STAT 216 Coursepack

Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

This resource was developed by Melinda Yager, Jade Schmidt, and Stacey Hancock in 2021 to accompany the online textbook: Hancock, S., Carnegie, N., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R*. Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/.

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Montana State Introductory Statistics with R, which can be found at https://mtstateintrostats.github.io/IntroStatTextbook/. The syllabus for the course (including the course calendar), data sets, and links to D2L Brightspace, Gradescope, and the MSU RStudio server can be found on the course webpage: https://math.montana.edu/courses/s216/. Other notes and review materials are linked in D2L.

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, video notes are provided to aid in taking notes while you complete the required videos. Bring this workbook with you to class each class period, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

All activities and labs in this coursepack will be completed during class time. Parts of each lab will be turned in on Gradescope. To aid in your understanding, read through the introduction for each activity before attending class each day.

STAT 216 is a 3-credit in-person course. In our experience, it takes six to nine hours per week outside of class to achieve a good grade in this class. By "good" we mean at least a C because a grade of D or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day's class. A typical week in the life of a STAT 216 student looks like:

- *Prior to class meeting*:
  - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
  - Watch the provided videos, taking notes in the coursepack.
  - Read through the introduction to the day's in-class activity.
  - Read through the week's homework assignment and note any questions you may have on the content.
- *During class meeting*:
  - Work through the guided activity, in-class activity or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
  - Complete any parts of the activity you did not complete in class.
  - Review the activity solutions in the Math and Stat Center, and take notes on key points.
  - Complete any remaining assigned readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Basics of Data and Sampling Methods

1.1 Vocabulary Review and Key Topics

At the beginning of each module is a list of new vocabulary terms and key topics for that module. As you read through the material in the text book and watch the videos prior to class, look for these terms. Reference the following definitions to guide your understanding.

1.1.1 Module 1 Vocabulary

- **Data**: observations used to answer research questions
- **Observational units (cases)**: the subjects or entities on which data are collected
  - The rows in a data set represent the observational units
- **Variable**: the characteristics collected on each observational unit
- **Types of variables**:
  - **Categorical**: cases are grouped into categories
  - **Quantitative**: numerical measurements, where performing arithmetic operations makes sense
- **Target population**: group of observational units of interest
- **Sample**: subset of the population
- **Sampling methods**:
  - **Unbiased sampling method (e.g., a random sample)**: on average, the sample will be representative of the target population; all observational units in the target population have the same chance of being selected
  - **Biased sampling method (e.g., convenience sample)**: on average, the sample will not be representative of the target population; some part of the target population will be over- or under-represented
- **Type of sampling bias**:
  - **Selection bias**: method of sampling is biased; some part of the target population is over- or under-represented
  - **Non-response bias**: part of a pre-selected sample does not respond or cannot be reached
  - **Response bias**: responses are not truthful (poor/leading question phrasing, social desirability)
- **Generalization**: to what group of observational units can the results be applied to?
  - If an unbiased method of selection was used and there is no non-response or response bias, we can generalize the results to the target population.
  - If a biased method of selection was used or if non-response or response bias is present, we can only generalize the result to the sample or similar observational units.

1.2 Activity 1: Intro to Data

1.2.1 Learning outcomes

- Creating a data set

1.2.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. This week in class you will be introduced to the following terms:

- Observational units or cases
- Variables: categorical or quantitative

For more on these concepts, read Chapter 1 in the textbook.

1.2.3 General information on the Coursepack

Information is provided throughout each activity and lab to guide students through that day's activity or lab. Be sure to read ALL the material provided at the beginning of the activity and between each question. At the end of each activity is a section called *Take-home messages* that contains key points from the day's activity. Use these to review the day's activity and make sure you have a full understanding of that material.

1.2.4 Steps of the statistical investigation process

As we move through the semester we will work through the six steps of the statistical investigation process.

1. Ask a research question.
2. Design a study and collect data.
3. Summarize and visualize the data.
4. Use statistical analysis methods to draw inferences from the data.
5. Communicate the results and answer the research question.
6. Revisit and look forward.

Today we will focus on the first two steps.

Step 1: The first step of any statistical investigation is to *ask a research question.* As stated in the textbook, "with the rise of data science, however, we might not start with a research question, and instead start with a data set." Today we will create a data set by collecting responses on students in class.

