$.

We want to know the probability of disease *given* positive test, $P(D|T)$. Using Bayes' Theorem,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{0.99 \times 0.005}{0.08} = 0.062$$

So if a person gets a positive result, they actually only have a 0.062 probability of having the disease (even though the test is very accurate!).

This may all seem a bit counterintuitive. The probability of disease given positive result seems really low! That's because most people who actually get tested have some reason to suspect they have the disease. If doctors started testing absolutely everyone, we would probably see a lot more false positives. This is why we typically only do routine screening for things like cancer for people who are already at risk.

The Monty Hall Problem

This is a neat problem with an unexpected result. This is a brain teaser based loosely on the game show *Let's Make a Deal* hosted at that time by Monty Hall. The problem is as follows:

Suppose you are participating in a game show where you are given the choice of one of three doors. Behind one of the doors is a prize. The other two doors have nothing behind them. After you choose a door, the host opens one of the other two doors, which always has nothing behind it. He then asks if you want to stick with your original door, or switch to the other unopened door. What should you do? Does it matter?

Most people's instinct is that it does not matter and that, after the host opens the door, there is a 50% chance of winning the prize either way. However, this problem turns out to be a *little* more complicated than that.

This time, let's start with the answer: you should switch doors. In fact, after the host has opened the door, the probability that the prize is behind your original door is 1/3, but the probability of its being behind the other unopened door is 2/3!

The host must show you a door with nothing behind it, and he must open one of the doors you did not select, so he is *not* choosing this door at random. That impacts what you should do.

Hard to believe? We can set this up using a table or with Bayes' Theorem. Let's start with the table. Suppose you initially selected Door 1.

Behind door 1

Behind door 2

Behind door 3

Result of staying

Result of switching

nothing

nothing

prize

loses

wins

nothing

prize

nothing

loses

wins

prize

nothing

nothing

wins

loses

Looking under "Result of staying", we can see we would lose 2/3 of the time (and we win 2/3 of the time if we switch)!

We will also demonstrate this using Bayes' Theorem. Suppose again you start with Door 1. Let A_i be the event that the prize is behind Door i. Then, without knowing anything else, $P(A_1) = P(A_2) = P(A_3) = 1/3$ since there are three doors.

We will let D be the event the host opened Door 2.

To get into $P(D)$, we need our Law of Total Probability. We selected Door 1, so

- if the prize is behind Door 1, the host will randomly select between Door 2 and Door 3. That is, $P(D|A_1) = \frac{1}{2}$.
- if the prize is behind Door 2, the host will always show us Door 3. That is, $P(D|A_2) = 0$.
- if the prize is behind Door 3, the host will always show us Door 2. That is, $P(D|A_3) = 1$.

So

$$\begin{aligned} P(D) &= P(D|A_1)P(A_1) + P(D|A_2)P(A_2) + P(D|A_3)P(A_3) \\ &= \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} \\ &= \frac{1}{2} \end{aligned}$$

Then the probability the prize is behind Door 1 (the one we selected), given the host opened Door 2, is

$$P(A_1|D) = \frac{P(D|A_1)P(A_1)}{P(D)} = \frac{1/2 \times 1/3}{1/2} = \frac{1}{3}$$

So if we stay, we have a 0.3333 chance of winning. Similarly, the probability the prize is behind Door 3, given the host opened Door 2, is

$$P(A_3|D) = \frac{P(D|A_3)P(A_3)}{P(D)} = \frac{1 \times 1/3}{1/2} = \frac{1}{3} = \frac{2}{3}$$

So if we switch, we have a 0.6667 change of winning.

Section Exercises

For the following conditional probability scenarios, determine the **condition** and the **outcome of interest**.

We want to find the probability that a randomly chosen student is enrolled in Stat 1, given that they are a health sciences major.

Knowing that a person texts while driving, we will find the probability that they got a speeding ticket in 2023.

We have data on the type of car people bought and on what additional features they purchased. We will find the probability that a randomly selected sports car buyer also bought bucket seats.

We have data on sex and paw preference for dogs. We will find the probability that a female dog is left-pawed.

Using the contingency table, find the indicated probabilities.

	A	A^c	Total
$B1$	0.31	0.05	0.36
$B2$	0.11	0.13	0.24
$B3$	0.08	0.32	0.40
Total	0.5	0.5	1.00

$P(A \text{ and } B2)$

$P(B2)$

$P(A)$

$P(A|B2)$

$P(B2|A)$

The following contingency table represents a sample of households asked about whether they had children and whether they had pets. Use the table to answer the following questions.

	Children	No children	Total
Pets	38	47	85
No pets	21	44	65
Total	59	91	150

Find the probability that a household has pets.

Find the probability a household has pets *and* has children.

Knowing that a household has children, find the probability they have pets.

Are having children and having pets independent events? Explain.

If A and B are disjoint events, what can we say about whether they are independent? Hint: think about how disjoint events will impact our tests for independence.

Chapter 5

Random Variables

In previous modules, we introduced the idea of variables and examined their distributions. We also began our discussion on probability theory. Now, we extend these concepts into what are called random variables. We will introduce the concept of random variables in general and will discuss a specific type of distribution - the binomial distribution. Then we will discuss a continuous probability distribution, the normal distribution. The normal distribution will provide a foundation for much of the inference we will complete throughout the rest of this course.

Module Learning Objectives/Outcomes

1. Discuss discrete random variables using key terminology.
2. Express cumulative probabilities using probability notation.
3. Calculate the expected value and standard deviation of a discrete random variable.
4. Calculate binomial probabilities.
5. Use z scores to compare observations on different scales.
6. Calculate probabilities for a normal distribution using area under the curve.
7. Calculate normal distribution percentiles.

R Objectives

1. Calculate binomial probabilities.
2. Find cumulative probabilities for the standard normal distribution.
3. Find percentiles.

This module's outcomes correspond to course outcomes (4) use the binomial distribution as a model for discrete variables and (5) use the normal distribution as a model for continuous variables.

5.1 Discrete Random Variables

A **random variable** is a quantitative variable whose values are based on chance. By "chance", we mean that you can't *know* the outcome before it occurs.

A **discrete random variable** is a random variable whose possible values can be listed.

Notation:

- x,y,z (lower case letters) denote variables.
- X, Y, Z (upper case letters) denote *random* variables.

In contrast to events, where we usually used letters toward the start of the alphabet, (random) variables are typically denoted by letters from the end of the alphabet.

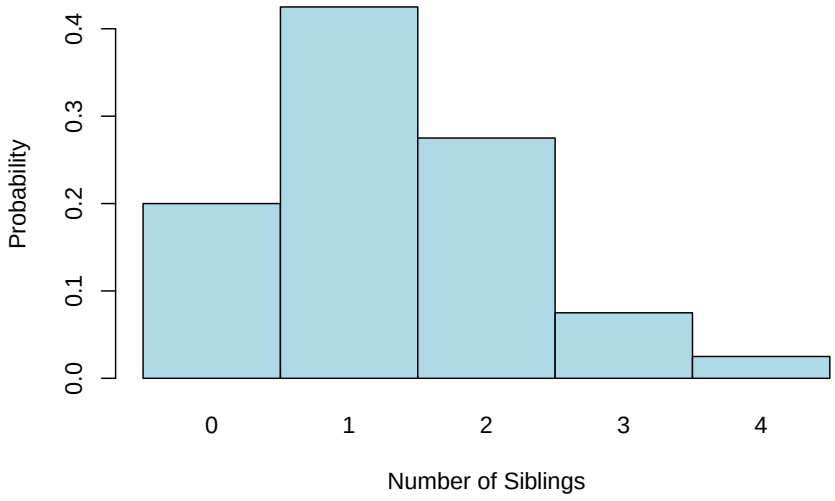
- $\{X = x\}$ denotes the event that the random variable X equals x.
- $P(X = x)$ denotes the probability that the random variable X equals x.

Recall: a probability distribution is a list of all possible values and their corresponding probabilities. (See Section 3.3 for a refresher.) A **probability histogram** is a histogram where the heights of the bars correspond to the probability of each value. (This is very similar to a relative frequency histogram!) For discrete random variables, each "bin" is one of the listed values.

Example:

Number of Siblings, x	0	1	2	3	4
Probability, $P(X = x)$	0.200	0.425	0.275	0.075	0.025

(Assume for the sake of the example that no one has more than 4 siblings.)



Interpretation: in a large number of independent observations of a random variable X, the proportion of times each possible value occurs will approximate the probability distribution of X.

5.1.1 The Mean and Standard Deviation

Mean of a Discrete Random Variable

The mean of a discrete random variable X is denoted μ_X. If it's clear which random variable we're talking about, we can drop the subscript and write μ.

$$\mu_X = \Sigma x P(X = x)$$

where Σ denotes "the sum over all values of x":

$$\Sigma x P(X = x) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_n P(X = x_n).$$

The mean of a random variable is also called the **expected value** or **expectation**. Recall that measures of center are meant to identify the most common or most likely, thus the value we can *expect* to see (most often).

Example: for the Siblings distribution,

$$\mu = 0(0.200) + 1(0.425) + 2(0.275) + 3(0.075) + 4(0.025) = 1.3$$

Make sure you understand how we used the formula for μ and the probability distribution to come up with this number.

The random variable X represents draws from some population, so μ is just the population mean. In the above example, the mean number of siblings (for people in that population) is 1.3.

The larger the number of observations, the closer their average tends to be to μ. This is known as the **law of large numbers**.

> **Example**: Suppose I took a random sample of 10 people and asked how many siblings they have.
>
> $$2, 2, 2, 2, 1, 0, 3, 1, 2, 0$$
>
> In my random sample of 10, $\bar{x} = 2$, which is a reasonable estimate but not that close to the true mean $\mu = 1.3$.
>
> - A random sample of 30 gave me a mean of $\bar{x} = 1.53$.
> - A random sample of 100 gave me a mean of $\bar{x} = 1.47$.
> - A random sample of 1000 gave me a mean of $\bar{x} = 1.307$.

We use concepts related to the law of large numbers as a foundation for statistical inference, but note that - although very large samples are nice to have - it's not necessary to take enormous samples all the time. Often, we can come to interesting conclusions with fewer than 30 observations!

Standard Deviation of a Discrete Random Variable

The variance of a discrete random variable X is denoted σ_X^2 (or σ^2 if it's clear which variable we're talking about).

$$\sigma_X^2 = \Sigma[(x - \mu_X)^2 P(X = x)]$$

OR

$$\sigma_X^2 = \Sigma[x^2 P(X = x)] - \mu_X^2$$

These formulas are *exactly* equivalent and you may use whichever you wish, but note that the second may be a little easier to work with.

As before, the standard deviation is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

> **Example**: Calculate the standard deviation of the Siblings variable.
>
> In general, a table is the best way to keep track of a variance calculation:

x	$P(X = x)$	$xP(X = x)$	x^2	$x^2 P(X = x)$
0	0.200	0	0	0
1	0.425	0.425	1	0.425
2	0.275	0.550	4	1.100
3	0.075	0.225	9	0.675

x	$P(X = x)$	$xP(X = x)$	x^2	$x^2P(X = x)$
4	0.025	0.100	16	0.400
		$\mu = 1.3$		Total = 2.6

Then the variance is

$$\sigma^2 = 2.6 - 1.3^2 = 0.9$$

and the standard deviation is

$$\sigma = \sqrt{0.9} = 0.9539.$$

Section Exercises

Consider the following probability distribution for some discrete random variable.

x	0	1	2	3	4
P(x)	0.1	0.2	0.3	0.3	0.1

Use probability notation to write the probability that X is less than 2. Then, use the table to find this probability.

Find the expected value of the random variable X.

Find the standard deviation of the random variable.

Consider the probability distribution of the number of dogs owned by people in the US. Assume for the sake of the problem that nobody has more than three dogs.

Number of dogs, x	0	1	2	3
Proportion of people, $p(x)$	0.635	0.212	0.131	0.022

Find the mean number of dogs owned by people in the US.

For a randomly selected household, what is the probability they have at least 2 dogs?

5.2 The Binomial Distribution

The binomial distribution is used to describe the number of successes in some fixed number of trials (or some set sample size). > **Example**: Suppose we will

roll three dice: a green (G), a blue (B), and a red (R). What is the probability that exactly one of them will result in a 1? $>>$ Let's start by considering one specific scenario where exactly one die results in a 1: $>$

$$\begin{aligned} P(G = 1 \text{ and } B \neq 1 \text{ and } R \neq 1) &= P(G = 1)P(B > 1)P(R > 1) \\ &= \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} \\ &= (0.1667)(0.8333)(0.8333) \\ &= (0.1667)^1(0.8333)^2 \\ &= 0.1157 \end{aligned}$$

$>$ Note that "not 1" in this scenario is the same as "greater than 1" and includes the values 2, 3, 4, 5, and 6, which is how we come up with those individual probabilities. Also notice that our first step was to use the multiplication rule for independent events! $>>$ But we also could have considered the scenarios where the Blue or Red die was the single die to roll a 1. In each case, the probability is the same: $(0.1667)^1(0.8333)^2$ so we need to multiply this probability by 3:

$$P(\text{exactly one four-sided die rolls a 1}) = 3 \times (0.1667)^1(0.8333)^2 = 0.3472$$

Next, think back to replication in an experiment. Each replication is what we call a **trial**. We will consider a setting where each trial has two possible outcomes.

> **Example**: Suppose you want to know if a coin is fair (both sides equally likely). You might flip the coin 100 times (thus running 100 trials). Each trial is a flip of the coin with two possible outcomes: heads or tails.

The product of the first k positive integers $(1, 2, 3, \dots)$ is called **k-factorial**, denoted $k!$:

$$k! = k \times (k-1) \times \cdots \times 3 \times 2 \times 1$$

We define $0! = 1$.

> **Example**: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

If n is a positive integer $(1, 2, 3, \dots)$ and x is a nonnegative integer $(0, 1, 2, \dots)$ with $x \leq n$, the **binomial coefficient** is

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

> **Example**:
>
> $$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)}$$

The binomial coefficient represents the number of ways to select x objects from a group of n objects, *without replacement*. When we say "without replacement", we mean that we cannot select an object more than once. (Note that this implies a binomial coefficient should *always* result in a whole number!) This turns out to be a handy thing to be able to calculate when working with probability and random variables.

> **Example**: When we worked out the probability of rolling exactly one 1 when rolling three dice, we found that there were 3 possible ways to do so. Using the binomial coefficient, $n = 3$ rolls and $x = 1$ roll of a 1, so
>
> $$\binom{3}{1} = \frac{3!}{1!(3-2)!} = \frac{3 \times 2 \times 1}{(1) \times (2 \times 1)} = \frac{6}{2} = 3$$
>
> In this case, it was fairly straightforward to list all the ways to roll exactly one 1, but sometimes it's not so simple and the binomial coefficient will save us a lot of headache.

Sometimes, we may want to simplify a binomial coefficient *before* taking all of the factorials. Why? Well,

$$20! = 2,432,902,008,176,640,000$$

Most calculators will not print this number! Instead, you'll get an error or a rounded version printed using scientific notation. Neither will help you accurately calculate the binomial coefficient.

> **Example**:
>
> $$\binom{20}{17} = \frac{20!}{17!3!} = \frac{20 \times 19 \times 18 \times 17 \times 16 \times \cdots \times 3 \times 2 \times 1}{(17 \times 16 \times \cdots \times 3 \times 2 \times 1)(3 \times 2 \times 1)}$$
>
> but notice that I can rewrite 20! as $20 \times 19 \times 18 \times 17!$, so
>
> $$\binom{20}{17} = \frac{20 \times 19 \times 18 \times 17!}{17!(3 \times 2 \times 1)} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = \frac{6840}{6} = 1140$$

Bernoulli trials are repeated trials of an experiment where:

1. Each trial has two possible outcomes: success and failure.
2. Trials are independent.
3. The probability of success (the **success probability**) p remains the same from one trial to the next:

$$P(X = \text{success}) = p$$

The **binomial distribution** is the probability distribution for the number of successes in a sequence of Bernoulli trials.

Fact: in n Bernoulli trials, the number of outcomes that contain exactly x successes equals the binomial coefficient $\binom{n}{x}$.

Example: Find the probability of rolling exactly one 1 in three dice rolls.

In this scenario, we are primarily interested in an outcome related to rolling a 1, so we will let this be our *success*. Rolling exactly one 1 means observing exactly one success. The other possibility (to rolling a 1) is *not* rolling a 1, so this must be our failure. From our previous examples, we know this probability can be written as

$$\binom{3}{1}(0.1667)^1(0.8333)^2$$

- $\binom{3}{1}$ is the number of ways to observe 1 success in 3 trials, or rolls of the dice.
- 0.1667 is the probability of success.
- The 1 in the binomial coefficient and the power is the number of successes.
- 0.8333 is the probability of failure.
  - By the complement rule, this value is equal to $1 - 0.1667$ (one minus the probability of success).
- The 2 in the power is the number of failures.
  - If each trial is a success or a failure, and we are interested in 1 success out of 3 failures, then the other $3-1 = 2$ trials must be failures.

This is a binomial probability! Let's see how this becomes the general binomial probability formula.

Binomial Probability Formula

Let x denote the total number of successes in n Bernoulli trials with success probability p. The probability distribution of the random variable X is given by

$$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

The random variable X is called a **binomial random variable** and is said to have the **binomial distribution**. Because n and p fully define this distribution, they are called the distribution's **parameters**.

To find a binomial probability formula:

Check assumptions.

Exactly n trials to be performed.

Two possible outcomes for each trial.

Trials are independent (each trial does not impact the result of the next)

Success probability p remains the same from trial to trial.

Identify a "success". Generally, this is whichever of the two possible outcomes we are most interested in.

Determine the success probability p.

Determine n, the number of trials.

Plug n and p into the binomial distribution formula.

Example: Approximately 44.5% of US households have dogs. If we are to take a random sample of 10 US households, what is the probability that exactly 4 of them will have dogs?

Solution:

1. Assumptions: (1) We will take a random sample of 10 households, so there are exactly 10 trials to be performed. (2) The two possible outcomes are "dogs" or "no dogs". (3) Since this is a random sample, we can assume that trials are independent. (4) The probability of a US household having a dog is (always) 0.445, so the success probability is constant.
2. We will let success = household has dog(s) since the probability of interest relates to having dogs. Therefore failure = household does not have dogs.
3. $P(\text{success}) = 0.445$ from the problem statement.
4. We will take a random sample of 10, so our number of trials is $n = 10$.
5. The formula looks like

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{10}{x} 0.445^x (1-0.445)^{10-x}$$

For exactly 4 successes, we set $x = 4$:

$$\begin{aligned} P(X = 4) &= \binom{10}{4} 0.445^4 (1-0.445)^{10-4} \\ &= \binom{10}{4} 0.445^4 (0.555)^6 \end{aligned}$$

Now, let's take a moment to work through that binomial coefficient:

$$\begin{aligned} \binom{10}{4} &= \frac{10!}{4!6!} \\ &= \frac{10 \times 9 \times 8 \times 7 \times 6!}{(4 \times 3 \times 2 \times 1) \times 6!} \\ &= \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} \\ &= \frac{5040}{24} \\ &= 210 \end{aligned}$$

Plugging that into our binomial probability formula,

$$\begin{aligned} P(X = 4) &= 210 \times 0.445^4 \times 0.555^6 \\ &= 210 \times 0.0392 \times 0.0292 \\ &= 0.241 \end{aligned}$$

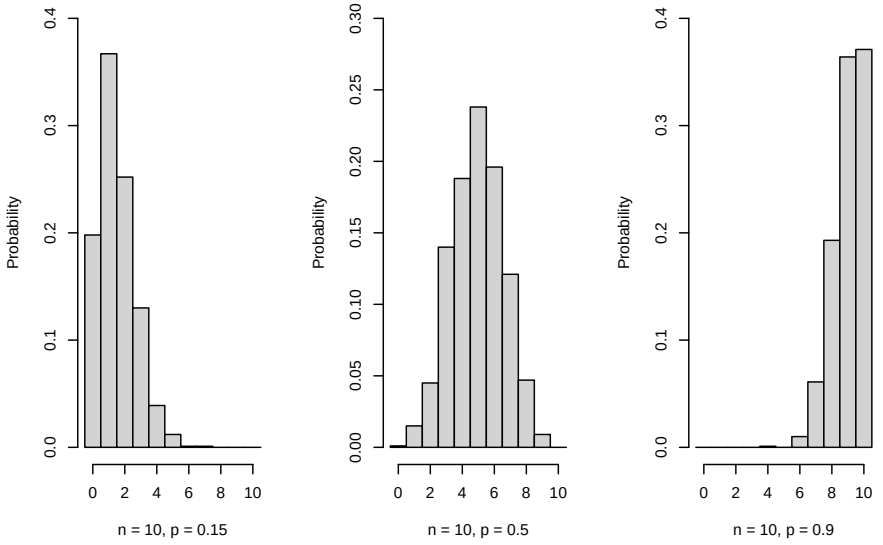
So the probability that exactly 4 of the 10 US households have dog(s) is 0.241.

5.2.1 Expected Value and Variance

The expected value (mean) of a binomial distribution is $\mu = np$. The variance is $\sigma^2 = np(1-p)$.

The shape of a binomial distribution is determined by the success probability:

- If $p \approx 0.5$, the distribution is approximately symmetric.
- If $p < 0.5$, the distribution is right-skewed.
- If $p > 0.5$, the distribution is left-skewed.