Step 2: To answer any research question, we must *design a study and collect data.* Our study will consist of answers from each student. Your responses will become our observed data that we will explore.

Observational units or **cases** are the subjects data are collected on. In a spreadsheet of the data set, each row will represent a single observational unit.

1. Open the Google Form linked in D2L and fill in the responses for the following questions. When creating a data set for use in R it is important to use single words or an underscore between words. Each outcome must be written the same way each time. Make sure to use all lowercase letters to create this data set to have consistency between responses. Do not give units of measure for numerical values within the data set. For `Residency` use in_state or out_state as the two outcomes.

- Major: what is your declared major?

- Residency: do you have in-state or out-of-state residency?

- Num_Credits: how many credits are you taking this semester?

- Dominant_hand: are you left or right-handed?

- Hand_span: what is the width of your dominant hand from the tip of your thumb to the tip of your pinky with your hand spread out measured in cm?

- Grip_dominant: what is the grip strength measured in lbs for your dominant hand?

- Grip_nondominant: what is the grip strength measured in lbs for your non-dominant hand?

1.2.5 Take-home messages

1. When creating a data set, each row will represent a single observational unit or case. Each column represents a variable collected. It is important to write each variable as a single word or use an underscore between words.
2. Make sure to be consistent with writing each outcome in the data set as R is case sensitive. All outcomes must be written exactly the same way.

1.2.6 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered, and to write down the names and contact information of your teammates.

1.3 Video Notes: Intro to data and Sampling Methods

Read through Sections 1.1 – 1.3 and 2.1 in the course textbook and watch the course videos prior to coming to class. Fill in the following questions to aid in your understanding of the material. Many of the following questions are asked on the video quiz on Gradescope.

1.3.1 Course Videos

- 1.2.1and1.2.2
- 1.2.3to1.2.4
- 2.1

Data basics: Video 1.2.1and1.2.2

Data: ____ used to answer research questions

Observational unit or case: the people or things we ___________________________ data from; represents the ______________ in each data set

Variable: characteristics measured on each ____.

Types of variables

- Categorical variable:

  - Ordinal: levels of the variable have a natural ordering

    Examples: 'Scale' questions, years of schooling completed

  - Nominal:levels of the variable do not have a natural ordering

    Examples: hair color, eye color, zipcode

- Quantitative variable:

  - Continuous variables: value can be any value within a range.

    Examples: percentage of students who are nursing majors

    - average hours of exercise per week
    - distance or time (measured with enough precision)

  - Discrete variables: can only be specific values, with jumps between

    Examples: SAT score

    - number of car accidents

Example: The Bureau of Transportation Statistics ("Bureau of Transportation Statistics" 2019) collects data on all forms of public transportation. The data set seen here includes several variables collect on flights departing on a random sample of 150 US airports in December of 2019.

```
airport <- read.csv("data/airport_delay.csv")
glimpse(airport)
#> Rows: 150
#> Columns: 19
#> $ airport            <chr> "ABI", "ABY", "ACV", "ACY", "ADQ", "AEX", "ALB", "~
#> $ city               <chr> "Abilene", "Albany", "Arcata/Eureka", "Atlantic Ci~
#> $ state              <chr> " TX", " GA", " CA", " NJ", " AK", " LA", " NY", "~
#> $ airport_name       <chr> " Abilene Regional", " Southwest Georgia Regional"~
#> $ hub                <chr> "no", "no", "no", "no", "no", "no", "no", "no", "n~
#> $ international      <chr> "no", "no", "no", "yes", "no", "yes", "yes", "yes"~
#> $ elevation_1000     <dbl> 1.7906, 0.1932, 0.2223, 0.0748, 0.0787, 0.0881, 0.~
#> $ latitude           <dbl> 32.4, 31.5, 41.0, 39.5, 57.7, 31.3, 42.7, 35.2, 45~
#> $ longitude          <dbl> -99.7, -81.2, -124.1, -74.6, -152.5, -92.5, -73.8,~
#> $ arr_flights        <int> 195, 81, 215, 293, 54, 282, 943, 410, 53, 32314, 6~
#> $ perc_delay15       <dbl> 16.410256, 13.580247, 23.255814, 15.358362, 12.962~
#> $ perc_cancelled     <dbl> 0.5128205, 0.0000000, 4.1860465, 0.6825939, 14.814~
#> $ perc_diverted      <dbl> 0.00000000, 0.00000000, 2.32558139, 0.68259386, 0.~
#> $ arr_delay          <int> 1563, 1244, 4763, 2905, 329, 1293, 15127, 9705, 25~
#> $ carrier_delay      <int> 459, 890, 1613, 476, 180, 302, 5627, 2253, 439, 10~
#> $ weather_delay      <int> 21, 43, 549, 124, 1, 58, 2346, 168, 1236, 13331, 2~
#> $ nas_delay          <int> 257, 39, 154, 771, 51, 112, 2096, 616, 746, 45674,~
#> $ security_delay     <int> 0, 0, 0, 25, 0, 0, 44, 0, 0, 375, 0, 83, 0, 23, 0,~
#> $ late_aircraft_delay <int> 826, 272, 2447, 1509, 97, 821, 5014, 6668, 108, 10~
```

- What are the observational units?

- Identify which variables are categorical.

- Identify which variables are quantitative.

Exploratory data analysis (EDA)

Summary statistic: a single number which ________________________________ an entire data set

- Also called the point estimate.

  Examples:

  proportion of people who had a stroke

  mean (or average) age

- The summary statistic and type of plot used depends on the type (categorical or quantitative) of variable(s)!

Roles of variables: 1.2.3to1.2.4

Explanatory variable: predictor variable

- The variable researchers think *may be* ____________________ the other variable.
- In an experiment, what the researchers ____________________ or _________________________.
- The groups that we are comparing from the data set.

Response variable:

- The variable researchers think *may be* ______________________________ by the other variable.
- Always simply _________________________ or ____________________________; never controlled by researchers.

Examples:

Can you predict a criminal's height based on the footprint left at the scene of a crime?

- Identify the explanatory variable:
- Identify the response variable:

Does marking an item on sale (even without changing the price) increase the number of units sold per day, on average?

- Identify the explanatory variable:
- Identify the response variable:

In the Physician's Health Study ("Physician's Health Study," n.d.), male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- Identify the explanatory variable:
- Identify the response variable:

Relationships between variables

- Association: the ____________________ between variables create a pattern; knowing something about one variable tells us about the other.
  - Positive association: as one variable ____________________, the other tends to _______________________ also.
  - Negative association: as one variable ____________________, the other tends to ____________________.
- Independent: no clear pattern can be seen between the ________________.

1.3.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What is the explanatory variable in the Male Physicians study?

2. What is the response variable in the Male Physicians study?

Sampling Methods: Video 2.1

The method used to collect data will impact

- Target population: all ______________________ or ____________________ of interest
- Sample:________________________ or ________________________ from which data is collected

Example: Many high schools moved to partial or fully online schooling in Spring of 2020. Did students who graduated in 2020 tend to have a lower GPA during freshman year of college than the previous class of college freshmen? A nationally representative sample of 1000 college students who were freshmen in AY19-20 and 1000 college students who were freshmen in AY20-21 was taken to answer this question.

- What is the target population?
- What is the sample?

Good vs. bad sampling

GOAL: to have a sample that is ______________________ of the _____________________ ______________________ on the variable(s) of interest

- Unbiased sample methods:

  Simple random sample

- Biased sampling method:

Types of Sampling Bias

- Selection bias:

Example of Selection Bias: Newspaper article from 1936 reported that Landon won the presidential election over Roosevelt based on a poll of 10 million voters. Roosevelt was the actual winner. What was wrong with this poll? Poll was completed using a telephone survey and not all people in 1936 had a telephone. Only a certain subset of the population owned a telephone so this subset was over-represented in the telephone survey. The results of the study, showing that Landon would win, did not represent the target population of all US voters.

- Non-response bias:

- To calculate the non-response rate:

$$\frac{\text{number of people who do not respond}}{\text{total number of people selected for the sample}} \times 100\%$$

- For non-response bias to occur must first select people to participate and then they choose not to.

Example of Non-response bias: A company randomly selects buyers to complete a review of an online purchase but some choose not to respond.