5.2.2 Probabilities with Inequalities

We can also use the binomial probability formula to calculate probabilities like $P(X \leq x)$. Notice that we can rewrite this using concepts from the previous module

$$P(X \leq k) = P(X = k \text{ or } X = k-1 \text{ or } \ldots \text{ or } X = 2 \text{ or } X = 1 \text{ or } X = 0)$$

Since X is a discrete random variable, each possible value is *disjoint*. We can use this!

$$P(X \leq k) = P(X = k) + P(X = k-1) + \cdots + P(X = 2) + P(X = 1) + P(X = 0)$$

Example:

$$\begin{aligned} P(X \leq 3) =& P(X = 3 \text{ or } X = 2 \text{ or } X = 1 \text{ or } X = 0) \\ =& P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0) \end{aligned}$$

We can also extend this concept to work with probabilities like $P(a < X \leq b)$.

Example: $P(2 < X \leq 5)$

First, notice that if $2 < X \leq 5$, then X can be 3, 4, or 5:

$$\begin{aligned} P(2 < X \leq 5) &= P(X = 3 \text{ or } X = 4 \text{ or } X = 5) \\ &= P(X = 3) + P(X = 4) + P(X = 5) \end{aligned}$$

Note: if going from $2 < X \leq 5$ to "X can be 3, 4, or 5" doesn't make sense to you, start by writing out the sample space. Suppose $n = 10$. Then the sample space for the binomial distribution is

$$S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Then I can check any number in this sample space by plugging it in for X. So for 1, I can check $2 < 1 \leq 5$. Obviously this is not true, so we won't include 1. Checking the number 2, I get $2 < 2 \leq 5$. Since $2 < 2$ is NOT true, we don't include 2. Etc.

These concepts apply to any other type of inequality, too! Figure out which values in the sample space satisfy the inequality, apply the addition rule for disjoint outcomes, and add up the individual probabilities.

Section Exercises

Suppose you are going to roll a 6-sided die seven times. What is the probability of getting exactly three 6s?

Over the course of a board game, a player is going to roll a 10-sized die 30 times. Find the expected value and standard deviation for the number of 10s rolled.

In the 2024 NFL season, Gardner Minshew, quarterback for the Raiders, completed 66.3% of his pass attempts. Suppose we will randomly select 10 passes by Minshew and see if they are completed.

What distribution could you use to model the probability that all 10 pass attempts are completed? Go through each condition and confirm is it satisfied.

Find the mean and standard deviation number of completed passes.

Find the probability that exactly 5 of the pass attempts are completed.

Find the probability that all 10 of the pass attempts are completed.

Find the probability that none of the pass attempts are completed.

A restaurant has a reservation on the books for a 22 person birthday party. They know that 32% of their customers order dessert. The restaurant works with a local bakery to provide desserts, and they want to think about how many to order.

What distribution could we use to model the number of birthday party goers who order dessert? Justify your answer by checking any necessary conditions.

What are the expected value and standard deviation number of party goes who will order dessert?

Find the probability that no more than 3 people order dessert.

Find the probability that between 5 and 7 people (inclusive) order dessert.

If you flip a fair coin 5 times, what is the probability of getting at least one heads? Hint: what is the complement of getting at least one heads?

For smokers, the probability of developing a severe lung infection at some point in their lifetime is 0.3.

Suppose we take a random sample of 20 smokers. What distribution could you use to model the probability of some number of them developing a severe lung infection? Justify your answer by checking any necessary conditions.

Find the mean and standard deviation number of smokers (from the sample of 20) who will develop a severe lung infection during their lifetimes.

For our 20 smokers, find the probability that none of them develop a severe lung infection during their lifetimes.

Consider the probability distribution of the number of dogs owned by people in the US. (For the sake of the problem, assume no one has more than 3 dogs.)

Number of dogs, x	0	1	2	3
Proportion of people, $p(x)$	0.635	0.212	0.131	0.022

What proportion of people have at least one dog?

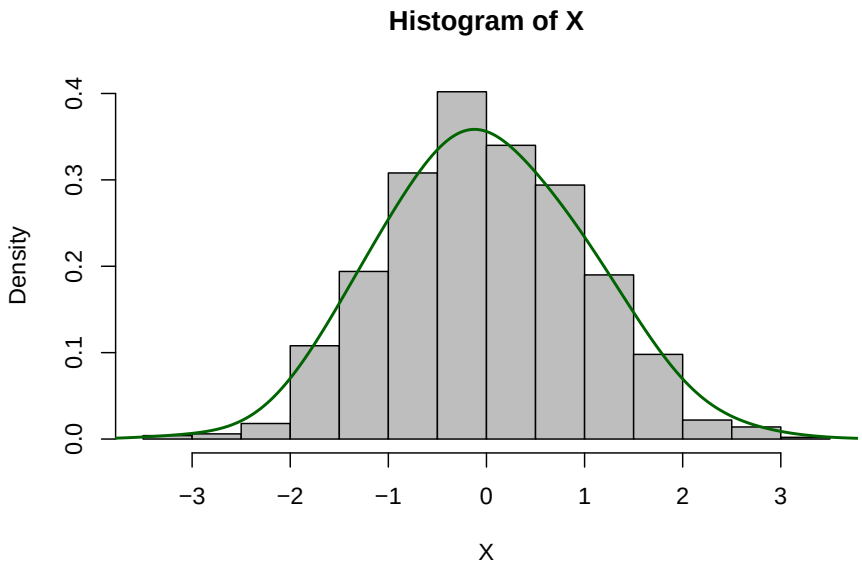
Consider two categories: (1) people with no dogs and (2) people with dog(s). A random sample of 10 people were asked about their dog ownership. Let Y be the random variable which counts number of people with (some number of) dogs. What is the distribution of Y? Justify your answer by checking any necessary conditions.

What is the probability that fewer than 3 people out of 10 will have dog(s)?

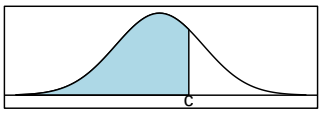
5.3 The Normal Distribution

If we can represent a discrete variable with a probability histogram, what can we do with a continuous variable?

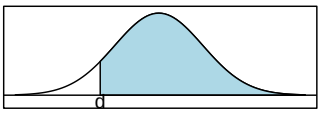
We represent the shape of a continuous variable using a **density curve**. This is like a histogram, but with a smooth curve:



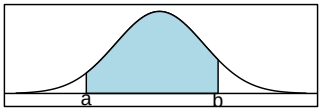
For a variable with a density curve, the proportion of all possible observations that lie within a specified range equals the corresponding area under the density curve.



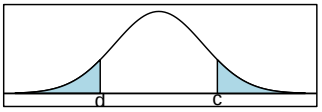
P(X < c)



P(X > d)



P(a < X < b)



P(X < c or X > d)

Properties of density curves:

1. The curve is always above the horizontal axis (because probabilities are always nonnegative).
2. The total area under the curve equals 1 (because $P(S) = 1$).

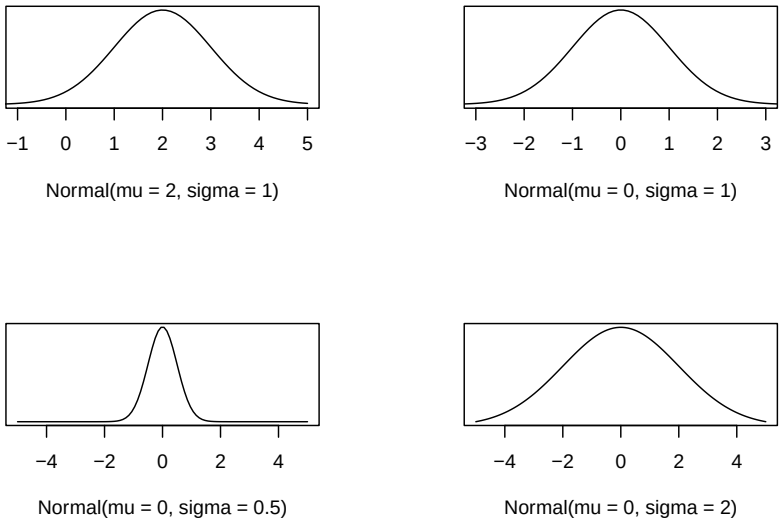
A **normal curve** is a special type of density curve that has a "bell-shaped" distribution. In fact, all of the density curves I've shown in this section have been normal curves! We say that a variable is **normally distributed** or has a **normal distribution** if its distribution has the shape of a normal curve.

Why "normal"? Because it appears so often in practice! Lots of things are more common around the average and less common as you get farther from the average: height, amount of sleep people get each night, standardized test scores, etc. (In practice, these things aren't *exactly* normally distributed… instead, they're **approximately normally distributed** - and that's ok.)

Normal distributions…

- have curves that extend indefinitely in both directions along the horizontal axis.
- are fully determined by parameters mean μ and standard deviation σ.
- are symmetric and centered at μ.
- have spreads that depend on σ.

Pay close attention to the horizontal axis and how spread out the densities are in each of the following plots:



Notice that the bottom left plot comes to a sharper peak, while the bottom right has a gentler slope. This is what we mean by "spread": the density on the bottom right is the most spread out.

Empirical Rule

For any (approximately) normally distributed variable,

1. Approximately 68% of all possible observations lie within one standard deviation of the mean: $\mu \pm \sigma$.
2. Approximately 95% of all possible observations lie within two standard deviations of the mean: $\mu \pm 2\sigma$.
3. Approximately 99.7% of all possible observations lie within three standard deviations of the mean: $\mu \pm 3\sigma$.

Given some data, you can check if approximately 68% of the data falls within $\bar{x} \pm s$, 95% within $\bar{x} \pm 2s$, and 99.7% within $\bar{x} \pm 3s$ to examine whether the data follow the empirical rule.

To check whether a variable is (approximately) normally distributed, we can check the histogram to see if it is symmetric and bell-shaped… or we can check to see if the variable conforms (approximately) to the empirical rule! If we decide it is approximately normal, we can estimate the parameters: μ using $\bar{x}$ and σ using s.

5.3.1 Z-Scores

We **standardize** a variable using

$$z = \frac{x - \mu}{\sigma}.$$

This is also called a **z-score**. Standardizing using this formula will *always* result in a variable with mean 0 and standard deviation 1 (even if it's not normal!).

Because z-scores always result in variables with mean 0 and standard deviation 1, they are also very useful for comparing values which are originally on different scales.

Note that a z-score tells us how many standard deviations an observation is from the mean. A positive z-score $z > 0$ is *above* the mean; a negative z-score $z < 0$ is *below* the mean. For example, if an observation has $z = -0.23$, that observation is 0.23 standard deviations below the mean.

> **Example**: ACT scores have mean 20.8 and standard deviation 5.8. SAT scores have mean 1500 and standard deviation 300. If Jose scored a 27 on his ACT and Navreet scored an 1850 on her SAT, who got a better score (relative to other test takers)?
>
> We cannot compare their scores directly because the ACT and SAT are on different scales. Instead, we will compare their z-scores.
>
> For the SAT, we are given $\mu = 1500$ and $\sigma = 300$. So Navreet's z-score is
>
> $$z_{\text{Navreet}} = \frac{1850 - 1500}{300} = 1.17$$
>
> standard deviations above the mean. Then for the ACT, we have that $\mu = 20.8$ and $\sigma = 5.8$. So Jose's z-score is
>
> $$z_{\text{Jose}} = \frac{27 - 20.8}{5.8} = 1.07$$
>
> standard deviations above the mean.
>
> Comparing their z-scores, we can see that Navreet's score is a little farther above the mean than Jose's and so we can conclude that Navreet got the better score.

Take a moment to connect z-scores back to the empirical rule. If a z-score is the number of standard deviations an observation is from the mean, we can rewrite the empirical rule as: For any (approximately) normally distributed variable,

1. Approximately 68% of all possible observations have z-scores between -1 and 1.
2. Approximately 95% of all possible observations have z-scores between -2 and 2.

3. Approximately 99.7% of all possible observations have z-scores between -3 and 3.

Note that, since nearly all observations fall within three standard deviations of the mean, this is another way to think about potential outliers: if an observation has a z-score less than -3 or greater than 3, it's a potential outlier.

5.3.2 Normal Distribution Probabilities

In order to make normal distributions easier to work with, we will standardize them. A **standard normal distribution** is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. If X is approximately normal, then the standardized variable Z will have a standard normal distribution.

Properties:

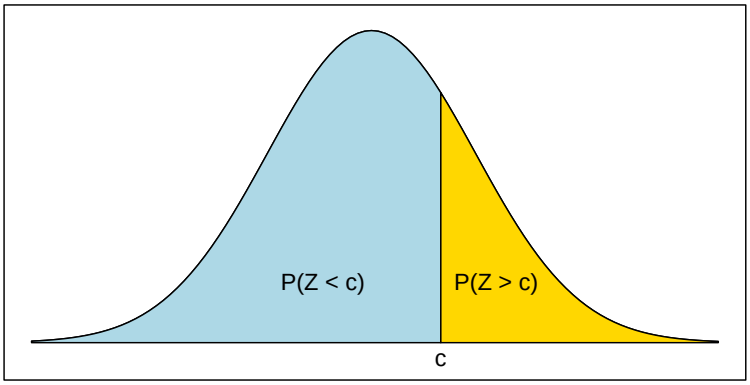
1. Total area under the curve is 1.
2. The curve extends infinitely in both directions, never touching the horizontal axis.
3. Symmetric about 0.
4. Almost all of the area under the curve is between -3 and 3.

Note: when we z-score a variable, we preserve the area under the curve properties! If X is Normal(μ, σ), then

$$P(X < c) = P\left(Z < \frac{c - \mu}{\sigma}\right) = P(Z < z).$$

We will think about area under the standard normal curve in terms of **cumulative probabilities** or probabilities of the form $P(Z < z)$.

We will use the fact that the total area under the curve is 1 to find probabilities like $P(Z > c)$:



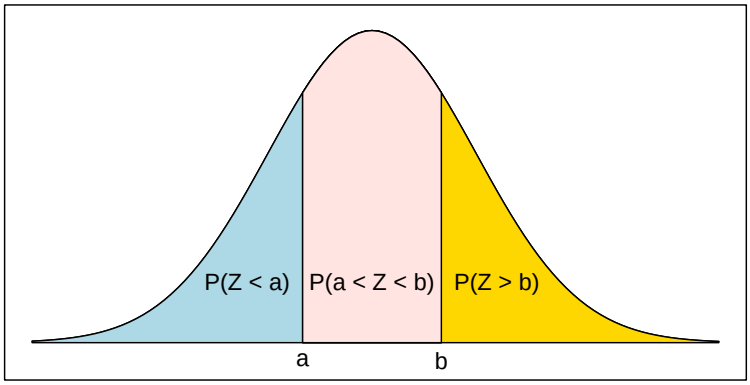
Using the graphic to help visualize, we can see that

$$1 = P(Z < c) + P(Z > c)$$

which we can then rewrite as

$$P(Z > c) = 1 - P(Z < c).$$

We can also use this concept to find $P(a < Z < b)$.



Total area = 1

Notice that

$$1 = P(Z < a) + P(a < Z < b) + P(Z > b),$$

which we can rewrite as

$$P(a < Z < b) = 1 - P(Z > b) - P(Z < a)$$

and since we just found that $P(Z > b) = 1 - P(Z < b)$, we can replace $1 - P(Z > b)$ with $P(Z < b)$, and get

$$P(a < Z < b) = P(Z < b) - P(Z < a).$$

Key Cumulative Probability Concepts

- $P(Z > c) = 1 - P(Z < c)$
- $P(a < Z < b) = P(Z < b) - P(Z < a)$

A final note, because the normal distribution is symmetric, $P(X < \mu) = P(X > \mu) = 0.5$. Notice this also implies that, when a distribution is symmetric (and unimodal), the mean and median are the same!

Now that we can get all of our probabilities written as *cumulative* probabilities, we're ready to use software to find the area under the curve!

Finding Area Under the Curve: Applets

One option for finding probabilities and z-scores associated with the normal curve is to use an online applet. The Rossman and Chance Normal Probability Calculator is my preferred applet. It's relatively straightforward to use, but

would be difficult to demonstrate in these course notes. Instead, we will demonstrate this applet in class. I recommend you bookmark any websites you use to find probabilities.

You can also find the area under a normal distribution using a Normal Distribution Table. These are outdated and not used anywhere but the statistics classroom. As a result, I do not teach them.

Using z-scores and area under the standard normal curve, we can now find probabilities for any normal distribution problem!

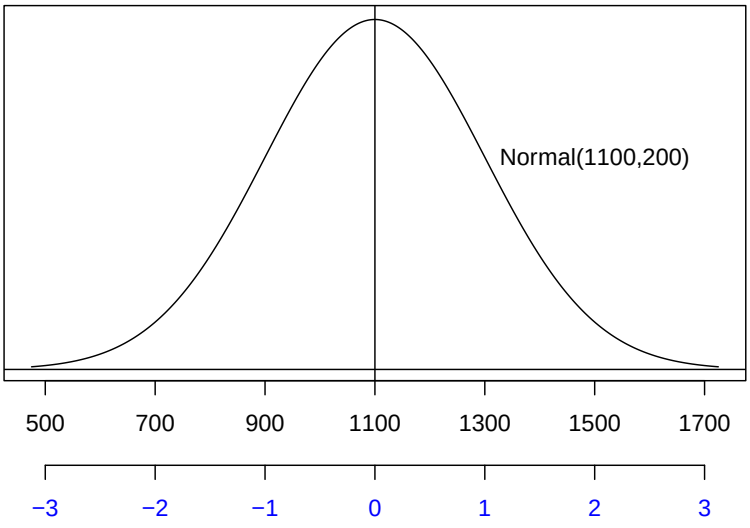
Determining Normal Distribution Probabilities

1. Sketch the normal curve for the variable.
2. Shade the region of interest and mark its delimiting x-value(s).
3. Find the z-score(s) for the value(s).
4. Use an applet (or the `pnorm` command in R) to find the associated area.

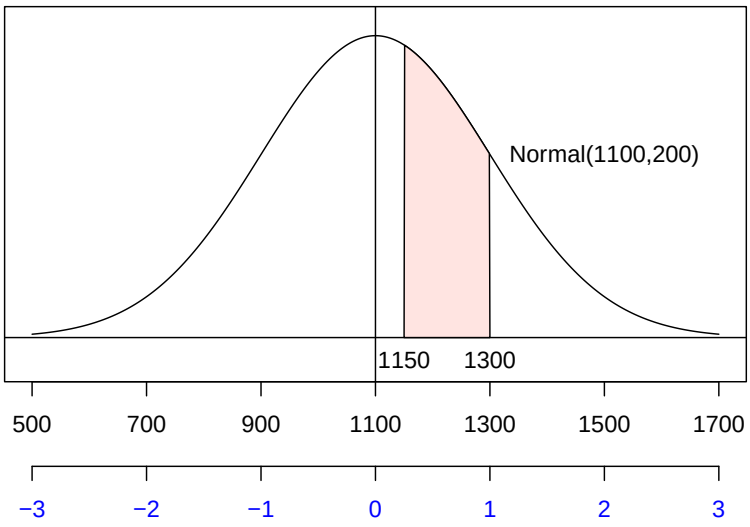
Example: Find the proportion of SAT-takers who score between 1150 and 1300. Assume that SAT scores are approximately normally distributed with mean $\mu = 1100$ and standard deviation $\sigma = 200$.

First, let's figure out what we want to calculate. Using area under the curve concepts, the proportion of test-takers who score *between* 1150 and 1300 will be $P(1150 < X < 1300)$.

1. Sketch. On the bottom axis, I am including μ, $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$,as well as the corresponding z-scores (in blue).



2. Shade and label:



3. Calculate z-scores:

$$x = 1150 \rightarrow z = \frac{1150 - 1100}{200} = 0.25$$

and

$$x = 1300 \rightarrow z = \frac{1300 - 1100}{200} = 1.$$

4. Use an applet to find $P(Z < 0.25) = 0.599$ and $P(Z < 1) = 0.841$ *or* use the `pnorm` command in R:

```
pnorm(0.25)
```

```
## [1] 0.5987063
```

```
pnorm(1)
```

```
## [1] 0.8413447
```

Note that

$$P(1150 < X < 1300) = P\left(\frac{1150 - 1100}{200} < Z < \frac{1300 - 1100}{200}\right) = P(0.25 < Z < 1)$$

and, using cumulative probability concepts,

$$P(0.25 < Z < 1) = P(Z < 1) - P(Z < 0.25).$$

We found $P(Z < 0.25) \approx 0.5987$ and $P(Z < 1) \approx 0.8413$, so

$$P(Z < 1) - P(Z < 0.25) \approx 0.8413 - 0.5987 = 0.2426.$$

That is, approximately 26.26% of test-takers score between 1150 and 1300 on the SAT.

5.3.3 Percentiles

We can also find the *observation* associated with a percentage/proportion.

The wth **percentile** p_w is the observation that is higher than w% of all observations

$$P(X < p_w) = \frac{w}{100}$$

Finding a Percentile

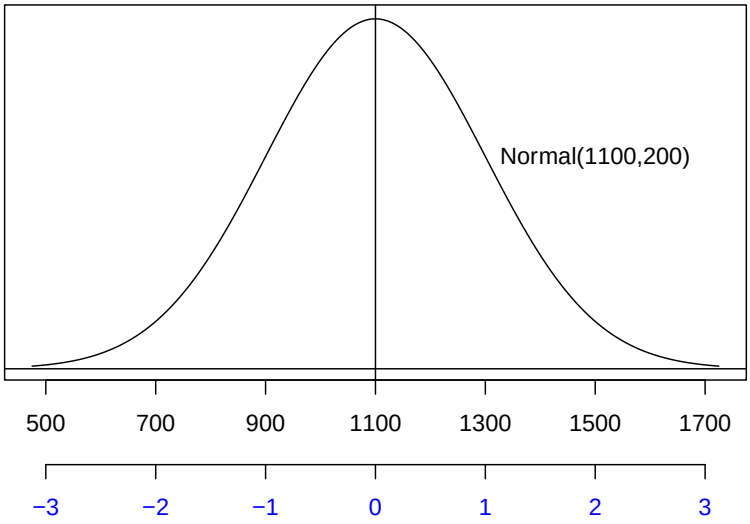
1. Sketch the normal curve for the variable.
2. Shade the region of interest and label the area.
3. Use the applet (or R - see below) to determine the z-score for the area.
4. Find the x-value using z, μ, and σ.