- Response bias:

Example of Response Bias: Police officer pulls you over and asks if you have been drinking. Expect people to say no, whether they have been drinking or not.

- Need to be able to predict how people will respond.

Words of caution:

- Convenience samples: gathering data for those who are easily accessible; online polls

    Selection bias?

    Non-response bias?

    Response bias?

- Random sampling reduces ________________________ bias, but has no impact on _______________________ or _____________________ bias.

Video Example

A radio talk show asks people to phone in their views on whether the United States should pay off its debt to the United Nations.

- Selection?
- Non-response?
- Response?

The Wall Street Journal plans to make a prediction for the US presidential election based on a survey of its readers and plans to follow-up to ensure everyone responds.

- Selection?
- Non-response?
- Response?

A police detective interested in determining the extent of drug use by high school students, randomly selects a sample of high school students and interviews each one about any illegal drug use by the student during the past year.

- Selection?
- Non-response?
- Response?

1.3.3 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What are the two types of variables?
2. Purpose of random selection:
3. Types of sampling bias:

1.4 Activity 2: Intro to Data Analysis and Sampling Bias

1.4.1 Learning outcomes

- Identify observational units, variables, and variable types in a statistical study.
- Creating a data set
- Identify biased sampling methods.

1.4.2 Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. This week in class you will be introduced to the following terms:

- Observational units or cases
- Variables: categorical or quantitative

For more on these concepts, read Chapter 1 in the textbook.

Further analysis of class data set

1. What are the observational units or cases for the data collected in class on day 1?

2. How many observations are reported in the data set? This is the **sample size**.

3. The header for each column in the data set describes each variable measured on the observational unit. For each column of data, fill in the following table identifying the type of each variable, and if the variable is categorical whether the variables is binary and if the variable is quantitative the units of measure used.

Column	Type of Variable	Binary?	Units?
Major			
Residency			
Num Credits			
Dominant hand			
Hand Span			
Grip strength dominant hand			
Grip strength non-dominant hand			

4. Review the completed data set with your table. Remember that when creating a data set for use in R it is important to use single words or an underscore between words. Each outcome must be written the same way each time to have consistency between responses. Do not include units of measure in the data set when reporting numerical values. Write down some issues found with the created class data set.

1.4.3 Sampling Methods

Discuss the following questions with your team.

5. Describe how the students were selected for this study.

6. Can we generalize the results of this study back to all University students? All MSU students? All Stat 216 students?

7. Explain your answer to question 6.

Types of bias

8. To determine if the proportion of out-of-state undergraduate students at Montana State University has increased in the last 10 years, a statistics instructor sent an email survey to 500 randomly selected current undergraduate students. One of the questions on the survey asked whether they had in-state or out-of-state residency. She only received 378 responses.

   Sample size:

   Observational units sampled:

   Target population:

   Justify why there is non-response bias in this study.

   Variables measured and their types:

9. A television station is interested in predicting whether or not local voters will pass a referendum to legalize marijuana for adult. The TV station asks its viewers to phone in and indicate whether they are in favor or opposed to the referendum. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

   Sample size:

   Observational units sampled:

   Target population:

   Justify why there is selection bias in this study.

   Variables measured and their types:

10. To gauge the interest of Bozeman City Voters in a new swimming pool, a local organization stood outside of the Bogart Pool in Bozeman, MT, during open hours. One of the questions they asked was, “Since the Bogart Pool is in such bad repair, don’t you agree that the city should fund a new pool?”

    Sample size:

    Observational units sampled:

    Target population:

    Justify why there is response bias in this study.

    Justify why there is selection bias in this study.

    Variables measured and their types:

1.4.4 Take-home messages

1. There are two types of variables: categorical (groups) and quantitative (numerical measures).
2. We will learn more about summarizing variable later in the semester. Categorical variables are summarized by calculating a proportion from the data and quantitative variables are summarized by finding the mean and the standard deviation.
3. There are three types of bias to be aware of when designing a sampling method: selection bias, non-response bias, and response bias.

1.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today’s activity and material covered, and to write down the names and contact information of your teammates.

1.5 Activity 3: American Indian Address

1.5.1 Learning outcomes

- Explain why a sampling method is unbiased or biased.
- Identify biased sampling methods.
- Explain the purpose of random selection and its effect on scope of inference.