Note that if $z = \frac{x-\mu}{\sigma}$, then $x = \mu + z\sigma$.

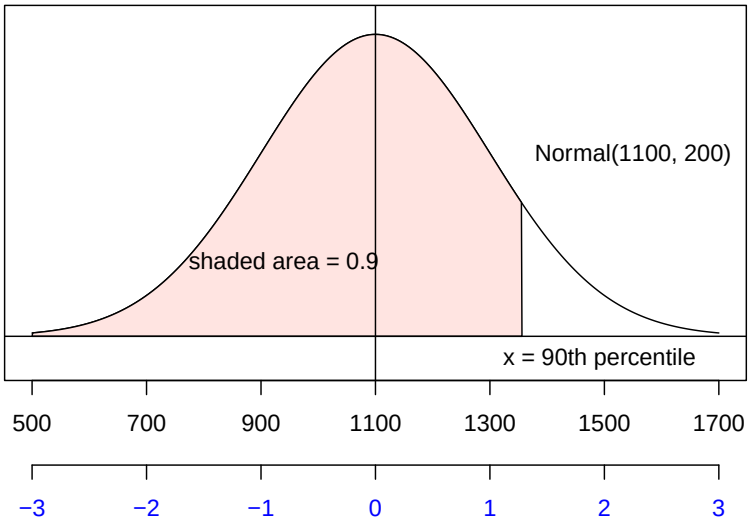
Example: Find the 90th percentile for SAT scores.

From the previous example, we know that SAT scores are approximately Normal($\mu = 1100$, $\sigma = 200$).

1. Sketch the normal curve.



2. Shade the region of interest and label the area.



3. Use the applet to determine the z-score for the area. This results in $z = 1.28$.

4. Find the x-value using $z \approx 1.28$, $\mu = 1100$, and $\sigma = 200$:

$$x = 1100 + 1.28(200) = 1356$$

so 90% of SAT test-takers score below 1356.

Section Exercises

Let X be the height of randomly selected adult men. Globally, the approximate mean and standard deviation of men's heights are $\mu = 68$ and $\sigma = 3$.

Use this information, along with the z-score formula, to fill in the table, below

x	62	64	_____	66	68	70	_____	72	74
z	_____	_____	**-1**	_____	_____	_____	**1**	_____	_____

Assume X is approximately normal. What percent of male heights fall between 65 and 71 inches?

2. Sketch the normal curve and shade the region of interest for the following probabilities
   a. $P(Z < 1.5)$
   b. $P(Z > -1)$
   c. $P(-0.75 < Z < 0.5)$
3. Use a computer to find the following probabilities
   a. $P(Z < 1.5)$
   b. $P(Z > -1)$
   c. $P(-0.75 < Z < 0.5)$
4. For what value of a will $P(-a < z < a) \approx 0.95$? Hint: remember the normal distribution is symmetric!
5. Let X be the height of a randomly selected adult man. Globally, $\mu = 68$ and $\sigma = 3$. Use this information to calculate the following probabilities. Hint: the table from (1) may be helpful.
   a. $P(X < 70)$
   b. $P(64 < X < 72)$
6. Suppose $\mu = 20$ and $\sigma = 4$. Draw, shade, and label the normal curve, and then determine the probability.
   a. $P(X < 23)$
   b. $P(X < 19)$
   c. $P(19 < X < 23)$
   d. $P(X < 15 \text{ OR } X > 22)$
7. Suppose $\mu = 15$ and $\sigma = 5$. Draw, shade, and label the normal curve, and then determine the percentile.
   a. 90th percentile
   b. 45th percentile
   c. 17th percentile

8. ACT scores are well-approximately by a normal distribution with mean 20.8 and standard deviation 5.8.
   a. Find the probability that a randomly selected test taker scores below a 19.
   b. Find the probability a randomly selected test taker score above a 23.
   c. Find the probability a randomly selected test taker scored between and 18 and a 22.
   d. Find the ACT score associated with the 90th percentile.
9. The average retirement age for an NFL player is 27.6 years old, with a standard deviation of 3.1.
   a. Draw a normal distribution for the age of retirement of NFL players. Mark μ, along with $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.
   b. Find the probability an NFL player retires before age 25.
   c. Tom Brady retired from the NFL in 2023 at age 45. Is it unusual for a player to stay in the NFL until such an age? Explain, and support your answer using math.
   d. Find the 75th percentile for NFL retirement age.
10. The notation z_α means the z-value such that the area of the *right* tail (the area to the right of z_α) is α. That is, $P(Z > z_\alpha) = \alpha$. For each z_α below, draw, shade, and label the normal curve, then determine the value of z_α.
   a. $z_{0.5}$

$$P(Z > z_{0.5}) = 0.5$$

   b. $z_{0.1}$
   c. $z_{0.05}$
   d. $z_{0.01}$

R Lab: Probabilities and Percentiles

Binomial Probabilities

When it comes to calculating binomial probabilities, hand calculations can be cumbersome. Fortunately, this is another thing we can do in R!

Approximately 66% of US adults take prescription medications. Find the probability that, in a sample of 100 adults, exactly 65 take prescription drugs.

We want to find $P(X = 65)$ where X has a binomial distribution with $n = 100$ and $p = 0.66$. (Take a moment on your own to make sure you can convince yourself that this satisfies the conditions for a binomial setting and that you understand how we got here from the prompt above.)

Instead of doing this by hand (the larger n is, the more difficult this tends to get!), we will use the `dbinom` command in R. The `dbinom` command takes in the following information:

- `x` the value x takes on in the expression $P(X = x)$
- `size` the value of n

- `prob` the probability p

```
dbinom(x = 65, size = 100, prob = 0.66)
```

```
## [1] 0.0815753
```

So without doing any hand calculations, I find that $P(X = 65) = 0.082$; the probability that exactly 65 of 100 randomly selected US adults take prescription medication is 0.082.

Suppose now we want to find $P(63 < X < 68)$. How can we manage that? We can figure out that this probability includes numbers between 63 and 68, but does not include 63 or 68. In fact, it is the numbers 64 through 67. SO we can break this up as

$$P(63 < X < 68) = P(X = 64) + P(X = 65) + P(X = 66) + P(X = 67)$$

In R, we can get a sequence of whole numbers using the format `a:b`. For example

```
64:67
```

```
## [1] 64 65 66 67
```

gives all whole numbers from 64 through 67.

I can then put this directly into the `dbinom` command!

```
dbinom(x = 64:67, size = 100, prob = 0.66)
```

```
## [1] 0.07587601 0.08157530 0.08397457 0.08272122
```

This produces each individual probability $P(X = 64)$, $P(X = 65)$, $P(X = 66)$, and $P(X = 67)$. To quickly add these up, I am going to use the `sum` command. Notice that I put the entire `dibnom` command *in the parentheses* of the `sum()`.

```
sum(
  dbinom(x = 64:67, size = 100, prob = 0.66)
)
```

```
## [1] 0.3241471
```

And so $P(63 < X < 68) = 0.324$; the probability that between 64 and 67 (inclusive) US adults in a sample of 100 take prescription medication is 0.324.

Normal Distribution Probabilities

Standard normal probabilities are found using the command `pnorm`. This command takes arguments

- `q`: the value of x.
- `mean`: the mean of the normal distribution. (If you leave this out, R will use $\mu = 0$.)

- `sd`: the standard deviation of the normal distribution. (If you leave this out, R will use $\sigma = 1$.)
- `lower.tail`: whether to find the lower tail probability.
  - When this equals `TRUE`, R will find $P(X < x)$.
  - When this equals `FALSE`, R will find $P(X > x)$.

Suppose X is a normal random variable with mean $\mu = 8$ and standard deviation $\sigma = 2$. To find $P(X > 4)$, I would write

```
pnorm(q=4, mean=8, sd=2, lower.tail=FALSE)
```

```
## [1] 0.9772499
```

So $P(X > 4) = 0.9772$

Because R will use the standard normal distribution if we leave out the `mean` and `sd` commands, it's even easier to find probabilities for Z, To find $P(Z < 1)$, I would type:

```
pnorm(q=1, lower.tail=TRUE)
```

```
## [1] 0.8413447
```

so $P(Z < 1) = 0.841$.

A quick note about R: R will print very large numbers and numbers close to 0 using *scientific notation*. However, R's scientific notation may not look the way you're used to! Check out the R output for $P(Z < -5)$:

```
pnorm(-5)
```

```
## [1] 2.866516e-07
```

When you see `e-07`, that means $\times 10^{-7}$... so $P(Z < -5) = 2.8665 \times 10^{-7} \approx 0.00000029$.

Normal Distribution Percentiles

Instead of using an applet, we can use the `qnorm` command in R to find the z-score corresponding to a percentile. In this case, we simply enter the percentile of interest *expressed as a proportion* in the `qnorm` command. That is, to find the z score for the 90th percentile, we would enter

```
qnorm(0.9)
```

```
## [1] 1.281552
```

which gives the same result as the applet in the example above. Then, we can use R as a calculator to find the value of x (recall $\mu = 1100$ and $\sigma = 200$)

```
1100 + 1.281552*200
```

```
## [1] 1356.31
```

This gives us the same result as before, that 90% of SAT test-takers score below 1356.

Chapter 6

Introduction to Confidence Intervals

This module will bridge the gap between our discussion on the normal distribution and our first forays into statistical inference. As it turns out, much of the statistical inference we will use relies on the normal distribution and the t-distribution, which we will introduce in this module. We begin our study of statistical inference by learning about confidence intervals.

Module Learning Objectives/Outcomes

1. Find the distribution of a sample mean.
2. Estimate probabilities for a sample mean.
3. Use the standard normal and t-distributions to find critical values.
4. Calculate and interpret confidence intervals for a population mean.

R Objectives

1. Find z and t critical values.
2. Generate complete confidence intervals for a population mean.

This module's outcomes correspond to course outcome (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

6.1 Sampling Distributions

6.1.1 Sampling Error

We want to use a sample to learn something about a population, but no sample is perfect! **Sampling error** is the error resulting from using a sample to estimate

a population characteristic.

If we use a sample mean $\bar{x}$ to estimate μ, chances are that $\bar{x} \neq \mu$ (they might be close but… they might not be!). We will consider

- How close *is* $\bar{x}$ to μ?
- What if we took many samples and calculated $\bar{x}$ many times?
  - How would that relate to μ?
  - What would be the distribution of these values?

The distribution of a statistic (across all possible samples of size n) is called the **sampling distribution**. We will focus primarily on the distribution of the sample mean.

For a variable x and given a sample size n, the distribution of $\bar{x}$ is called the **sampling distribution of the sample mean** or the **distribution of** $\bar{x}$.

> **Example**: Suppose our population is the five starting players on the Sacramento Kings during their game on April 2, 2024. We are interested in their heights (measures in inches). The full population data is

Player	De'Aaron Fox	Keon Ellis	Harrison Barnes	Keegan Murray	Domantas Sabonis
Height	73	73	78	78	80

> The population mean is $\mu = 76.4$. Consider all possible samples of size $n = 2$:

Sample	$\bar{x}$
Fox, Ellis	73
Fox, Barnes	75.5
Fox, Murray	75.5
Fox, Sabonis	76.5
Ellis, Barnes	75.5
Ellis, Murray	75.5
Ellis, Sabonis	76.5
Barnes, Murray	78
Barnes, Sabonis	79
Murray, Sabonis	79

> There are 10 possible samples of size 2. Of these samples, *none* have means exactly equal to μ.

In general, the larger the sample size, the smaller the sampling error tends to be in estimating μ using $\bar{x}$.

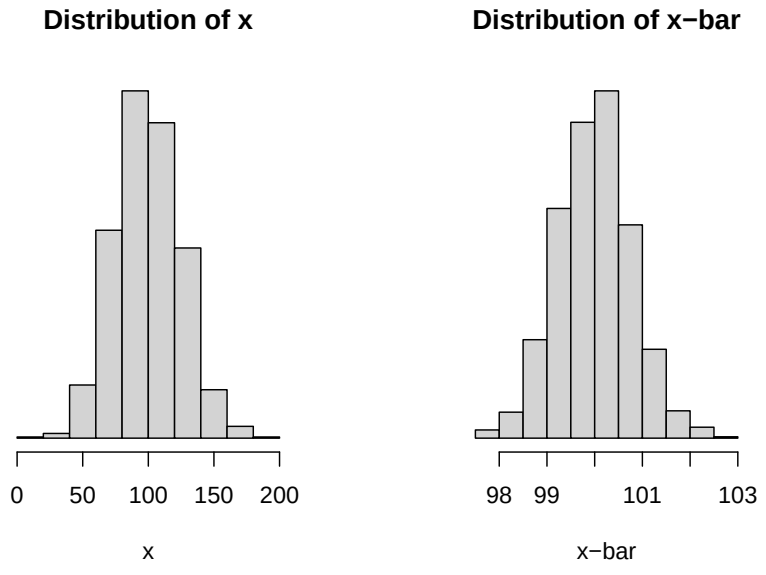
In practice, we have one sample and μ is unknown. We also have limited resources to collect data, so it may not be feasible to collect a very large sample.

The mean of the distribution of $\bar{x}$ is $\mu_{\bar{X}} = \mu$ and the standard deviation is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. We refer to the standard deviation of a sampling distribution as **standard error**. (Note that this standard error formula is built for very large populations, so it will not work well for our basketball players. This is okay! We usually work with populations so large that we treat them as "infinite".)

> **Example**: The mean house size in the United States is 2164 ft^2 with a standard deviation of 568 square feet. For samples of 25 homes, determine the mean and standard error of $\bar{x}$.
>
> Using our formulae:
>
> $$\mu_{\bar{X}} = \mu = 2164$$
>
> and
>
> $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{568}{\sqrt{25}} = 113.6.$$

6.1.2 The Central Limit Theorem

Consider the setting where X is Normal(μ, σ). The plots below show (A) a random sample of 1000 from a Normal(100, 25) distribution and (B) the approximate sampling distribution of $\bar{X}$ when X is Normal(100, 25).



Notice how the x-axis changes from one plot to the next.

In fact, if X is Normal(μ, σ), then $\bar{X}$ is Normal($\mu_{\bar{X}} = \mu$, $\sigma_{\bar{X}} = \sigma/\sqrt{n}$).

Surprisingly, we see a similar result for $\bar{X}$ even when X is not normally distributed!

Central Limit Theorem

For relatively large sample sizes, the random variable $\bar{X}$ is approximately normally distributed *regardless of the distribution of* X:

$$\bar{X} \text{ is Normal}(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n}).$$

Notes

- This approximation improves with increasing sample size.
- In general, "relatively large" means sample sizes $n \geq 30$.

Section Exercises

1. The mean house size in Hong Kong is 484 ft^2 with a standard deviation of 163 ft^2 (this mean is accurate as of 2024, but the standard deviation is a guesstimate for the sake of the problem).
   a. Determine $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ for samples of size 22.
   b. Determine $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ for samples of size 100.
   c. For which of the above ($n = 22$ or $n = 100$) can we assume the sampling distribution is normal, *without knowing anything else about the data*?
   d. For a random sample of 100 houses, what is the probability the mean size is less than 400 ft^2?
2. Suppose we have some population with $\sigma = 5$. Determine the proportion of all samples of size 50 that will have means within 0.25 of the population mean.
   a. Draw and label a normal curve with μ, $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.
   b. What is the sample size?
   c. Determine $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$.
   d. Write an expression to represent the values 0.125 above and below the mean, $\mu_{\bar{X}}$. (Since we don't know the value of $\mu_{\bar{X}}$, you should use $\mu_{\bar{X}}$ directly in your expression.)
   e. Determine the z-score for each of the values you found in part (d).
   f. Use your result from part (e) to determine the proportion of all samples of size 50 that will have means within 0.25 of the population mean.

6.2 Developing Confidence Intervals

Recall: A **point estimate** is a single-value estimate of a population parameter. We say that a statistic is an **unbiased estimator** if the mean of its distribution is equal to the population parameter. Otherwise, it is a **biased estimator**.

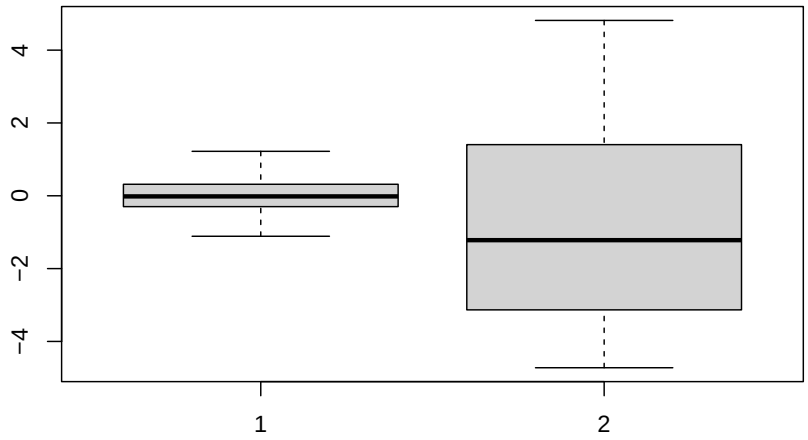
Comment: Remember how our formula for sample variance, the "mean squared deviance" divides by $n-1$ instead of n? We do this so that s is an *unbiased* estimate of σ.

Ideally, we want estimates that are unbiased with small standard error. For example, a sample mean (unbiased) with a large sample size (results in smaller standard error).

Point estimates are useful, but they only give us so much information. The variability of an estimate is also important!

> **Example** Think about estimating what tomorrow's weather will be like. If it's May in Sacramento, the average high temperature is 82 degrees Fahrenheit, but it's not uncommon to have highs anywhere from 75 to 90! Since the highs are so *variable*, it's hard to be confident using 82 to predict tomorrow's weather.
>
> On the flip side, think about July in Phoenix. The average high is 106 degrees Fahrenheit. In Phoenix, it's uncommon to have a July day with a high below 100. Since the highs are *not variable*, you could feel pretty confident using 106 to predict tomorrow's weather.

Take a look at these two boxplots:



Both samples are size $n = 100$ and have $\bar{x} = 0$, which would be our point estimate for μ... but Variable 1 has a standard deviation of $\sigma = 0.5$ and Variable 2 has standard deviation $\sigma = 5$. As a result, we can be more confident in our estimate of the population mean for Variable 1 than for Variable 2.

We want to formalize this idea of confidence in our estimates. A **confidence interval** is an interval of numbers based on the point estimate of the parameter. Say we want to be 95% confident about a statement. In Statistics, this means that we have arrived at our statement using a method that will give us a correct statement 95% of the time.

Our best point estimate for μ (based on a random sample) is $\bar{x}$, so that value will make up the center (or midpoint) of the interval. To create an interval around $\bar{x}$, we will construct what is called the **margin of error**. We will use the variability of the data along with some normal distribution properties. This will look like

$$z \times \frac{\sigma}{\sqrt{n}}$$

The value z will come from the normal distribution and will be based on how confident we want to be, e.g., 95% confident.

Putting everything together, the 95% confidence interval is

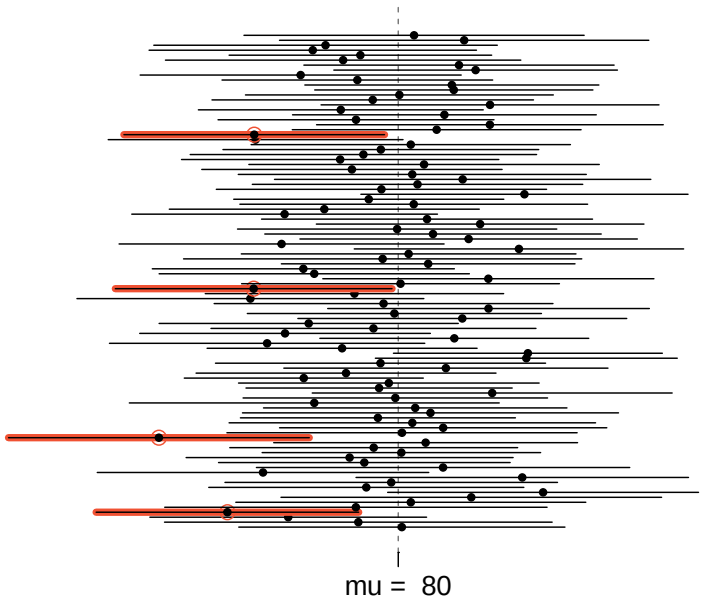
$$\left(\bar{x} - z_* \frac{\sigma}{\sqrt{n}}, \bar{x} + z_* \frac{\sigma}{\sqrt{n}}\right)$$

where $z_* = 1.96$. The value 1.96 is chosen because $P(-1.96 < Z < 1.96) = 0.95$ (this is what makes it a 95% confidence interval!).

Note: A more detailed mathematical explanation of how we get this interval is available in Appendix C.

6.2.1 Interpreting a Confidence Interval

To interpret a confidence interval, we need to think back to our definition of probability as "the proportion of times an event would occur if the experiment were run infinitely many times". In the confidence interval case, if an experiment is run infinitely many times, the true value of μ will be contained in 95% of the intervals.



The graphic above shows 95% confidence intervals for 100 samples of size $n = 60$ drawn from a population with mean $\mu = 80$ and standard deviation $\sigma = 25$. Each sample's confidence interval is represented by a horizontal line. The dot in the middle of each is the sample mean. When a confidence interval does *not* capture the population mean μ, the line is printed in red. Based on this concept of repeated sampling, we would expect about 95% of these intervals to capture μ. In fact, 96 of the 100 intervals capture μ.

Finally, when you interpret a confidence interval, it is important to do so in the context of the problem.

> **Example** Suppose I took a random sample of 50 Sac State students and asked about their SAT scores and found a mean score of 1112. Prior experience with SAT scores in the CSU system suggests that SAT scores are well-approximated by a normal distribution with standard deviation known to be 50.
>
> 1. Find and interpret a 95% confidence interval for Sac State SAT scores.
> 2. What is the width of your interval? If you want a narrower interval, what could you do?
>
> *Solution* Here, $\bar{x} = 1112$, $\sigma = 50$, and $n = 50$.
>
> The interval is
>
> $$\bar{x} \pm z_{*\alpha/2*} \frac{\sigma}{\sqrt{n}} = 1112 \pm 1.96 \times \frac{50}{\sqrt{50}} = 1112 \pm 13.86 = (1098.1, 1125.9).$$

> Interpretation: We can be 95% confident that the true mean SAT score for Sac State students is between 1098.1 and 1125.9.

Notice that I kept the interpretation simple! "We can be [x]% confident that the true mean [fill in from problem statement] is between [lower value] and [upper value]." As appropriate, we should also be sure to include any units. A simple interpretation is fine - just be sure you are *also* able to explain what it means to be 95% confident (using the concept of repeated sampling).

Common mistakes:

- It is NOT accurate to say that "the probability that μ is in the confidence interval is 0.95". The parameter μ is some fixed quantity and it's either in the interval or it isn't.
- We are NOT "95% confident that $\bar{x}$ is in the interval". The value $\bar{x}$ is some known quantity and it's always in the interval.

Section Exercises

1. Interpret each confidence interval in the context of the setting.
   a. A 95% confidence interval for the mean ACT score of Sac State students was (21.2, 25.1).
   b. A 90% confidence interval for the mean number of burritos eaten by Californian adults each week was (0.5, 2.1).
   c. A 98% confidence interval for the mean height (in inches) of corgis was (10.1, 11.7).
2. Describe in your own words what is meant by "95% confident".

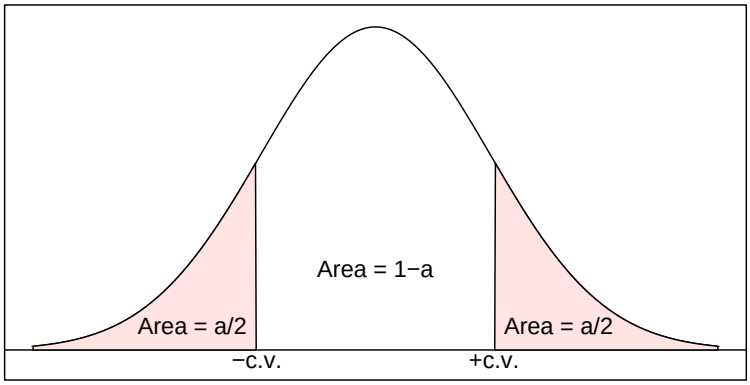
6.3 Other Levels of Confidence

While the 95% confidence interval is common in research, there's nothing inherently special about it. You could calculate a 90%, a 99%, or - if you're feeling spicy - something like a 43.8% confidence interval. These numbers are called the **confidence level** and they represent the proportion of times that the parameter will fall in the interval (if we took many samples).

The 100(1-α)% confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z-score associated with the $[1-(\alpha/2)]$th percentile of the standard normal distribution. The value $z_{\alpha/2}$ is called the **critical value** ("c.v." on the plot, below).



Using the those normal distribution properties, we can also say that $P(Z > z_{\alpha/2}) = \alpha/2$. Depending on how you've been using software to find normal distribution probabilities, you may find you prefer this approach to the percentile approach.

Common Critical Values

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
98%	0.02	2.326
99%	0.01	2.575

A good exercise: make sure you are comfortable using software to find critical values by finding the common critical values given in the table above.

> **Example** In the previous section, we worked with a random sample of 50 Sac State students with mean SAT score 1112. Prior experience with SAT scores in the CSU system suggests that SAT scores are well-approximated by a normal distribution with standard deviation known to be 50.
>
> a. Find and interpret a 98% confidence interval.
> b. Find and interpret a 90% confidence interval.
> c. Comment on how the intervals change as you change the confidence level.

Solution Again, $\bar{x} = 1112$, $\sigma = 50$, and $n = 50$.

a. For a 98% interval, $\alpha = 0.02$ and $z_{0.02/2} = 2.326$. So the interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1112 \pm 2.326 \times \frac{50}{\sqrt{50}} = 1112 \pm 16.4 = (1095.6, 1128.4)$$

We can be 98% confident that the true mean SAT score for Sac State students is between 1095.6 and 1128.4.

b. For a 90% interval, $\alpha = 0.1$ and $z_{0.1/2} = 1.645$. So the interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1112 \pm 1.645 \times \frac{50}{\sqrt{50}} = 1112 \pm 11.6 = (1100.4, 1123.6)$$

We can be 90% confident that the true mean SAT score for Sac State students is between 1100.4 and 1123.6

c. As the level of confidence *increases*, the interval width *decreases*.

6.3.1 Breaking Down a Confidence Interval

Consider

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

The key values are

- $\bar{x}$, the sample mean
- σ, the population standard deviation
- n, the sample size
- $z_{\alpha/2}$, the critical value

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

The value of interest is μ, the (unknown) population mean; the confidence interval gives us a reasonable range of values for μ.

In addition, the formula includes

- The standard error, $\frac{\sigma}{\sqrt{n}}$
- The margin of error, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Section Exercises

1. Find α, then use a computer to find the corresponding critical value $z_{\alpha/2}$ for
   a. A 92% level of confidence.
   b. A 37% level of confidence.
   c. An 85% level of confidence.

2. We saw in the last example that the level of confidence impacts the width of the confidence interval. When we think about collecting data and calculating a confidence interval, what else can we do to change the width of the interval?

6.4 Confidence Level, Precision, and Sample Size

If we can be 99% confident (or even higher), why do we tend to "settle" for 95%?? Take a look at the common critical values (above) and the confidence interval formula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

What will higher levels of confidence do to this interval? Think back to the intuitive interval width explanation with the weather. Mathematically, the same thing will happen: the interval will get wider! And remember, a narrow interval is a more informative interval. There is a trade off here between interval width and confidence. In general, the scientific community has settled on 95% as a compromise between the two, but different fields may use different levels of confidence.

There is one other thing we can control in the confidence interval: the sample size n. One strategy is to specify the confidence level and the maximum acceptable interval width and use these to determine sample size. We know that

$$\text{interval width} = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

When determining sample size, $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the *maximum* acceptable interval width, so we will consider

$$\text{interval width} \geq 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

since we would still be happy if this value turned out to be smaller! Letting interval width equal w, we then solve for n:

$$\begin{aligned} w &\geq 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ w\sqrt{n} &\geq 2z_{\alpha/2}\sigma \\ \sqrt{n} &\geq 2z_{\alpha/2}\frac{\sigma}{w} \\ n &\geq \left(2z_{\alpha/2}\frac{\sigma}{w}\right)^2 \end{aligned}$$

Alternately, we may specify a maximum margin of error m instead:

$$\begin{aligned} m &\geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ m\sqrt{n} &\geq z_{\alpha/2}\sigma \\ \sqrt{n} &\geq z_{\alpha/2}\frac{\sigma}{m} \\ n &\geq \left(z_{\alpha/2}\frac{\sigma}{m}\right)^2 \end{aligned}$$

Once we've done this calculation, we need a whole number for n. Since $n \geq$ something, we will *always round up.*

> **Example** Suppose we want a 95% confidence interval for the mean of a normally distributed population with standard deviation $\sigma = 10$. It is important for our margin of error to be no more than 2. What sample size do we need?
>
> Using the formula for sample size with a desired margin of error, I can plug in $z_{0.05/2} = 1.96$, $m = 2$ and $\sigma = 10$:
>
> $$n = \left(1.96 \times \frac{10}{2}\right)^2 = 96.04$$
>
> So (rounding up!) I need a sample size of *at least 97.*

A few comments:

- As desired width/margin of error decreases, n will increase.
- As σ increases, n will also increase. (More population variability will necessitate a larger sample size.)
- As confidence level increases, n will also increase.

Section Exercises

1. Find the sample size necessary in each of the following settings.
   a. Suppose $\sigma = 4$. We wish to construct a 90% confidence interval with a maximum width of 10.
   b. Suppose $\sigma = 12$. We wish to construct a 98% confidence interval with a maximum width of 15.
   c. Suppose $\sigma = 2$. We wish to construct a 92% confidence interval with a maximum margin of error of 5.
   d. Suppose $\sigma = 10$. We wish to construct a 99% confidence interval with a maximum margin of error of 3.
2. Prior experience with SAT scores in the CSU system suggests that SAT scores are well-approximated by a normal distribution with standard deviation known to be 50. Find the sample size required for a 98% confidence interval with maximum margin of error 10.

3. In practice, we want contradictory things: small margin of error, high level of confidence, and small sample size (often due to time/monetary constraints).
   a. Come up with a scenario where it will be especially important to have a small margin of error and a high level of confidence, and therefore worth spending a lot of resources on gathering a large sample size.
   b. Come up with a second research scenario where it may be impractical to gather a large sample size. In your scenario, is it more important to prioritize interval width or confidence level? Explain your thought process.

6.5 Confidence Intervals for a Mean

In practice, the value of σ is almost never known… but we know that we can estimate σ using s. Can we plug in s for σ? Sometimes!

Remember the Central Limit Theorem from earlier in this this module? For samples of size $n \geq 30$, $\bar{X}$ will be approximately normal even if X isn't. In this case, we can plug in s for σ:

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

That setting is pretty straightforward! Now we need to consider the setting where $n < 30$, which will require a bit of additional work.

> **Example**: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly selected sample of adults. One of the questions on the survey is "After an average workday, about how many hours do you have to relax or pursue activities that you enjoy?". The average time spent relaxing was 3.68 hours, with a standard deviation of 2.6 hours.
>
> Find and interpret a 95% confidence interval for the average time spent relaxing after work.
>
> *Solution*: Clearly, $n = 1154 \geq 30$, so we can use the normal distribution for our critical value. We want a 95% interval, so we will use critical value $z_{0.05/2} = z_{0.025} = 1.96$.
>
> Further, we can see in the prompt that the sample resulted in sample mean $\bar{x} = 3.68$ and standard deviation $s = 2.6$. So the 95% interval is
>
> $$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 3.68 \pm 1.96\frac{2.6}{\sqrt{1154}} = 3.68 \pm 0.15$$
>
> which results in the interval $(3.53, 3.83)$.

Interpretation: we can be 95% confident that the true mean number of hours spent relaxing after work is between 3.53 and 3.83.

6.5.1 The T-Distribution

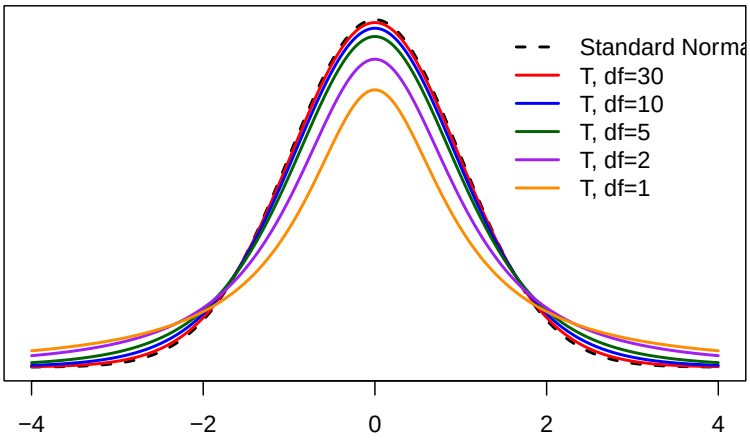
Enter: the t-distribution. If

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution (for X normal or $n \geq 30$), the slightly modified

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has what we call the **t-distribution** with $n-1$ **degrees of freedom** (even when $n < 30$!). The only thing we need to know about degrees of freedom is that $\text{df} = n - 1$ is the t-distribution's only parameter.



The t-distribution is symmetric and always centered at 0. When $n \geq 30$, the t-distribution is approximately equivalent to the standard normal distribution. For smaller sample sizes, the t-distribution has more area in the tails (and therefore less area in the center of the distribution). Therefore, we can always use t confidence intervals (even if n is large) because the critical values are essentially equivalent between the t and standard normal distributions for large values of n. This is usually what happens in practice.

In practice, we plug in s for σ and almost always use a t critical value (instead of a z critical value): $t_{\text{df},\,\alpha/2}$. The t critical value is the $[1-\alpha/2]$th percentile of

the t-distribution with $n-1$ degrees of freedom. The resulting 95% confidence interval is

$$\bar{x} \pm t_{\text{df},\,\alpha/2}\frac{s}{\sqrt{n}}.$$

You may use the applet, Rossman and Chance t Probability Calculator to find t critical values. For this applet, enter the degrees of freedom $n-1$ next to "df". Then check the top box under "t-value probability" and make sure the inequality is clicked to ">" . Enter the value of $\alpha/2$ for the probability. Click anywhere else on the page and the applet will automatically fill in the box under "t-value". This is your t critical value.

Example: The following table gives a few t critical values. You should make sure you can use the degrees of freedom and α to replicate those t critical values using a computer.

n	α	$t_{\text{df},\,\alpha/2}$
14	0.05	2.160
17	0.1	1.746
24	0.02	2.500

Example: The weights of 6 randomly selected NFL linebackers were $243, 238, 229, 253, 248, 225$. Find a 90% confidence interval for the average weight of NFL linebackers.

Solution: Right away, we can see that $n = 6 < 30$ and so we will need to use a confidence interval with a t critical value. Next, we calculate the mean and standard deviation from the data: $\bar{x} = 239.3$ and $s = 10.9$.

The t-critical value for a 90% interval has $\alpha = 0.1$ and $\text{df} = 6-1 = 5$. Using a computer, $t_{\text{df},\,\alpha/2} = t_{5,\,0.05} = 2.015$. Then

$$\bar{x} \pm t_{\text{df},\,\alpha/2}\frac{s}{\sqrt{n}} = 239.3 \pm 2.015 \times \frac{10.9}{\sqrt{6}} = 239.3 \pm 8.97$$

which yields the interval $(230.3, 248.3)$.

Interpretation: we can be 90% confident that the average weight of NFL linebackers is between 230.3 and 248.3 pounds.

Section Exercises

1. Find the t critical values for each combination of sample size and confidence level.
   a. 95%, $n = 19$
   b. 90%, $n = 27$
   c. 80%, $n = 7$

2. In 1-2 sentences, explain the difference between the Z (standard normal) and the t-distribution. What properties do these distributions have in common? How do they differ?
3. Refer to the example above on NFL linebackers. Find and interpret
   a. A 95% confidence interval for mean weight of NFL linebackers.
   b. A 98% confidence interval for mean weight of NFL linebackers.
   c. A 99% confidence interval for mean weight of NFL linebackers.
4. Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7.
   a. Should we use a z or a t critical value in this setting? Explain.
   b. Find and interpret a 95% confidence interval for the mean number of eggs in a green sea turtle nest.
   c. Find and interpret a 98% confidence interval for the mean number of eggs in a green sea turtle nest.
   d. Suppose $\mu = 105$. Did the confidence intervals do a good job of estimating μ? Explain.
5. Some Sac State students wanted to know how: long is an average commute to campus? A random sample of 32 students resulted in a mean of 31 minutes with standard deviation 18 minutes.
   a. Should we use a z or a t critical value in this setting? Explain.
   b. Find and interpret a 95% confidence interval for mean commute time.
   c. Find and interpret a 90% confidence interval for mean commute time.
6. A city council member suspects that people speed through a certain intersection. They want to know the average speed at which drivers enter the intersection. They set up a speed camera and randomly sample 250 vehicles at different times of day, which results in a mean speed of 52 mph with standard deviation 8 mph.
   a. Find and interpret a 95% confidence interval for the mean speed at which drivers take the intersection.
   b. Suppose the posted speed limit is 45mph. Does the interval suggest that people speed? Explain your thought process.

R Lab: Confidence Intervals

Finding Z Critical Values

We can find critical values in R using the same command we used to find percentiles: `qnorm`. Recall that $z_{\alpha/2}$ is the z-score associated with the $[1-(\alpha/2)]$th percentile of the standard normal distribution. So for a $(1-\alpha)100\%$ confidence interval, we need to find the value of $1-\alpha/2$ to input into the `qnorm` command.

For example, to find $z_{\alpha/2}$ for a 93% confidence interval, we would use

$$(1-\alpha)100\% = 93\%$$

to solve for α and get $\alpha = 0.07$. Then we need the $[1 - (\alpha/2)]$th percentile:

$$1 - \frac{\alpha}{2} = 1 - \frac{0.07}{2} = 0.965$$

Finally, we enter this value into the `qnorm` command:

```
qnorm(0.965)
```

```
## [1] 1.811911
```

and the critical value for a 93% confidence interval is $z_{\alpha/2} = 1.812$.

Finding T Critical Values

To find a t critical value, we will again use R, now with the command `qt`. (Notice that this is similar to the command for the standard normal distribution, but instead of "norm" for normal it has "t" for the t-distribution.) The `qt` command takes in the following:

- `p`, the probability
- `df`, the degrees of freedom

For example, for a 98% interval with a sample size of 15,

$$100(1 - \alpha) = 98 \implies \alpha = 0.02$$

Then $1 - \alpha/2 = 0.99$ and $\text{df} = 15 - 1 = 14$.

```
qt(p = 0.99, df = 14)
```

```
## [1] 2.624494
```

which gives the t critical value $t_{14,\alpha/2} = 2.625$.

Confidence Intevals for a Mean

To generate complete confidence intervals for a mean in R, we will use our confidence interval formula,

$$\bar{x} \pm (\text{critical value}) \times \frac{s}{\sqrt{n}}$$

where the critical value is either $z_{\alpha/2}$ or $t_{\text{df},\alpha/2}$.

We will continue to use the `Loblolly` pine tree data.

```
attach(Loblolly)
```

In this case, the variable of interest is `x = height` and the desired confidence level is `conf.level = 0.95`. We can check how many observations we have (the sample size) by using the command `length()` and putting the variable name in the parentheses:

```
length(height)
```

```
## [1] 84
```

Then we can find the sample mean and sample standard deviation of the `height` variable:

```
## [1] 32.3644
```

```
## [1] 20.6736
```

Since $n = 84 > 30$, we will use the standard normal distribution. Let's find the critical value for a 95% confidence interval. Here,

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975$$

so the R command will be

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Finally, we can put this all together. As when you use your calculator, pay close attention to order of operations in R. If in doubt, use parentheses!

```
32.36 - 1.96*(20.67/sqrt(84))
```

```
## [1] 27.93965
```

```
32.36 + 1.96*(20.67/sqrt(84))
```

```
## [1] 36.78035
```

So a 95% confidence interval for `height` is $(27.94, 36.78)$. We can say that we are 95% confident that the true mean height of the loblolly pines is between 27.94 and 36.78 feet (assuming the researchers took a random sample of trees).

Alternately, if you're starting to feel comfortable with R, we can put all of this together into one:

```
mean(height) - qnorm(0.975)*(sd(height)/sqrt(length(height)))
```

```
## [1] 27.94336
```

```
mean(height) + qnorm(0.975)*(sd(height)/sqrt(length(height)))
```

```
## [1] 36.78545
```

The numbers are slightly different only because R does not round until the very end.

```
detach(Loblolly)
```

Chapter 7

Introduction to Hypothesis Testing

In this module, we will continue our discussion on statistical inference with a discussion on hypothesis testing. In hypothesis testing, we take a more active approach to our data by asking questions about population parameters and developing a framework to answer those questions. We will root this discussion in confidence intervals before learning about several other approaches to hypothesis testing.

Module Learning Outcomes/Objectives

1. Test one sample means using
   a. confidence intervals.
   b. the critical value approach.
   c. the p-value approach.

R Objectives

1. Generate hypothesis tests for a mean.
2. Interpret R output for tests of a mean.

This module's outcomes correspond to course outcomes (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

7.1 Logic of Hypothesis Testing

This section is framed in terms of questions about a population mean μ, but the same logic applies to p (and other population parameters).

One of our goals with statistical inference is to make decisions or judgements about the value of a parameter. A confidence interval is a good starting point, but we might also want to ask questions like

- Do cans of soda actually contain 12 oz?
- Is Medicine A better than Medicine B?

A **hypothesis** is a statement that something is true. A hypothesis test involves two (competing) hypotheses:

1. The **null hypothesis**, denoted H_0, is the hypothesis to be tested. This is the "default" assumption.
2. The **alternative hypothesis**, denoted H_A is the alternative to the null.

Note that the subscript 0 is "nought" (pronounced "not"). A **hypothesis test** helps us decide whether the null hypothesis should be rejected in favor of the alternative.

> **Example**: Cans of soda are labeled with "12 FL OZ". Is this accurate?
>
> The default, or uninteresting, assumption is that cans of soda contain 12 oz.
>
> - H_0: the mean volume of soda in a can is 12 oz.
> - H_A: the mean volume of soda in a can is NOT 12 oz.

We can write these hypotheses in words (as above) or in statistical notation. The null specifies a single value of μ

- H_0: $\mu = \mu_0$

We call μ_0 the **null value**. When we run a hypothesis test, μ_0 will be replaced by some number. For the soda can example, the null value is 12. We would write $H_0 : \mu = 12$.

The alternative specifies a *range* of possible values for μ:

- H_A: $\mu \neq \mu_0$. "The true mean is different from the null value."

The Logic of Hypothesis Testing

Take a random sample from the population. If the data area consistent with the null hypothesis, do not reject the null hypothesis. If the data are inconsistent with the null hypothesis *and* supportive of the alternative hypothesis, reject the null in favor of the alternative.

> **Example**: One way to think about the logic of hypothesis testing is by comparing it to the U.S. court system. In a jury trial, jurors are told to assume the defendant is "innocent until proven guilty". Innocence is the default assumption, so
>
> - H_0: the defendant is innocent.
> - H_A: the defendant is guilty.

> Like in hypothesis testing, it is not the jury's job to decide if the defendant is innocent. That should be their default assumption. They are only there to decide if the defendant is guilty or if there is not enough evidence to override that default assumption. The *burden of proof* lies on the alternative hypothesis.

Notice the careful language in the logic of hypothesis testing: we either reject, or fail to reject, the null hypothesis. We never "accept" a null hypothesis.

7.1.1 Decision Errors

- A **Type I Error** is rejecting the null when it is true. (Null is true, but we conclude null is false.)
- A **Type II Error** is not rejecting the null when it is false. (Null is false, but we do not conclude it is false.)

H_0 is

True

False

Decision

Do not reject H_0

Correct decision

Type II Error

Reject H_0

Type I Error

Correct decision

> **Example**: In our jury trial,
>
> - H_0: the defendant is innocent.
> - H_A: the defendant is guilty.
>
> A Type I error is concluding guilt when the defendant is innocent.
> A Type II error is failing to convict when the person is guilty.

How likely are we to make errors? Well, $P(\text{Type I Error}) = \alpha$, the **significance level**. (Yes, this is the same α we saw in confidence intervals!) For Type II error, $P(\text{Type II Error}) = \beta$. This is related to the sample size calculation from the previous module, but is otherwise something we don't have time to cover.

We would like both α and β to be small but, like many other things in statistics, there's a trade off! For a fixed sample size,

- If we decrease α, then β will increase.
- If we increase α, then β will decrease.

In practice, we set α (as we did in confidence intervals). We can improve β by increasing sample size. Since resources are finite (we can't get enormous sample sizes all the time), we will need to consider the consequences of each type of error.

> **Example** We could think about assessing consequences through the jury trial example. Consider two possible charges:
>
> 1. Defendant is accused of stealing a loaf of bread. If found guilty, they may face some jail time and will have a criminal record.
> 2. Defendant is accused of murder. If found guilty, they will have a felony and may spend decades in prison.
>
> Since these are moral questions, I will let you consider the consequences of each type of error. However, keep in mind that we do make scientific decisions that have lasting impacts on people's lives.

Hypothesis Test Conclusions

- If the null hypothesis is rejected, we say the result is **statistically significant**. We can interpret this result with:
  - At the α level of significance, the data provide sufficient evidence to support the alternative hypothesis.
- If the null hypothesis is *not* rejected, we say the result is **not statistically significant**. We can interpret this result with:
  - At the α level of significance, the data do *not* provide sufficient evidence to support the alternative hypothesis.

Notice that these conclusions are framed in terms of the alternative hypothesis, which is either supported or not supported. We will *never* conclude the null hypothesis. Finally, when we write these types of conclusions, we will write them in the context of the problem.

Section Exercises

1. For each of the following scenarios, describe what it means to make a Type I and a Type II error.
   a. Researchers want to test if a new surgical procedure helps prevent recurring heart attacks. (Null hypothesis: the procedure does not prevent recurring heart attacks.)
   b. Cans of soda are labeled with "12 FL OZ". A company will test if this claim is accurate.
   c. For each of the scenarios above, comment on the consequences of making each error type.

7.2 Confidence Interval Approach to Hypothesis Testing

We can use a confidence interval to help us weigh the evidence against the null hypothesis. A confidence interval gives us a range of *plausible* values for μ. If the null value is in the interval, then μ_0 is a plausible value for μ. If the null value is *not* in the interval, then μ_0 is *not* a plausible value for μ.

1. State null and alternative hypotheses.
2. Decide on significance level α. Check assumptions (decide which confidence interval setting to use).
3. Find the critical value.
4. Compute confidence interval.
5. If the null value is *not* in the confidence interval, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results in the context of the problem.

Example: Is the average mercury level in dolphin muslces different from $2.5\mu g/g$? Test at the 0.05 level of significance. A random sample of 19 dolphins resulted in a mean of $4.4\mu g/g$ and a standard deviation of $2.3\mu g/g$.

1. $H_0 : \mu = 2.5$ and $H_A : \mu \neq 2.5$.
2. Significance level is $\alpha = 0.05$. The value of σ is unknown and $n = 19 < 30$, so we are in setting 3.
3. For setting 3, the critical value is $t_{df,\alpha/2}$. Here, $df = n-1 = 18$ and $\alpha/2 = 0.025$:

```
qt(0.025, 18)
```

```
## [1] -2.100922
```

4. The confidence interval is

$$\bar{x} \pm t_{df,\alpha/2}\frac{s}{\sqrt{n}} \tag{7.1}$$

$$4.4 \pm 2.101\frac{2.3}{\sqrt{19}} \tag{7.2}$$

$$4.4 \pm 1.109 \tag{7.3}$$

or $(3.29, 5.51)$.

5. Since the null value, 2.5, is not in the interval, it is *not* a plausible value for μ (at the 95% level of confidence). Therefore we reject the null hypothesis.
6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the true mean mercury level in dolphin muscles is *greater than* $2.5\mu g/g$.

Note: The alternative hypothesis is "not equal to", but we conclude

"greater than" because all of the plausible values in the confidence interval are greater than the null value.

Section Exercises

1. The weights of 6 randomly selected NFL linebackers were 243, 238, 229, 253, 248, 225. Suppose we want to test if the mean weight of NFL linebackers differs from 230 pounds. We will test at the 0.05 level of significance. (Note that $\bar{x} = 239.3$ and $s = 10.9$.)
   a. Write out the null and alternative hypotheses.
   b. Using the confidence interval approach, what decision should you make about rejecting or not rejecting the null hypotheses?
   c. Interpret your results in the context of the problem.
2. Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7. Use the confidence interval approach to test at the 0.1 level of confidence if the true mean number of eggs is 110.
3. Some Sac State students claim the average commute to campus is 45 minutes. A random sample of 32 students resulted in a mean of 31 minutes with standard deviation 18 minutes. Use the confidence interval approach to test their claim at the 0.05 level of confidence.
4. A city council member suspects that people speed through a certain intersection. They want to know the average speed at which drivers enter the intersection. They set up a speed camera and randomly sample 250 vehicles at different times of day, which results in a mean speed of 52 mph with standard deviation 8 mph. Suppose the posted speed limit is 45mph. Use the confidence interval approach at the 0.05 level of confidence to test the claim that drivers do not follow the speed limit.

7.3 Critical Value Approach to Hypothesis Testing

We learned about critical values when we discussed confidence intervals. Now, we want to use these values directly in a hypothesis test. We will compare these values to a value based on the data, called a **test statistic**.

Idea: the null is our "default assumption". If the null is true, how likely are we to observe a sample that looks like the one we have? If our sample is very inconsistent with the null hypothesis, we want to reject the null hypothesis.

7.3.1 Test statistics

Test statistics are similar to z- and t-scores:

$$\text{test statistic} = \frac{\text{point estimate} - \text{null value}}{\text{standard error}}.$$

In fact, they serve a similar function in converting a variable $\bar{X}$ into a distribution we can work with easily.

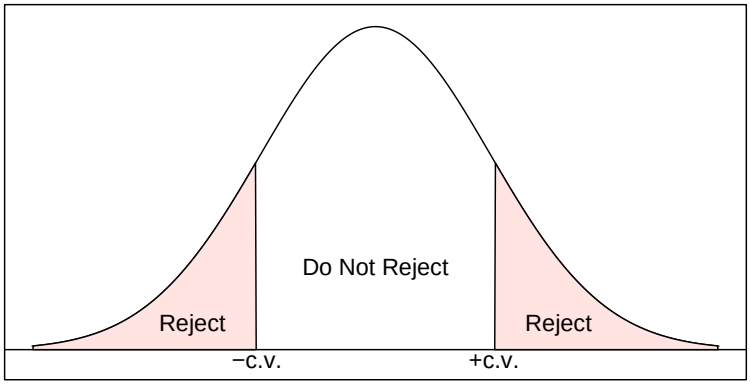
- **Large Sample Setting**: μ is target parameter, $n \geq 30$

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- **Small Sample Setting**: μ is target parameter, $n < 30$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The set of values for the test statistic that cause us to reject H_0 is the **rejection region**. The remaining values are the **nonrejection region**. The value that separates these is the critical value!



Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the critical values.
5. If the test statistic is in the rejection region, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

Example: Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7. Using the critical value approach at the 0.1 level of significance, test if the mean number of eggs in green sea turtle clutches is 110.

Solution: From the problem statement, $n = 20$, $\bar{x} = 107.3$, and $s = 13.7$.

1. $H_0 : \mu = 110$ and $H_A : \mu \neq 110$.
2. Significance level is $\alpha = 0.1$. The value of σ is unknown and $n = 20 < 30$, so we are in setting 3.
3. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.4}$$

$$= \frac{107.3 - 110}{13.7/\sqrt{20}} \tag{7.5}$$

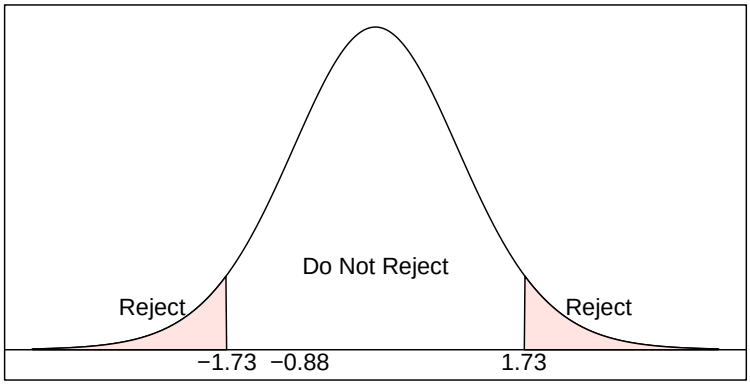
$$= -0.881 \tag{7.6}$$

4. The critical value is $t_{df,\alpha/2}$. Here, $df = n - 1 = 19$ and $\alpha/2 = 0.05$:

```
qt(0.05, 19, lower.tail = FALSE)
```

```
## [1] 1.729133
```

So the critical value is $t_{19,0.05} = 1.729$. 5. The test statistic is *not* in the rejection region, so we will fail to reject the null hypothesis:



6. At the 0.05 level of significance, the data provide insufficient evidence to conclude that the true mean number of eggs in a green sea turtle next differs from 110.

Section Exercises

1. The weights of 6 randomly selected NFL linebackers were $243, 238, 229, 253, 248, 225$. Suppose we want to test if the mean weight of NFL linebackers differs from 230 pounds. We will test at the 0.05 level of significance. (Note that $\bar{x} = 239.3$ and $s = 10.9$.)
   a. Write out the null and alternative hypotheses.
   b. Find the test statistic.
   c. Using the critical value approach, what decision should you make about rejecting or not rejecting the null hypotheses?
   d. Interpret your results in the context of the problem.
2. Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7. Use the critical value approach to test at the 0.1 level of confidence if the true mean number of eggs is 110.
3. Some Sac State students claim the average commute to campus is 45 minutes. A random sample of 32 students resulted in a mean of 31 minutes with standard deviation 18 minutes. Use the critical value approach to test their claim at the 0.05 level of confidence.
4. A city council member suspects that people speed through a certain intersection. They want to know the average speed at which drivers enter

the intersection. They set up a speed camera and randomly sample 250 vehicles at different times of day, which results in a mean speed of 52 mph with standard deviation 8 mph. Suppose the posted speed limit is 45mph. Use the critical value approach at the 0.05 level of confidence to test the claim that drivers do not follow the speed limit.

7.4 P-Value Approach to Hypothesis Testing

If the null hypothesis is true, what is the probability of getting a random sample that is as inconsistent with the null hypothesis as the random sample we got? This probability is called the **p-value**.

Example: Is the average mercury level in dolphin muscles different from $2.5\mu g/g$? Test at the 0.05 level of significance. A random sample of 19 dolphins resulted in a mean of $4.4\mu g/g$ and a standard deviation of $2.3\mu g/g$.

Probability of a sample *as inconsistent* as our sample is $P(t_{df}$ is as extreme as the test statistic). Consider

$$P(t_{18} > 3.6) = 0.001$$

but we want to think about the probability of being "as extreme" in *either direction* (either tail), so

$$\text{p-value} = 2P(t_{18} > 3.6) = 0.002$$

If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.

7.4.1 P-Values

- **Large Sample Setting**: μ is target parameter, $n \geq 30$,

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Small Sample Setting**: μ is target parameter, $n < 30$,

  $$2P(t_{df} > |t|)$$

  where t is the test statistic.

Note: $|a|$ is the "absolute value" of a. The absolute value takes a number and throws away the sign, so $|2| = 2$ and $|-3| = 3$.

Steps:

1. State the null and alternative hypotheses.

2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the p-value.
5. If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

We often use p-values instead of the critical value approach because they are meaningful on their own (they have a direct interpretation).

> **Example**: Is the average meerkat height different from 25cm? A random sample of 31 meerkats yielded a mean of 26.5cm and a standard deviation of 4.07cm. Use the p-value approach to test at the 0.05 level of significance.
>
> *Solution*: From the problem statement, $n = 31$, $\bar{x} = 26.5$ and $s = 4.07$. Then,
>
> 1. $H_0 : \mu = 25$ and $H_A : \mu \neq 25$.
> 2. Significance level is $\alpha = 0.05$. The value of σ is unknown and $n = 31 \geq 30$, so we are in setting 2 (standard normal).
> 3. The test statistic is
>
> $$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.7}$$
> $$= \frac{26.5 - 25}{4.07/\sqrt{31}} \tag{7.8}$$
> $$= 2.052 \tag{7.9}$$
>
> 4. The p-value is
>
> $$2P(z > |z|) - 2P(z > 2.052) = 0.040$$
>
> 5. Since p-value $= 0.04 < \alpha = 0.05$, reject the null hypothesis.
> 6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the true mean height of meerkats is greater than 25cm.

As before, this is the same conclusion we came to when we used the confidence interval and critical value approaches. All of these approaches are exactly equivalent.

Section Exercises

1. The weights of 6 randomly selected NFL linebackers were $243, 238, 229, 253, 248, 225$. Suppose we want to test if the mean weight of NFL linebackers differs from 230 pounds. We will test at the 0.05 level of significance. (Note that $\bar{x} = 239.3$ and $s = 10.9$.)
   a. Write out the null and alternative hypotheses.

   b. Find and interpret the p-value.
   c. Using the p-value approach, what decision should you make about rejecting or not rejecting the null hypotheses?
   d. Interpret your results in the context of the problem.
2. Researchers took a random sample of 20 green sea turtle nests and counted the number of eggs in each. They found a mean of 107.3 eggs with standard deviation 13.7. Use the p-value approach to test at the 0.1 level of confidence if the true mean number of eggs is 110.
3. Some Sac State students claim the average commute to campus is 45 minutes. A random sample of 32 students resulted in a mean of 31 minutes with standard deviation 18 minutes. Use the p-value approach to test their claim at the 0.05 level of confidence.
4. A city council member suspects that people speed through a certain intersection. They want to know the average speed at which drivers enter the intersection. They set up a speed camera and randomly sample 250 vehicles at different times of day, which results in a mean speed of 52 mph with standard deviation 8 mph. Suppose the posted speed limit is 45mph. Use the p-value approach at the 0.05 level of confidence to test the claim that drivers do not follow the speed limit.

R Lab: Hypothesis Tests for a Mean

To conduct hypothesis tests for a mean in R, we will use the `t.test` command. The arguments we will use for hypothesis testing are

- `x`: the variable that contains the data we want to use to construct a confidence interval.
- `mu`: the null value, μ_0.
- `conf.level`: the desired confidence level $(1 - \alpha)$.

We will again to use the `Loblolly` pine tree data.

```
attach(Loblolly)
```

Let's test if the average height of Loblolly pines differs from 40 feet. We will test at a 0.01 level of significance ($\alpha = 0.01$). So $H_0 : \mu = 40$ and $H_A : \mu \neq 40$ and the R command will look like

```
t.test(x = height, mu = 40, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  height
## t = -3.3851, df = 83, p-value = 0.001089
## alternative hypothesis: true mean is not equal to 40
## 99 percent confidence interval:
```

```
##  26.41761 38.31120
## sample estimates:
## mean of x
##   32.3644
```

The output from this test shows the following (top to bottom):

- the variable used in the hypothesis test.
- the value of the test statistic ($t = -3.3851$), the degrees of freedom (83), and the p-value (0.001).
- the alternative hypothesis ($H_0 : \mu \neq 40$).
- the confidence interval $(29.42, 38.31)$.
- the sample mean ($\bar{x} = 32.36$).

Based on this output, we have everything we need to conduct a hypothesis test using (A) the confidence interval approach, (B) the critical value approach, or (C) the p-value approach! In practice, we might include results from multiple approaches: At the 0.01 level of significance, there is sufficient evidence to reject the null hypothesis and conclude that the true mean height of Loblolly pines is less than 40 feet ($t = -3.385$, p-value$= 0.001$).

Comment: Notice are we using `t.test` and not a test using the normal distribution? Remember that when $n \geq 30$, the t and standard normal distributions are basically the same! Therefore, when using a computer, we often use the t-distribution regardless of sample size.

```
detach(Loblolly)
```

Chapter 8

Inference for a Proportion

In this module, we will continue our discussion on statistical inference with a discussion on hypothesis testing. In hypothesis testing, we take a more active approach to our data by asking questions about population parameters and developing a framework to answer those questions. We will root this discussion in confidence intervals before learning about several other approaches to hypothesis testing.

Module Learning Outcomes/Objectives

1. Perform and interpret inference for a population proportion.

R Objectives

1. Generate hypothesis tests for a proportion.
2. Interpret R output for tests of a proportion.

This module's outcomes correspond to course outcomes (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

8.1 Confidence Intervals for a Proportion

Inference for a proportion is really similar to inference for a mean! It turns out we can apply the Central Limit Theorem to the sampling distribution for a proportion. But wait - isn't our Central Limit Theorem only for means?

Think back to the binomial distribution (Section 4.3). A binomial experiment is made up of a series of Bernoulli trials, which result in 0s and 1s. If we add up these values, we get the number of successes x. If we take the mean of these successes, we get the *proportion* of successes. In short, $\bar{x} = \hat{p}$ and we can work with the sampling distribution for a sample mean!

The mean of a Bernoulli random variable is $\mu = p$ and the standard deviation is $\sigma = \sqrt{p(1-p)}$. So if we apply the Central Limit Theorem, $\hat{p}$ is approximately normally distributed with mean

$$\mu_{\hat{p}} = p$$

and standard error

$$\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Each of the confidence intervals for a mean uses the same logic:

$$\text{estimate } \pm \text{ critical value } \times \text{ standard error}$$

Confidence intervals for a proportion will do the same. We do not know the true value of p for the standard error, so we will plug in $\hat{p}$.

A $100(1-\alpha)\%$ confidence interval for p.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To use this formula, we need to check that $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$. (Note that $n\hat{p}$ is the number of successes and $n(1-\hat{p})$ is the number of failures, so this is another way to check this condition!)

Why? This relies on a normal approximation that does not work well if either of those quantities is less than or equal to 10. (This a topic which we have skipped, but the theory behind it is similar to the theory presented here for why we can use the Central Limit Theorem with proportions.)

> **Example:** Suppose we take a random sample of 27 US households and find that 15 of them have dogs. Find a 95% confidence interval for the proportion of US households with dogs.
>
> *Solution:* From the problem statement, $\alpha = 0.05$. Also, $\hat{p} = 15/27 = 0.56$. The number of successes (households with dogs) in the sample is 15 and the number of failures is 12, both greater than 10, so our assumptions are satisfied.
>
> The critical value is $z_{\alpha/2}$. Using the normal distribution applet with $\alpha = 0.05$, this yields a value of 1.96. Plugging everything in,
>
> $$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.56 \pm 1.96\sqrt{\frac{0.56 \times 0.44}{27}} = 0.56 \pm 0.19$$
>
> or a 95% confidence interval of (0.37, 0.75).
>
> Based on our sample, we can be 95% confident that the proportion of US households with dogs is between 0.37 and 0.75.

The concepts and interpretation behind these confidence intervals are the same as those for confidence intervals for a mean. Refer back to Module 6 for details.

Section Exercises

A survey of Stat 1 students resulted in the following data:

Year	Count
Freshman	16
Sophomore	11
Junior	3
Senior	5

1. We wish to find a 95% confidence interval for the proportion of freshmen.
   a. Based on the same data provided above, find $\hat{p}$, the sample proportion of freshmen.
   b. Confirm the condition that $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$.
   c. Determine the critical value.
   d. Find the 95% confidence interval for the proportion of freshmen.
   e. Interpret your interval in the context of the problem.
2. Find a 98% confidence interval for the proportion of sophomores.
3. Can you use the methods from this section to find a confidence interval for the proportion of juniors? Explain your thought process.

8.2 Hypothesis Tests for a Proportion

For a single proportion, the null and alternative hypotheses are

- $H_0 : p = p_0$
- $H_A : p \neq p_0$

We can perform a hypothesis test for p using the confidence interval, critical value, or p-value approach we covered previously. The concepts and interpretation are the same as those described in Module 7. You will also notice that the steps for each approach have not changed! The only modifications we need to make are to our setting, assumption, and a couple of formulas.

Setting and Assumptions: p is target parameter, $np_0 > 10$, $n(1-p_0) > 10$.

8.2.1 Confidence Interval Approach

The $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$$

Notice that we use p_0 in the standard error and *not* the sample proportion. This is different from how we dealt with the standard error when calculating confidence intervals outside of a hypothesis testing context. We do this because

the standard error is calculated based on the distribution based on the null hypothesis, which says that $p = p_0$.

Steps:

1. State null and alternative hypotheses.
2. Decide on significance level α. Check assumptions.
3. Find the critical value.
4. Compute confidence interval.
5. If the null value is *not* in the confidence interval, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results in the context of the problem.

Example: A quick internet search suggests that 38.4% of US households have dogs. Based on the sample described previously, is it reasonable to assume that the internet search is correct? Test at the 0.05 level of significance using a confidence interval approach.

Solution: We know from the previous example that $\hat{p} = 0.56$ and $n = 27$.

1. We want to see if the internet search is correct, so the null and alternative hypotheses are

$$H_0 : p = 0.384$$

$$H_A : p \neq 0.384$$

2. From the problem statement, $\alpha = 0.05$. Also, $np_0 = 27(0.384) = 10.4$ and $n(1 - p_0) = 27(0.616) = 16.6$, both greater than 10, so our assumptions are satisfied.
3. The critical value is $z_{\alpha/2}$. Using the normal distribution applet with $\alpha = 0.05$, this yields a value of 1.96.
4. Plugging everything in,

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} = 0.56 \pm 1.96\sqrt{\frac{0.384 \times (1-.384)}{27}} = 0.56 \pm 0.18$$

or a 95% confidence interval of (0.38, 0.74).
5. The null value is in the interval, so we fail to reject H_0.
6. At the 0.05 level of significance, the data provide *insufficient evidence* to conclude that the proportion of US Households with dogs differs from 0.384.

8.2.2 Critical Value Approach

The critical value is $z_{\alpha/2}$ and the test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Notice that we again plug in p_0 for the standard error!

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions.
3. Compute the value of the test statistic.
4. Determine the critical values.
5. If the test statistic is in the rejection region, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

Example: In 2007, the proportion of US adults who had ever had chickenpox was 61.4%. Since the chickenpox vaccine was introduced in 1995, it is reasonable to wonder if this value has decreased over time. A 2020 random sample of 100 US adults resulted in 13 with chickenpox. Use the critical value approach to test (at the 0.01 level of significance) whether the proportion of US adults who have ever had chickenpox is still 61.4%.

Solution: From the problem statement, $n = 100$ and $\widehat{p} = \frac{13}{100} = 0.13$.

1. We want to know if the true proportion of US adults who have ever had chickenpox is 0.614, so the null and alternative hypotheses are
$$H_0 : p = 0.614$$
$$H_A : p \neq 0.614$$
2. From the problem statement, $\alpha = 0.01$; for our assumptions, $np_0 = 100 \times 0.614 = 61.4$ and $n(1 - p_0) = 100 \times 0.386 = 38.6$, both greater than 10.
3. The test statistic is
$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.13 - 0.614}{\sqrt{\frac{0.614(1-0.614)}{100}}} = -9.94$$
4. The critical value is $z_{\alpha/2} = z_{0.01/2} = 2.58$
5. The rejection region is represented by values which are outside of $(-2.58, 2.58)$. So, the test statistic $z = -9.94$ is in the rejection region and we will reject the null hypothesis.
6. At the 0.01 level of significance, the data provide sufficient evidence to conclude that the true proportion of US adults who have ever had chickenpox is *less than* the 0.614 observed in 2007.

8.2.3 P-Value Approach

The p-value is
$$2P(Z > |z|)$$

where z is the test statistic described above.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions.
3. Compute the value of the test statistic.
4. Determine the p-value.
5. If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

Example: The 2020 US census suggested that 18.9% of the population identifies as Hispanic or Latino. A random sample of 53 Californians resulted in 20 who identified as Hispanic or Latino. Is the proportion of Hispanic or Latino Californians different from that of the US as a whole? Test at the 0.05 level of significance using the p-value approach.

Solution: From the problem statement, $n = 53$ and $\hat{p} = 20/53 = 0.377$.

1. We want to know if the proportion of Californians who identify as Hispanic or Latino differs from 0.189, so the null and alternative hypotheses are

$$H_0 : p = 0.189$$

$$H_A : p \neq 0.189$$

2. From the problem statement, $\alpha = 0.05$; For our assumptions, $np_0 = 53 \times 0.189 = 10.02$ and $n(1 - p_0) = 53 \times 0.811 = 42.98$, both greater than 10, so our assumptions are satisfied.
3. The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.377 - 0.189}{\sqrt{\frac{0.189(1-0.189)}{53}}} = 3.50$$

4. The p-value is

$$2P(Z > |z|) = 2P(Z > 3.50) = 2 \times 0.0002 = 0.0004$$

.

5. Since the p-value $0.0004 < \alpha = 0.05$, reject the null hypothesis.
6. At the 0.05 level of significance, the data provide sufficient evidence to conclude the proportion of Californias who identify as Hispanic or Latino is greater than that of the US as a whole.

Section Exercises

Another survey of Stat 1 students resulted in the following data:

Year	Count
Freshman	26
Sophomore	21
Junior	13
Senior	15

1. Suppose we wish to test the hypothesis that 50% of all Stat 1 students are freshmen using the confidence interval approach. We will test at the 0.05 level of significance.
   a. State the null and alternative hypothesis.
   b. Find $\hat{p}$, the sample proportion of freshmen.
   c. Determine α and check assumptions.
   d. Determine the critical value.
   e. How will this confidence interval formula differ from the one you found in Section 8.1 Exercise 1?
   f. Compute the confidence interval.
   g. What is your conclusion?
   h. Interpret your results in the context of the problem.
2. Suppose we want to know if 25% of Stat 1 students are sophomores. Use the confidence interval approach to complete this test at the 0.1 level of significance.
3. Use the critical value approach at the 0.05 level of significance to test the hypothesis that 25% of all Stat 1 students are juniors.
   a. State the null and alternative hypothesis.
   b. Find $\hat{p}$, the sample proportion of juniors.
   c. Determine α and check assumptions.
   d. Compute the test statistic.
   e. Determine the critical values.
   f. Sketch a normal curve, label the critical values, and indicate the rejection regions.
   g. What is your conclusion?
   h. Interpret your results in the context of the problem.
4. Use the critical value approach at the 0.02 level of significance to test the hypothesis that 20% of all Stat 1 students are seniors.
5. Use the p-value approach at the 0.01 level of significance to test the hypothesis that 40% of all Stat 1 students are freshmen.
   a. State the null and alternative hypothesis.
   b. Find $\hat{p}$, the sample proportion of freshmen.
   c. Determine α and check assumptions.
   d. Compute the test statistic.
   e. Use an applet to determine the p-value.
   f. What is your conclusion?
   g. Interpret your results in the context of the problem.
6. Use the p-value approach at the 0.05 level of significance to test the hypothesis that 10% of all Stat 1 students are seniors.

R: Hypothesis Tests for a Proportion

To generate confidence intervals and hypothesis tests for a proportion, we will use the command `binom.test`. This will give us slightly different results than the z-test we used throughout this module, but it is actually going to be more exact! This approach also does not have any limitations on the values $n\hat{p}$ or np_0. We use the z-test when working by hand because the exact binomial test is difficult to do on paper. The arguments we need are:

- `x`: the number of successes.
- `n`: the number of trials.
- `p`: the null value p_0.
- `conf.level`: the desired confidence level $(1 - \alpha)$.

Let's continue to use the example seen throughout this module. We have a random sample of 27 US households and 15 of them have dogs. We also have the claim that, in fact, 38.4% of US households have dogs. We will use a significance level of $\alpha = 0.05$.

Based on the prompt, there are $x = 15$ successes; $n = 27$ trials; and $p_0 = 0.384$. So the R command will look like

```
binom.test(x = 15, n = 27, p = 0.384, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  15 and 27
## number of successes = 15, number of trials = 27, p-value = 0.07591
## alternative hypothesis: true probability of success is not equal to 0.384
## 95 percent confidence interval:
##  0.3532642 0.7452012
## sample estimates:
## probability of success
##              0.5555556
```

The output shows (top to bottom):

- a summary of the data we entered, along with the p-value.
- the alternative hypothesis.
- a 95% confidence interval for p.
- the sample proportion $\hat{p}$.

Since this is slightly different from the test used when we discussed doing these calculations by hand, when we do hypothesis tests for a proportion using R, we will *not* use the critical value approach. Based on the confidence interval and p-value, at the 0.05 level of significance, the data provide insufficient evidence to conclude that the proportion of US Households with dogs differs from 0.384. (In general, we will come to the same conclusion whether we do these tests by

hand or using R.)

Chapter 9

Inference: Comparing Parameters

In this module, we extend the concepts from Module 6 to answer questions like "is there a difference between these means?" We will also consider hypothesis tests for whether a sample represents the population or closely matches a particular distribution.

Module Learning Outcomes/Objectives

1. Perform and interpret inference for
    a. the difference of two proportions.
    b. paired data and two sample means.

R Objectives

1. Generate hypothesis tests for the difference of two proportions.
2. Generate hypothesis tests for the difference of two means.
3. Interpret R output for tests of two proportions and two means.

This module's outcomes correspond to course outcomes (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

9.1 Hypothesis Tests for Two Proportions

Sometimes, we might like to *compare* two proportions. We do this by looking at their *difference*: $p_1 - p_2$. This is going to be fairly similar to the tests we used for a single proportion. Let n_1 be the sample size for the first group and p_1 the proportion for the first group. Similarly, let n_2 be the sample size for the second group and p_2 the proportion for the second group.

Conditions:

1. Independence within and between groups (generally satisfied if the data are from random samples or a randomized experiment).
2. We need $n_1p_1 > 10$ and $n_1(1-p_1) > 10$ **and** $n_2p_2 > 10$ and $n_2(1-p_2) > 10$
   - Recall that np is the number of successes and $n(1-p)$ is the number of failures, so we can also check if both groups have at least 10 successes and at least 10 failures.

If these conditions are satisfied, the standard error is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

and we can calculate confidence intervals and perform hypothesis tests on $p_1 - p_2$.

9.1.1 Confidence Intervals for Two Proportions

A $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Example: Last semester, a professor taught two sections of the same class. The first section had 30 students, 19 of whom passed. The second section had 43 students, 30 of whom passed. Find and interpret at 95% confidence interval for the difference in pass rates (proportion of students who passed) for the two sections.

Solution: First, we should note that $n_1 = 30$ and $\hat{p}_1 = 19/30 = 0.633$ and that $n_2 = 43$ and $\hat{p}_2 = 30/43 = 0.698$. For a 95% confidence interval, the critical value is $z_{0.025} = 1.96$.

Now for the first section, there were 19 passes and $30-19-11$ failures. For the second section, there were 30 passes and $43-30-13$ failures. All our "successes" and failures are at least 10, so our conditions are satisfied.

Then

$$\begin{aligned}
\hat{p_1} - \hat{p_2} \pm z_{\alpha/2} &\times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \\
0.633 - 0.698 \pm 1.96&\sqrt{\frac{0.633(1-0.633)}{30} + \frac{0.698(1-0.698)}{43}} \\
-0.065 &\pm 1.96\sqrt{0.0126} \\
-0.065 &\pm 0.0248 \\
(-0.065 - 0.0248, &- 0.065 + 0.0248) \\
(-0.090, &- 0.040)
\end{aligned}$$

We can be 95% confident that the true difference in proportion of students who passed section 1 versus section 2 was between -0.09 and -0.04.

9.1.2 Critical Values, Test Statistics, and P-Values

Often, we are interested in checking whether $p_1 = p_2$, which results in a null hypothesis of $H_0 : p_1 - p_2 = 0$ (where the null value is zero). In this case, we use a *pooled proportion* to estimate p in the standard error.

This pooled proportion is calculated as

$$\hat{p}_{\text{pooled}} = \frac{\text{total number of successes}}{\text{total number of cases}} = \frac{\hat{p_1} n_1 + \hat{p_2} n_2}{n_1 + n_2}$$

which makes the standard error in this case

$$\text{Standard Error} = \sqrt{\frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_2}}$$

The critical value is $z_{\alpha/2}$. The test statistic is

$$\begin{aligned} z &= \frac{\hat{p_1} - \hat{p_2}}{\text{standard error}} \\ &= \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}} \end{aligned}$$

and the p-value is

$$2P(Z > |z|)$$

where z is the test statistic.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions, $n_1 p_1 > 10$ and $n_1(1 - p_1) > 10$ **and** $n_2 p_2 > 10$ and $n_2(1 - p_2) > 10$.
3. Compute the value of the test statistic.
4. Determine the critical value or p-value.
5. Decision.
   - For the *critical value approach*: If the test statistic is in the rejection region, reject the null hypothesis.
   - For the *p-value approach*: If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

Example: A researcher wants to know if female dogs are more likely to bark when someone comes to the door than male dogs. He takes

a random sample of 50 male and 50 female dogs and tracks whether they bark at the door. Of the male dogs, 37 barked at the door. Of the female dogs, 39 barked at the door. Test at the 0.1 level of significance whether female dogs bark at the door more often.

Solution: Let male dogs be 1 and female dogs be 2. Then from the problem statement, $n_1 = 50$ and $\hat{p}_1 = 37/50 = 0.74$, and $n_2 = 50$ and $\hat{p}_2 = 39/50 = 0.78$.

1. $H_0 : p_1 - p_2 = 0$ vs $H_A : p_1 - p_2 \neq 0$
2. Our level of significance is $\alpha = 0.1$. For male dogs, there were 37 successes and $50 - 37 = 13$ failures; for female dogs, there were 39 successes and $50 - 39 - 11$ failures. So our conditions are satisfied.
3. For the test statistic, we need the pooled proportion.

$$\begin{aligned}\hat{p}_{\text{pooled}} &= \frac{\text{total number of successes}}{\text{total number of cases}} \\ &= \frac{37 + 39}{50 + 50} \\ &= \frac{76}{100} \\ &= 0.76\end{aligned}$$

and the standard error is

$$\begin{aligned}\text{SE} &= \sqrt{\frac{0.76(1 - 0.76)}{50} + \frac{0.76(1 - 0.76)}{50}} \\ &= \sqrt{\frac{0.1824}{50} + \frac{0.1824}{50}} \\ &= \sqrt{0.007296} \\ &= 0.085\end{aligned}$$

Then the test statistic is

$$z = \frac{\hat{p_1} - \hat{p_2}}{\text{SE}} = \frac{0.74 - 0.78}{0.085} = -0.47$$

4. For the critical value approach, we would find $z_{0.1/2} = 1.645$. For the p-value approach, we would find

$$2P(Z > |-0.47|) = 2(0.3192) = 0.6384$$

5. Since the test statistic is in the rejection region (and the p-value $= 0.6384 > \alpha = 0.1$), we fail to reject the null hypothesis.
6. At the 0.1 level of significance, the data provide insufficient evidence to conclude that female or male dogs are more likely to bark when someone comes to the door.

Calculating the pooled standard error and test statistic can be pretty tedious! In practice, we almost always let a computer handle these types of tests.

9.2 Hypothesis Tests for Two Means

What if we wanted to compare two means? We begin by discussing paired samples. This will feel very familiar, since it's essentially the same as hypothesis testing for a single mean. Then we will move on to independent samples, which will require a couple of adjustments.

9.2.1 Paired Samples

Sometimes there is a special correspondence between two sets of observations. We say that two sets of observations are **paired** if each observation has a natural connection with exactly one observation in the other data set. Consider the following data from 30 students given a pre- and post-test on a course concept:

Student	Pre-Test	Post-Test
1	52	70
2	71	98
3	13	65
...	...	...
30	48	81

The natural connection between "pre-test" and "post-test" is the student who took each test! Often, paired data will involve similar measures taken on the *same item or individual*. We *pair* these data because we want to compare two means, but we also want to account for the pairing.

Why? Consider: If a student got a 13% on the pre-test, I would love to see them get a 60% on the post-test - that's a huge improvement! But if a student got an 82% on the pre-test, I would *not* like to see them get a 60% on the post-test. Pairing the data lets us account for this connection.

So what do we do with paired data? Fortunately, this part is easy! We start by taking the difference between the two sets of observations. In the pre- and post-test example, I will take the pre-test score and subtract the post-test score:

Student	Pre-Test	Post-Test	**Difference**
1	52	70	**18**
2	71	98	**27**
3	13	65	**52**
...	...	...	...
30	48	81	**33**

Then, we do a test of a *single mean* on the differences where

- $H_0 : \mu_{\text{d}} = 0$

- $H_A : \mu_{\text{d}} \neq 0$

Note that the subscript "d" denotes "difference". We will use the exact same test(s) as in the previous sections:

- **Large Sample Setting**: μ_{d} is target parameter, $n_{\text{d}} \geq 30$,

  $$z = \frac{\bar{x}_{\text{d}}}{s_{\text{d}}/\sqrt{n_{\text{d}}}}$$

  and the p-value is

  $$2P(Z > |z|)$$

  where z is the test statistic.

- **Small Sample Setting**: μ_{d} is target parameter, $n_{\text{d}} < 30$,

  $$t = \frac{\bar{x}_{\text{d}}}{s_{\text{d}}/\sqrt{n_{\text{d}}}}$$

  and the p-value is

  $$2P(t_{df} > |t|)$$

  where t is the test statistic.

Here, n_{d} is the number of pairs.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions (decide which setting to use).
3. Compute the value of the test statistic.
4. Determine the critical values or p-value.
5. For the *critical value approach*: If the test statistic is in the rejection region, reject the null hypothesis. For the *p-value approach*: If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

9.2.2 Independent Samples

In **independent samples**, the sample from one population does not impact the sample from the other population. In short, we take two *separate samples* and compare them.

- $H_0 : \mu_1 = \mu_2 \quad \rightarrow \quad H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 \neq \mu_2 \quad \rightarrow \quad H_A : \mu_1 - \mu_2 \neq 0$

If we use $\bar{x}$ to estimate μ, intuitively we might use $\bar{x}_1 - \bar{x}_2$ to estimate $\mu_1 - \mu_2$. To do this, we need to know something about the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Consider: if X_1 is Normal(μ_1, σ_1) and X_2 is Normal(μ_2,σ_2) with σ_1 and σ_2 are known, then for independent samples of size n_1 and n_2,

- $\bar{X}_1 - \bar{X}_2$ is Normal($\mu_{\bar{X}_1-\bar{X}_2}$, $\sigma_{\bar{X}_1-\bar{X}_2}$).
- $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$
- $\sigma_{\bar{X}_1-\bar{X}_2} = \sigma_1 - \sigma_2$

so then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1/n_1 - \sigma_2/n_2}}$$

has a standard normal distribution. But, as we mentioned earlier, we rarely work in that setting where the population standard deviation is known. Instead, we will use s_1 and s_2 to estimate σ_1 and σ_2. For independent samples of size n_1 and n_2,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1/n_1 - s_2/n_2}}$$

has a t-distribution with degrees of freedom

$$\Delta = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

rounded *down* to the nearest whole number. (Note that Δ is the uppercase Greek letter, "delta".) If $n_1 = n_2$, this simplifies to

$$\Delta = (n-1)\left(\frac{(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4}\right)$$

Tip: Generally, people do not calculate Δ by hand. Instead, we use a computer to do these kinds of tests.

The Two-Sample T-Test

Assumptions:

- Simple random samples.
- Independent samples.
- Normal populations or large ($n \geq 30$) samples.

Steps for Critical Value Approach:

1. $H_0 : \mu_1 - \mu_2 = 0$ and $H_A : \mu_1 - \mu_2 \neq 0$
2. Check assumptions; select the significance level α.
3. Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1/n_1 - s_2/n_2}}$$

   Note that we assume under the null hypothesis that $\mu_1 - \mu_2 = 0$, which is why we replace this quantity with 0 in the test statistic.
4. The critical value is $\pm t_{df,\alpha/2}$ with $df = \Delta$.
5. If the test statistic falls in the rejection region, reject the null hypothesis.

6. Interpret in the context of the problem.

Steps for P-Value Approach:

1. $H_0 : \mu_1 - \mu_2 = 0$ and $H_A : \mu_1 - \mu_2 \neq 0$
2. Check assumptions; select the significance level α.
3. Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1/n_1 - s_2/n_2}}$$

   Note that we assume under the null hypothesis that $\mu_1 - \mu_2 = 0$, which is why we replace this quantity with 0 in the test statistic.
4. The p-value is $2P(t_{df} > |t|)$ with $df = \Delta$.
5. If p-value $< \alpha$, reject the null hypothesis.
6. Interpret in the context of the problem.

Notice that the only difference between the critical value and p-value approaches are steps 4 and 5.

Example: Researchers wanted to detemine whether a dymanic or static approach would impact the time needed to complete neurosurgeries. The experiment resulted in the following data from simple random samples of patients:

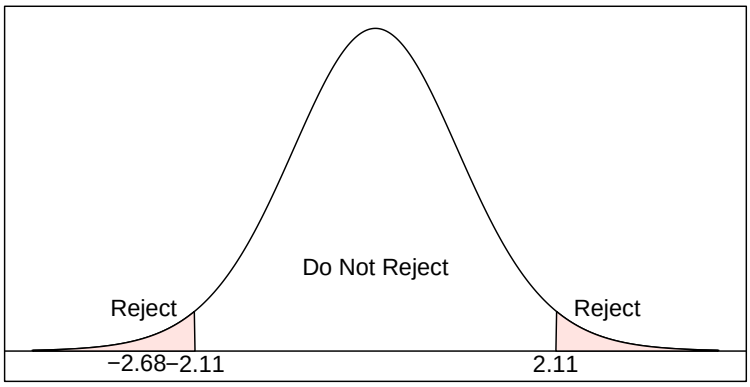
Dynamic	Static
$\bar{x}_1 = 394.6$	$\bar{x}_2 = 468.3$
$s_1 = 84.7$	$s_2 = 38.2$
$n_1 = 14$	$n_2 = 6$

Times are measured in minutes. Assume X_1 and X_2 are reasonably normal.

1. $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_2$
2. Let $\alpha = 0.05$ (this will be our default when a significance level is not given)
   - We are told these are simple random samples.
   - There's no reason that time for a neurosurgery with the dynamic system would impact time for the static system (or vice versa), so it's reasonable to assume these samples are independent.
   - We are told to assume that X_1 and X_2 are reasonably normal.
3. The test statistic is

$$t = \frac{394.6 - 468.3}{84.7^2/14 + 38.2^2/6} = -2.681$$

4. Then
$$df = \Delta = \frac{(84.7^2/14) + (38.2^2/6)^2}{\frac{(84.7^2/14)^2}{14-1} + \frac{(38.2^2/6)^2}{6-1}} = 17$$
when rounded down. The critical value is
$$t_{17,0.025} = 2.110$$
and the p-value is
$$2P(t_{17} > |-2.681|) = 2(0.0079) = 0.0158$$
5. For the critical value approach,



Since the test statistic is in the rejection region, we reject the null hypothesis. For the p-value approach, since p-value $= 0.158 < \alpha = 0.05$, reject the null hypothesis.

6. At the 0.05 level of significance, the data provide sufficient evidence to conclude that the mean time for the dynamic system is less than the mean time for the static system.

We can also construct a $(1-\alpha)100\%$ **confidence interval** for the difference of the two population means:
$$(\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
which we interpret as we interpret other confidence intervals, including in our interpretation that we are now considering the **difference of two means**.

R Lab: Comparing Parameters

Hypothesis Tests for Two Proportions

To compare two proportions, we will use the command `prop.test`. This is similar to `binom.test`, but the latter command does not allow us to compare two proportions. We will need the following arguments:

- `x`: a listing of the numbers of successes in each of the two groups. This will take the form `x = c(x1, x2)`.
- `n`: a listing of the numbers of trials for each group. This will take the form `n = c(n1, n2)`.
- `conf.level`: the confidence level $(1 - \alpha)$.

Note that order matters in `c(x1, x2)` and `c(n1, n2)`. Make sure to keep track of which variable you have set as 1 an which as 2. This test also assumes a null hypothesis of $p_1 = p_2$.

This test has a few behind-the-scenes tweaks relative to what we do by hand. This means that the results might be slightly different than the results you get when running these tests by hand. That's ok!

The `sleep` dataset in R contains data on two groups (10 in each) of patients given soporific drugs (drugs designed to induce sleep). We want to examine whether the *proportion* of patients who experienced an increase in hours of sleep differs between the two groups.

I have this set up with two variables, `d1` and `d2`, which represent drug 1 and drug 2. Each variable is 1 if the patient experienced an increase in hours of sleep and 0 if they did not. Let's print these out and find out how many successes were in each group.

```
d1
```

```
##  [1] 1 0 0 0 0 1 1 1 0 1
```

```
d2
```

```
##  [1] 1 1 1 1 0 1 1 1 1 1
```

We can find the total number of successes for each by summing the values in each variable. Let's do that in R using the `sum` command:

```
sum(d1)
```

```
## [1] 5
```

```
sum(d2)
```

```
## [1] 9
```

So the numbers of successes are $x_1 = 5$ and $x_2 = 9$ for group sizes $n_1 = n_2 = 10$. For the `prop.test` command, this will look like `x = c(5, 9)` and `n =`

`c(10,10)`. We will use an $\alpha = 0.1$ level of significance. Then

```
prop.test(x = c(5, 9), n = c(10,10), conf.level = 0.9)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(5, 9) out of c(10, 10)
## X-squared = 2.1429, df = 1, p-value = 0.1432
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.803296023  0.003296023
## sample estimates:
## prop 1 prop 2
##    0.5    0.9
```

The output of this test is (top to bottom)

- The data provided in the input.
- A test statistic and degrees of freedom (these are part of the behind-the-scenes tweaks and you can ignore them!) along with a p-value.
- When a hypothesis test says "two sided" that means the null hypothesis represents the "not equal" condition that we work with.
- The requested confidence interval.
- The sample proportions.

Although the sample proportions appear to be different, the sample sizes are very small! Therefore it is unsurprising that the data provide insufficient evidence to conclude that the drugs differ in their ability to increase hours slept ($p = 0.143$ and the confidence interval includes 0).

Hypothesis Tests for Two Means

The math has only gotten more cumbersome! Let's use R to quickly run these types of tests without having to do any calculations by hand.

There is data built into R that shows the effect of Vitamin C on tooth growth in guinea pigs through (A) ascorbic acid or (B) orange juice. (Each guinea pig was randomly assigned to either ascorbic acid *or* orange juice.) We want to compare the ascorbic acid group to the orange juice group to see if one has more tooth growth than the other. This is currently in a data set called `teeth`, which contains two variables: `aa`, the tooth length for guinea pigs in the ascorbic acid group and `oj` the tooth length for the orange juice group.

```
attach(teeth)
```

To run a two-sample test comparing means in R, we continue to use the command `t.test`. The arguments we need in this case are:

- `x`: the first variable.

- `y`: the other variable.
- `mu`: the null value, usually $\mu_1 - \mu_2 = 0$.
- `paired`: set this equal to `TRUE` for paired t tests; set it equal to `FALSE` for independent samples.
- `conf.level`: the desired confidence level $(1 - \alpha)$.

In this case, we are interested in variables `x = aa` and `y = oj`. The null value is `mu = 0`. Guinea pigs were randomly assigned to each treatment group, so these are independent samples and `paired = FALSE`. Finally, we will go ahead and test this at a 0.05 level of significance, so `conf.level = 0.95`. Putting that all together, the R command looks like

```
t.test(x = aa, y = oj, mu = 0, paired = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  aa and oj
## t = -1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

The R output shows (top to bottom)

- variables entered.
- the test statistic, degrees of freedom, and p-value.
- the alternative hypothesis.
- a confidence interval for the difference of the two means.
- sample means for each variable.

Based on the output, at the 0.05 level of significance, the data provide insufficient evidence to conclude that the mean tooth length for guinea pigs receiving ascorbic acid differs from the guinea pigs receiving orange juice ($p = 0.061$ and the confidence interval includes 0).

Chapter 10

More on Regression

In this module, we return to regression to discuss some of the nuances of the linear regression model. We also talk about two hypothesis tests related to regression.

Module Learning Outcomes/Objectives

1. Test the statistical significance of a predictor variable.
2. Test overall model significance.
3. Use plots to check regression model assumptions.

Recall that our regression model

$$\hat{y} = b_0 + b_1 x$$

describes the linear relationship between some predictor variable x and some outcome or response variable y.

10.1 A Hypothesis Test for a Predictor Variable

In the regression framework, we want to ask if the predictor variable x is useful in predicting y. If it's not at all useful, then the best we can do in this framework is to predict y using its mean $\hat{y} = \bar{y}$. This is also what happens in the linear regression model when $b_1 = 0$, so if x is not useful in predicting y, then b_1 should be 0.

With that in mind, our null hypothesis, that x is not useful in predicting y, can be translated into statistical notation as

$$H_0 : \beta_1 = 0$$

It turns out this framework is remarkably similar to the hypothesis test for a mean we discussed in Module 7. Recall the test statistic was

$$t = \frac{\text{estimate} - \text{null value}}{\text{standard error}}$$

which now looks like

$$t = \frac{b_1 - 0}{\text{SE}(b_1)}$$

This quantity follows a t-distribution with $n - 1$ degrees of freedom.

> Example: Suppose we have a dataset with $n = 40$ observations and find $b_1 = 2.5$ with $\text{SE}(b_1) = 1.2$. We will let $\alpha = 0.05$.
>
> Then, using our hypothesis testing framework from Section 7.4, we can calculate $t = 2.5/1.2 = 2.08$ and the p-value $2P(t_{39} > |2.08|) = 0.022$.
>
> Then, since the p-value$= 0.022 < \alpha = 0.05$, we reject the null hypothesis and conclude that $\beta_1 \neq 0$. That is, x is useful in predicting y.

In practice, we use a computer to generate thee values. These outputs typically look something like this

```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) -1.87401599 0.160143302 -11.70212  7.359171e-26
## waiting      0.07562795 0.002218541  34.08904 8.129959e-100
```

10.2 A Hypothesis Test for a Regression Model

10.2.1 The F-Distribution

The **F-test** relies on something called the F distribution. The F distribution has two parameters: $df_1 = df_G$ and $df_1 = df_E$. The F distribution always takes on positive values, so an *extreme* or *unusual* value for the F distribution will correspond to a large (positive) number.

When we run these types of tests, we almost always use the p-value approach. If you are using `R` for your distributions, the command is `pf(F, df1, df2, lower.tail=FALSE)` where `F` is the test statistic.

> *Example:* Suppose I have a test with 100 observations and 5 groups. I find $MSG = 0.041$ and $MSE = 0.023$. Then
>
> $$df_G = k - 1 = 5 - 1 = 4$$
>
> and
>
> $$df_E = n - k = 100 - 5 = 95$$

The test statistic is

$$f = \frac{0.041}{0.023} = 1.7826$$

To find the p-value using R, I would write the command

```
pf(1.7826, 4, 95, lower.tail=FALSE)
```

```
## [1] 0.1387132
```

and find a p-value of 0.1387.

Here is a nice F-distribution applet. For this applet, $\nu_1 = df_1$ and $\nu_2 = df_2$. Plug in your F test statistic where it indicates "x =" and your p=value will appear in the red box next to "P(X>x)". When you enter your degrees of freedom, a visualization will appear similar to those in the Rossman and Chance applets we used previously.

10.3 Model Assumptions

We have some assumptions we require in order for our hypothesis tests to be valid. These are

1. A linear equation adequately describes the relationship between x and y.
2. The errors are approximately normally distributed.
3. The errors have constant variance.
4. The errors are not correlated.

In general, if a linear equation does not do a good job describing the relationship between x and y, then we have no reason to run this type of model. Instead, we could develop a slightly more complex regression model or use another modeling technique, topics which are outside the scope of this class.

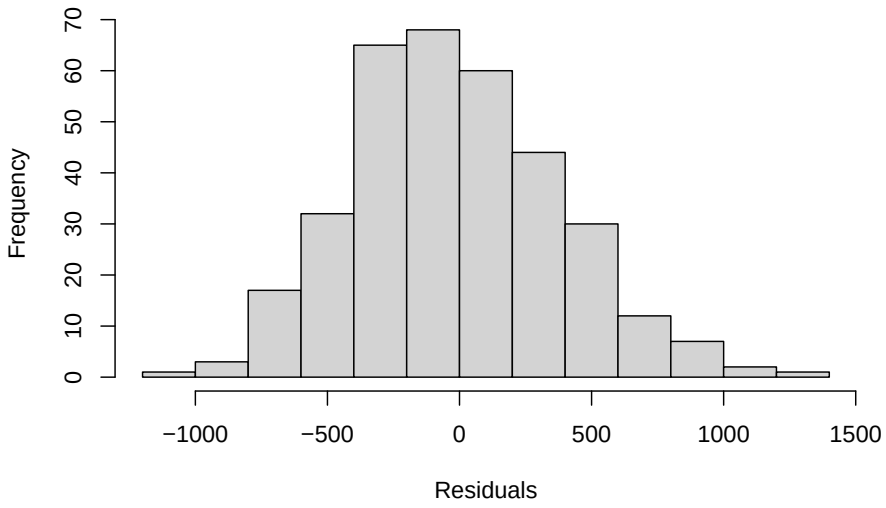
The rest of our assumptions have to do with the errors, which we approximate using our residuals $r = y - \hat{y}$.

10.3.1 Normality of Errors

In Module 3, we generated a regression model that used a penguin's flipper length (x, in mm) to predict its weight (y, in g):

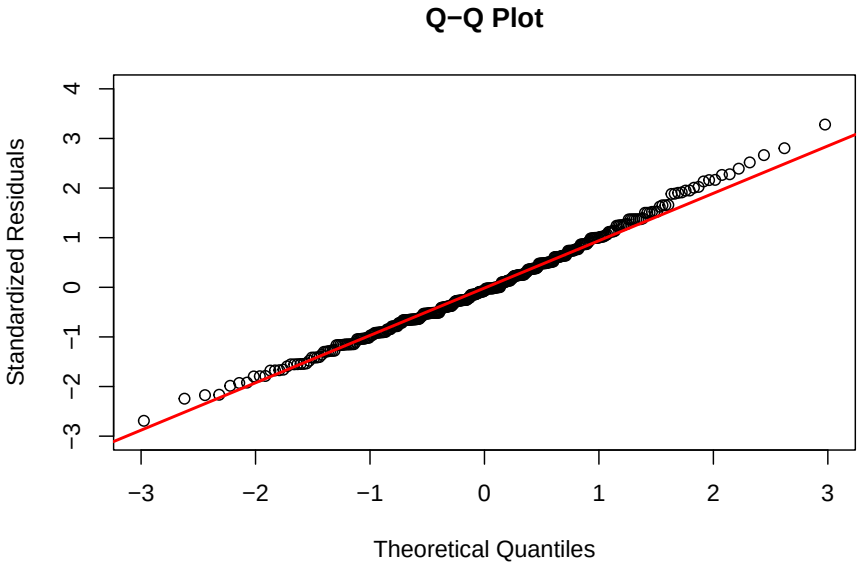
$$\hat{y} = -5780.83 + 49.69x$$

We could examine the distribution of this model's residuals using a histogram



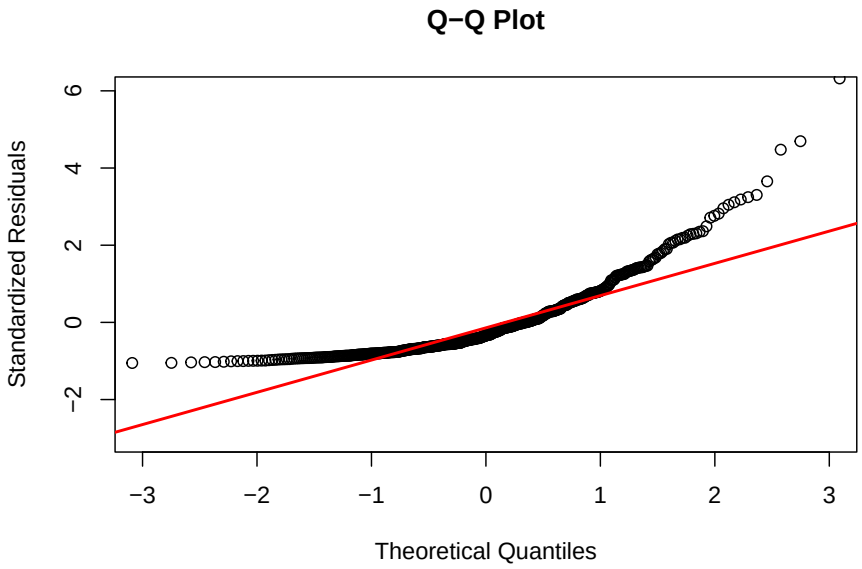
However, it can be kind of difficult to use a histogram to accurately determine normality.

Instead, we typically use what we call a **Q-Q Plot**. A Q-Q Plot is a scatterplot that plots the model's standardized residuals against the quantiles of a standard normal distribution. (Recall that *standardized* means we have z-scored everything.)

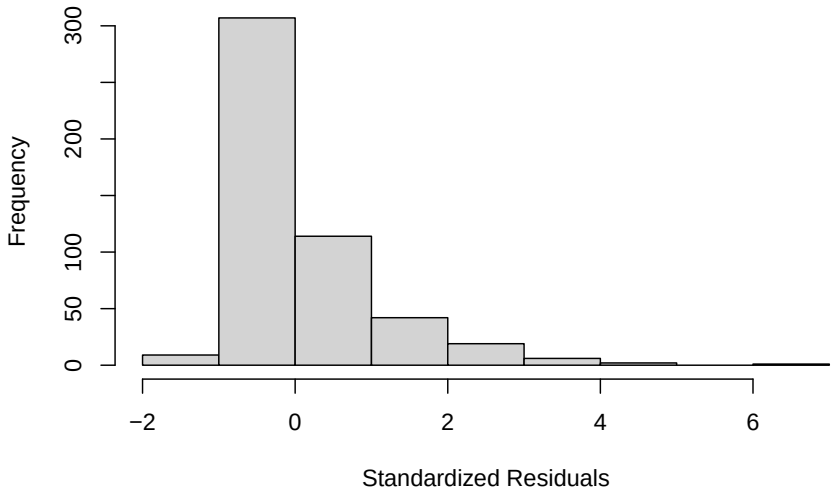


If the points fall along the line $y = x$, then the standardized residuals match the quantiles of the standard normal distribution, which means they are normally distributed! Here, the line $y = x$ has been added to the plot in red to make it easier to visually confirm normality.

If a lot of the points are far from this line, then we have violated our normality assumption.



In this example, our points are far from the $y = x$ line in both tails. In fact, these residuals are heavily skewed!



In settings where our residuals deviate significantly from normality, we should not use our linear regression model as-is. Techniques to "fix" this issue include

transformations on y and other modeling approaches, both of which are outside the scope of this class.

10.4 Constant Variance

When we calculate

10.5 Uncorrelated Errors

Chapter 11

Chi-Square Tests

In this module, we will continue our discussion on statistical inference with a discussion on hypothesis testing. In hypothesis testing, we take a more active approach to our data by asking questions about population parameters and developing a framework to answer those questions. We will root this discussion in confidence intervals before learning about several other approaches to hypothesis testing.

Module Learning Outcomes/Objectives

1. Perform and interpret inference for
    a. a population variance.
    b. the ratio of two variances.
    c. tests of goodness of fit and contingency tables.

This module's outcomes correspond to course outcomes (6) apply statistical inference techniques of parameter estimation such as point estimation and confidence interval estimation and (7) apply techniques of testing various statistical hypotheses concerning population parameters.

11.1 Inference for a Population Variance

Sometimes, it may be of interest to examine directly the variability of a population. Why? Suppose we have some medication that comes in a pill form. We know that each pill has an average of 10mg of active ingredient. For this medication to be consistently effective, we want to make sure that the amount of active ingredient does not *vary* too much from one pill to the next. We examine this using tests for population variance.

11.1.1 The Chi-Square Distribution

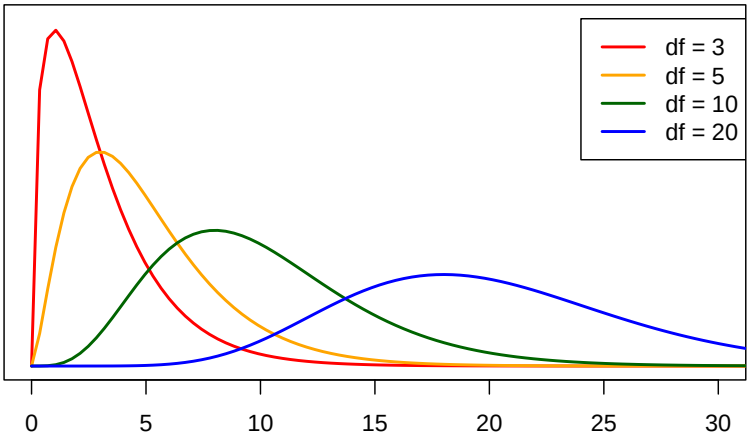
Numerically, the variance is different from the mean because it cannot be negative… so it won't make sense to use a normal or t distribution. In order to do hypothesis testing for a variance, we need to learn a little bit about a new distribution, the **chi-square distribution**.

Chi-square distributions

- have curves that start at 0 and extend indefinitely in the positive direction.
- are fully determined by parameter df.
- are right-skewed.
- have means equal to df and variances equal to $2 \times$ df.

These makes it a great choice for modeling continuous random variables that can only take on positive values, like the variance!

The chi-square distribution is denoted χ^2_{df}, where χ is the Greek letter "chi" and df is the degrees of freedom. The plot below shows several examples of chi-square distributions with different degrees of freedom.



Note that we are not able to see the full distributions, so that right-skew may not always be apparent. The chi-square distribution goes on forever in the positive direction, though, so eventually each curve *has* to skew to the right. We should also notice that, for larger values of df, the curve looks a bit like the normal curve.

11.1.2 Confidence Intervals for σ

A $(1-\alpha)100\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right)$$

where $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$ are the critical values based on significance level α and degrees of freedom $n-1$. We have to consider two separate critical values because the chi-square distribution is not symmetric, which means that - unlike the normal and t distributions - they will *not* be the same value with different signs.

To get a $(1-\alpha)100\%$ confidence interval for σ, we will take the square root of each side:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}\right)$$

This follows a slightly different pattern than the confidence intervals we saw previously, but we can use and interpret it in exactly the same way.

11.1.3 Hypothesis Tests for σ

Think back to the example given at the start of this section. Suppose we want the amount of medication in those pills to have a standard deviation no more than 0.5mg (a variance of 0.25mg^2).

In this case, we will test whether a variance is equal to some quantity or not. The null and alternative hypotheses are

- H_0: the variance is 0.25mg^2.
- H_A: the variance is NOT 0.25mg^2.

In general, and using statistical notation, this will look like

- H_0: $\sigma^2 = \sigma_0^2$
- H_A: $\sigma^2 \neq \sigma_0^2$

Setting and assumptions: σ^2 (or σ) is the target parameter, the population from which the sample was taken is normally distributed.

Confidence Interval Approach

Steps:

1. State null and alternative hypotheses.
2. Decide on significance level α. Check assumptions.
3. Find the critical values $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$

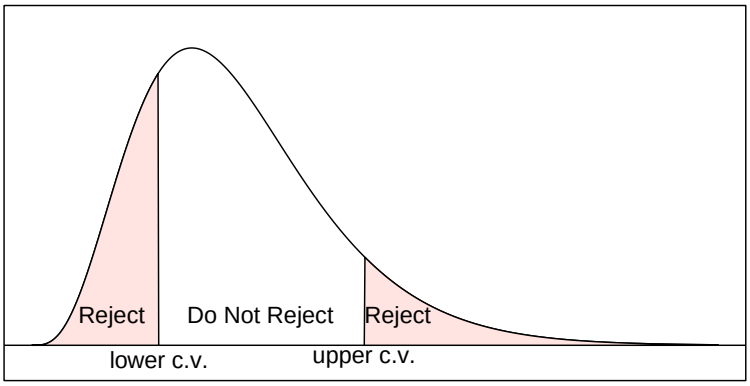
4. Compute confidence interval.
5. If the null value is *not* in the confidence interval, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results in the context of the problem.

Critical Value Approach

The critical values are $\chi^2_{1-\alpha/2,n-1}$ and $\chi^2_{\alpha/2,n-1}$. The test statistic is

$$T = \frac{(n-1)s}{\sigma_0^2}$$

where s is the sample standard deviation. The rejection region for this test is $T < \chi^2_{\alpha/2,n-1}$ OR $T > \chi^2_{1-\alpha/2,n-1}$.



Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions.
3. Compute the value of the test statistic.
4. Determine the critical values.
5. If the test statistic is in the rejection region, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

P-Value Approach

To find the p-value:

- If your test statistic $T < df$, the p-value is $2P(\chi^2_{df} < T)$.
- If your test statistic $T \geq df$, the p-value is $2P(\chi^2_{df} > T)$.

Steps:

1. State the null and alternative hypotheses.
2. Determine the significance level α. Check assumptions.
3. Compute the value of the test statistic.
4. Determine the p-value.
5. If p-value $< \alpha$, reject the null hypothesis. Otherwise, do not reject.
6. Interpret results.

11.2 The Ratio of Two Variances

11.3 Goodness of Fit

11.4 Contingency Tables

Chapter 12

ANOVA

In this module, we extend the concepts from Module 6 to answer questions like "is there a difference between these means?" We will also consider hypothesis tests for whether a sample represents the population or closely matches a particular distribution.

Module Learning Outcomes/Objectives

Perform and interpret inference for

1. Interpret an ANOVA.
2. Use the Bonferroni correction to conduct multiple comparisons.

12.1 What is the Analysis of Variance (ANOVA)

Now that we've examined tests for one and two means, it's natural to wonder about three or more means. For example, we might want to compare three different medications: treatment 1 (t_1), treatment 2 (t_2), and treatment 3 (t_3). Based on what we've learned so far, we might think to do pairwise comparisons, examining t_1 vs t_2, then t_2 vs t_3, then t_1 vs t_3. Unfortunately, this tends to increase our Type I error!

Think of it this way: if I set my confidence level to 95%, I'm setting my Type I error rate to $\alpha = 0.05$. In general terms, this means that about 1 out of every 20 times I run my experiment, I would make a type I error. If I went ahead and ran, say, 20 tests comparing two means, my *overall* Type I error rate is going to increase - there's a pretty significant chance that at least one of those comparisons will results in a Type I error!

Instead, we will use a test that allows us to ask: "Are all these means the same?" This is called the **an**alysis **o**f **va**riance, or ANOVA.

- H_0: The mean outcome is the same across all groups.

- H_A: At least one mean differs from the rest.

In statistical notation, these hypotheses look like:

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_A : \mu_i \neq \mu_j$ for at least one pair (i, j)

where k is the number of means being compared and the notation μ_i represents the mean for the ith group (i can take on any whole number value between 1 and k).

For ANOVA, we have three key conditions:

1. Observations are independent within and across groups.

Independence within groups is the way we've been thinking about independence already. We want to convince ourselves that for any particular group, the observations do not impact each other. For independence across groups, we want to convince ourselves that the groups do not impact each other. Note: if we have a simple random sample, this assumption is always satisfied.

2. Data within each group are approximately normal.

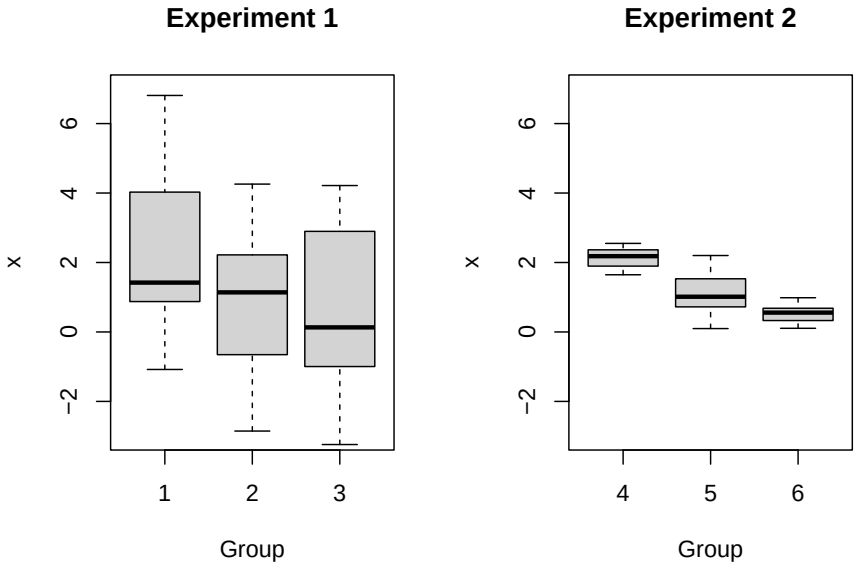
If you make a histogram of the data for each group, each histogram will look approximately bell-shaped.

3. Variability is approximately equal across groups.

Take the standard deviation for each group and check if they are approximately equal. A boxplot is an appropriate way to do this visually.

Why Variance?

You may have seen the name "analysis of variance" and wondered what the variance has to do with comparing many means. Consider the following boxplots:



Is there a difference in the means for Experiment 1? What about Experiment 2?

In fact, the means are $\mu_1 = \mu_4 = 2$, $\mu_2 = \mu_5 = 1$, and $\mu_3 = \mu_6 = 0.5$. But the variances for the Experiment 1 groups are much larger than for the Experiment 2 groups! The larger variances in Experiment 1 obscure any differences between the group means. It is for this reason that we analyze variance as part of our test for differences in means.

> Aside: Why can't we look at the data first and just test the two means that have the largest difference?
>
> When we look at the data *and then choose a test*, this inflates our Type I error rate! It's bad practice and not something we want to engage in as scientists.

In order to perform an ANOVA, we need to consider whether the sample means differ more than we would expect them to based on natural variation (remember that we expect random samples to produce slightly different sample statistics each time!). This type of variation is called **mean square between groups** or MSG. It has associated degrees of freedom $df_G = k - 1$ where k is the number of groups. Note that

$$MSG = \frac{SSG}{df_G}$$

where SSE is the **sum of squares group**. If the null hypothesis is true, variation in the sample means is due to chance. In this case, we would expect the MSG to be relatively small.

When I say "relatively small", I mean we need to compare this quantity to something. We need some quantity that will give us an idea of how much variability to expect if the null hypothesis is true. This is the **mean square error** or MSE, which has degrees of freedom $df_E = n - k$. Again, we have the relationship that

$$MSE = \frac{SSE}{df_E}$$

where SSE is the **sum of squares error**. These calculations are very similar to the calculation for variance (and standard deviation)! (Note: we will not calculate these quantities by hand, but if you are interested in the mathematical details they are available in the OpenIntro Statistics textbook in the footnote on page 289.)

We compare these two quantities by examining their ratio:

$$F = \frac{MSG}{MSE}$$

This is the test statistic for the ANOVA. (See Module 10 for an introduction to the F distribution.)

The ANOVA Table

Generally, when we run an ANOVA, we create an ANOVA table (or we have software create one for us!). This table looks something like this

	df	Sum of Squares	Mean Squares	F Value	P-Value
group	df_G	SSG	MSG	F	p-value
error	df_E	SSE	MSE		

Example: chick weights

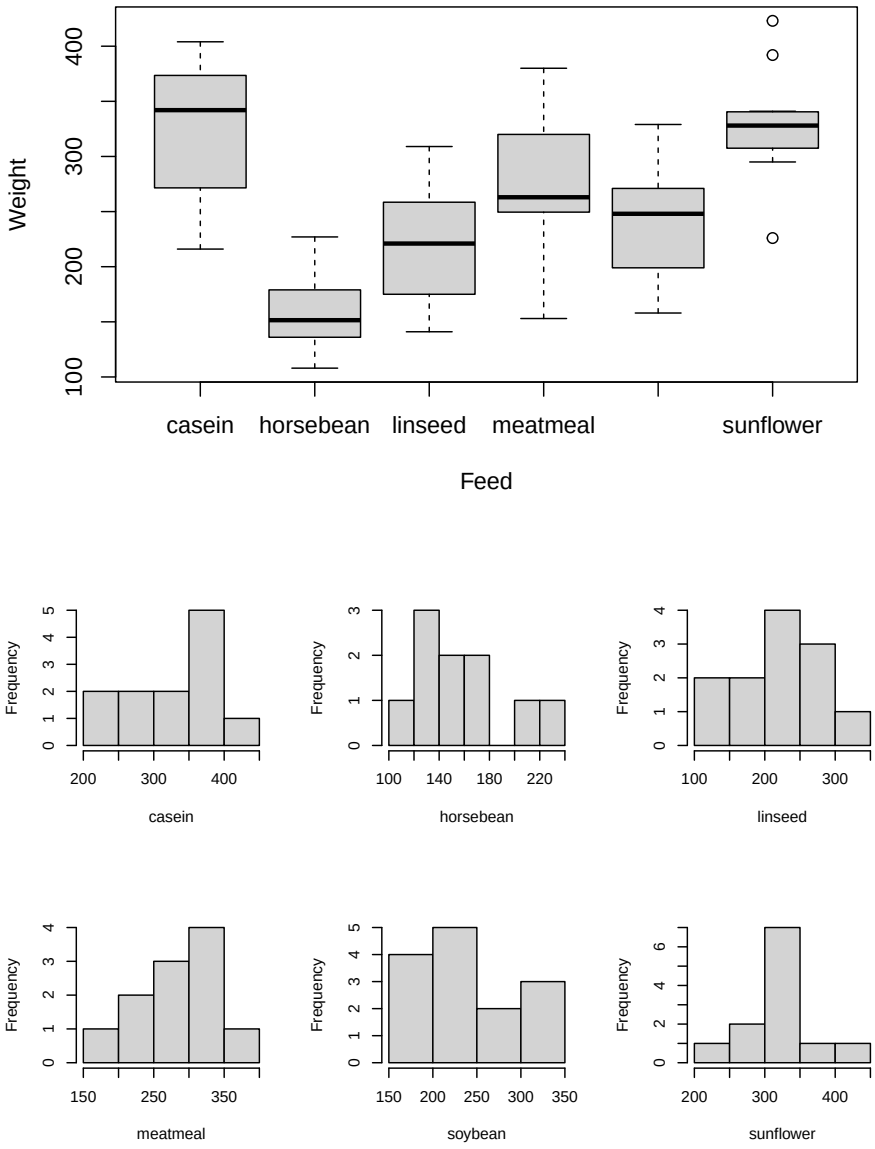
`R` has data on the weights of chicks fed six different feeds (diets). Assume these data are based on a random sample of chicks. There are $n = 71$ total observations and $k = 6$ different feeds. Let's assume we want to test with a 0.05 level of significance.

The ANOVA hypotheses are

- H_0: the mean weight is the same for all six feeds.
- H_A: at least one feed has a mean weight that differs.

The summaries for these data are

```
##         casein horsebean linseed meatmeal soybean sunflower
## n        12.00     10.00   12.00    11.00   14.00     12.00
## Mean    323.58    160.20  218.75   276.91  246.43    328.92
## Std Dev  64.43     38.63   52.24    64.90   54.13     48.84
```



The group sizes are relatively small, so it's difficult to determine how far from normality these data are based on the histograms. We may also run into some issues with constant variance. However, for the sake of the example, let's push ahead with the ANOVA! Since we usually use software to calculate ANOVAs, I've used `R` to create the

following ANOVA table:

```
## Analysis of Variance Table
##
## Response: chickwts$weight
##               Df Sum Sq Mean Sq F value    Pr(>F)
## chickwts$feed  5 231129   46226  15.365 5.936e-10 ***
## Residuals     65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table, we can confirm that $df_G = 6 - 1 = 5$ and $df_E = 71 - 6 = 65$. The F test statistic is

$$MSG/MSE = 46226/3009 = 15.365$$

Finally, the p-value is 5.936×10^{-10}. Clearly $5.936 \times 10^{-10} < \alpha = 0.05$, so we will reject the null hypothesis and conclude that at least one of the feed groups has a mean weight that differs.

12.2 Multiple Comparisons and Type I Error Rate

Let's return for a moment to our ANOVA hypotheses:

- H_0: The mean outcome is the same across all groups.
- H_A: At least one mean differs from the rest.

If we reject H_0 and conclude that "at least one mean differs from the rest", how do we determine which mean(s) differ? *If* we reject H_0, we will perform a series of two-sample t-tests. But wait! What about the Type I error? Isn't this exactly what we decided we couldn't do when we introduced ANOVA?

In order to avoid this increased Type I error rate, we run these **multiple comparisons** with a modified significance level. There are several ways to do this, but the most common way is with the **Bonferroni correction**. Here, if we want to test at the $100(1-\alpha)$ level of significance, we run each of our pairwise comparisons with

$$\alpha^* = \alpha/K$$

where K is the number of comparisons being considered. For k groups, there are

$$K = \frac{k(k-1)}{2}$$

possible pairwise comparisons.

For these comparisons, we use a special pooled estimate of the standard deviation, s_{pooled} in place of s_1 and s_2:

$$\text{standard error} = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}}$$

Other than changing α to α^* and the standard error to this new formula, the test is exactly the same as that discussed in the previous section. Note that

$$s_{\text{pooled}} = \sqrt{MSE}$$

and the degrees of freedom is df_E.

Example: chick weights

Let's extend our discussion on the chick weights to multiple comparisons. Since we were able to conclude that at least one feed has a weight that differs, we want to find out where the difference(s) lie!

We will test all possible pairwise comparisons. This will require $K = \frac{6(6-1)}{2} = 15$ tests. The pooled standard deviation is $s_{pooled} = \sqrt{3009} \approx 54.85$. Let's walk through the test of casein ($\bar{x}_1 = 323.58, n = 12$) vs horsebean ($\bar{x}_2 = 160.20, n = 10$):

- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$

The estimated difference and standard error are

$$\bar{x}_1 - \bar{x}_2 = 323.58 - 160.20 = 163.38 \qquad SE = \sqrt{\frac{54.85^2}{11} + \frac{54.85^2}{9}} = 25.65$$

which results in a test statistic of $t = 6.37$ and a p-value of 1.11×10^{-8}. We then compare this to $\alpha^* = 0.05/15 = 0.0033$. Since the p-value of $1.11 \times 10^{-8} < \alpha^* = 0.0033$, we reject the null hypothesis and conclude there is a significant difference in mean chick weight between the casein and horsebean feeds.

In order to complete the pairwise comparisons, we would then run the remaining 14 tests. I will leave this as an optional exercise for the particularly motivated student.

Note: occasionally, we may reject H_0 in the ANOVA but may fail to find any statistically significant differences when performing multiple comparisons with the Bonferroni correction. This is ok! It just means we were unable to identify which specific groups differ.

Appendices

Appendix A: Important Links and Additional Resources

Applets

- Normal Distribution Calculator
- Rossman and Chance Applets
- Simulating the Central Limit Theorem

Run R Online

- Run R Online
- WebR
- RStudio Cloud

Appendix B: Average Deviance

The deviance of an observation from its mean is $x - \bar{x}$. We denote the deviation for the ith observation as $x_i - \bar{x}$. So the sum over all n deviances is

$$\begin{aligned}
\text{Sum of Deviances} &= \Sigma_{i=1}^{n}(x_i - \bar{x}) && (12.1)\\
&= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_{n-1} - \bar{x}) + (x_n - \bar{x}) && (12.2)\\
&= x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_{n-1} - \bar{x} + x_n - \bar{x} && (12.3)\\
&= x_1 + x_2 + \cdots + x_{n-1} + x_n - \bar{x} - \bar{x} - \cdots - \bar{x} - \bar{x} && (12.4)\\
&= (x_1 + x_2 + \cdots + x_{n-1} + x_n) - (\bar{x} + \bar{x} + \cdots + \bar{x} + \bar{x}) && (12.5)
\end{aligned}$$

where the first half is the sum over all of the x values and the term $(\bar{x})$ appears n times. So we can rewrite this as

$$\text{Sum of Deviances} = \Sigma_{i=1}^{n}(x_i) - n\bar{x}$$

Now notice that, because $\bar{x} = \frac{\Sigma_{i=1}^{n}(x_i)}{n}$, we can multiply $\Sigma_{i=1}^{n}(x_i)$ by $\frac{n}{n}$ to get $\frac{n\Sigma_{i=1}^{n}(x_i)}{n} = n\bar{x}$ and rewrite the sum over the deviances as

$$\begin{aligned} \text{Sum of Deviances} &= n\bar{x} - n\bar{x} && (12.6)\\ &= 0 && (12.7) \end{aligned}$$

Appendix C: Deriving a Confidence Interval

Assume we are taking a sample from a normal distribution with mean μ and standard deviation σ. We will assume the value of σ is known to us. Then $\bar{X}$ is Normal$(\mu, \sigma/\sqrt{n})$. If we standardize $\bar{X}$, we get

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We want some interval (a, b). We will start by considering $a < Z < b$, so $a < Z$ and $Z < b$ (or $b > Z$). Then

$$\begin{aligned} Z &< b \\ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} &< b \\ \bar{X} - \mu &< b\sigma/\sqrt{n} \\ \bar{X} - b\sigma/\sqrt{n} &< \mu \end{aligned}$$

and

$$\begin{aligned} a &< Z \\ a &< \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \\ a\sigma/\sqrt{n} &< \bar{X} - \mu \\ \mu &< \bar{X} - a\sigma/\sqrt{n} \end{aligned}$$

putting these together,

$$\bar{X} - b\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - a\frac{\sigma}{\sqrt{n}}.$$

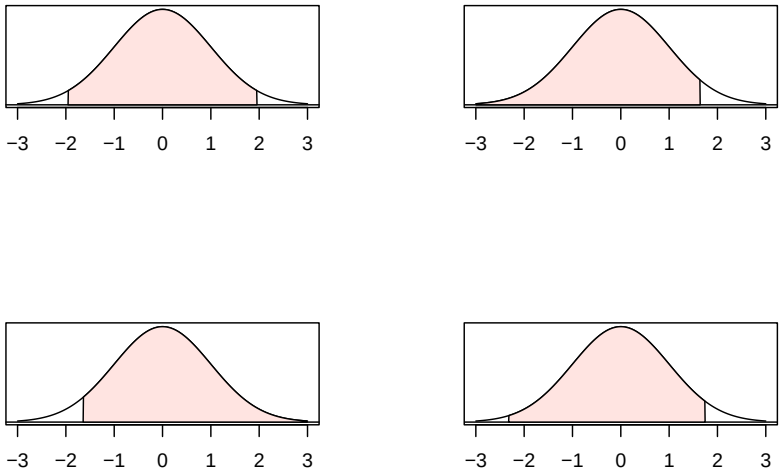
If we want to be 95% confident, then we want $P(a < Z < b) = 0.95$:

$$P\left(\bar{X} - b\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - a\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

To calculate the 95% confidence interval, we need to find a and b such that $P(a < Z < b) = 0.95$.

We want this interval to be as narrow (small) as possible. Why? Narrower intervals are more informative. If I say I'm 95% confident that tomorrow's high will be between -100 and 200 degrees Fahrenheit, that's a useless interval. If I change it to between 70 and 100, that's a little better. Changing it to between 85 and 90 is even better. This is what we mean by more informative.

It turns out that with a symmetric distribution like the normal distribution, the way to make a confidence interval as narrow as possible is to take advantage of this symmetry. Each of the plots below show a shaded area of 0.95. The narrowest interval (along the horizontal axis) is the first interval, which is shaded on $(-1.96 < Z < 1.96)$.



Using the symmetry of the normal distribution, we find that the narrowest interval uses $a = -1.96$ and $b = 1.96$, which results in the 95% confidence interval

$$\left(\bar{x} - z_* \frac{\sigma}{\sqrt{n}}, \bar{x} + z_* \frac{\sigma}{\sqrt{n}}\right)$$

where $z_* = 1.96$.

Works Cited

These works cited include the textbooks I referenced when writing my personal introductory statistics notes, which eventually morphed into this text. They also include all of the R packages used in the writing and publishing of this text.

Textbooks

Diez DM, Barr CD, & Çetinkaya-Rundel M (2019). OpenIntro Statistics (4th ed). OpenIntro. https://www.openintro.org/book/os/

Weiss NA, & Weiss CA (2017). Introductory statistics (10th ed). Pearson.

R Packages

Auguie B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3, https://CRAN.R-project.org/package=gridExtra.

Bryan J (2025). *gapminder: Data from Gapminder*. R package version 1.0.1, https://CRAN.R-project.org/package=gapminder.

Chen H (2022). *VennDiagram: Generate High-Resolution Venn and Euler Plots*. R package version 1.7.3, https://CRAN.R-project.org/package=VennDiagram.

Horst AM, Hill AP, Gorman KB (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi:10.5281/zenodo.3960218.

Neuwirth E (2022). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3, https://CRAN.R-project.org/package=RColorBrewer.

Pebesma E & Bivand R (2023). *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Pebesma E (2018). *Simple Features for R: Standardized Support for Spatial Vector Data*. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Venables WN & Ripley BD (2002) *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

Wickham H, Pedersen T, Seidel D (2023). *scales: Scale Functions for Visualization*. R package version 1.3.0, https://CRAN.R-project.org/package=scales.

Xie Y (2024). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.40, https://github.com/rstudio/bookdown.

Xie Y (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138700109, https://bookdown.org/yihui/bookdown.