1.5.2 Terminology review

In this activity, we will examine unbiased and biased methods of sampling. Some terms covered in this activity are:

- Random sample
- Unbiased vs biased methods of selection
- Generalization

To review these concepts, see Chapter 2 in the textbook.

1.5.3 Class Preparation

Prior to the next class, complete questions 1–3.

1.5.4 American Indian Address

For this activity, you will read a speech given by Jim Becenti, a member of the Navajo American Indian tribe, who spoke about the employment problems his people faced at an Office of Indian Affairs meeting in Phoenix, Arizona, on January 30, 1947 (Moquin and Van Doren 1973). His speech is below:

It is hard for us to go outside the reservation where we meet strangers. I have been off the reservation ever since I was sixteen. Today I am sorry I quit the Santa Fe [Railroad]. I worked for them in 1912–13. You are enjoying life, liberty, and happiness on the soil the American Indian had, so it is your responsibility to give us a hand, brother. Take us out of distress. I have never been to vocational school. I have very little education. I look at the white man who is a skilled laborer. When I was a young man I worked for a man in Gallup as a carpenter's helper. He treated me as his own brother. I used his tools. Then he took his tools and gave me a list of tools I should buy and I started carpentering just from what I had seen. We have no alphabetical language.

We see things with our eyes and can always remember it. I urge that we help my people to progress in skilled labor as well as common labor. The hope of my people is to change our ways and means in certain directions, so they can help you someday as taxpayers. If not, as you are going now, you will be burdened the rest of your life. The hope of my people is that you will continue to help so that we will be all over the United States and have a hand with you, and give us a brotherly hand so we will be happy as you are. Our reservation is awful small. We did not know the capacity of the range until the white man come and say "you raise too much sheep, got to go somewhere else," resulting in reduction to a skeleton where the Indians can't make a living on it. For eighty years we have been confused by the general public, and what is the condition of the Navajo today? Starvation! We are starving for education. Education is the main thing and the only thing that is going to make us able to compete with you great men here talking to us.

By eye selection

1. Circle ten words in Jim Becenti's speech which are a representative sample of the length of words in the entire text. Describe your method for selecting this sample.

2. Fill in the table below with your selected words from the previous question and the length of each word (number of letters/digits in the word):

Observation	Word	Length
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

3. Calculate the mean (average) word length in your selected sample. Is this value a parameter or a statistic?

1.5.5 Class Activity

1. Report your mean word length in the Google sheet. Your instructor will create a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive x-axis label. Report the mean of the sample mean word lengths.

2. Calculate how far your sample mean from Q1 is from the mean of the sample mean word lengths. Report this difference in the Google sheet. Your instructor will show you how to calculate the standard deviation.

Interpret the standard deviation of the statistics in context of the problem.

The plot created in the question 1 is a sampling distribution of statistics. This sampling distribution plots the mean word length from many samples taken from the population of words.

3. Based on the plot and summary statistics of the sample mean word lengths, what is your best guess for the average word length of the population of all 359 words in the speech?

4. The true mean word length of the population of all 359 words in the speech is 3.95 letters. Is this value a parameter or a statistic?

   Where does the value of 3.95 fall in the plot given? Near the center of the distribution? In the tails of the distribution?

5. If the class samples were truly representative of the population of words, what proportion of sample means in the sampling distribution would you expect to be below 3.95?

6. Using the graph in Q1, estimate the proportion of students' computed sample means that were lower than the true mean of 3.95 letters?

7. Based on your answers to questions 5 and 6, would you say the sampling method used by the class is biased or unbiased? Justify your answer.

8. If the sampling method is biased, what type of sampling bias (selection, response, non-response) is present? What is the direction of the bias, i.e., does the method tend to overestimate or underestimate the population mean word length?

9. Should we use results from our "by eye" samples to make a statement about the word length in the population of words in Becenti's address? Why or why not?

Random selection

Suppose instead of attempting to select a representative sample by eye (which did not work), each student used a random number generator to select a simple random sample of 10 words. A **simple random sample** relies on a random mechanism to choose a sample, without replacement, from the population, such that every sample of size 10 is equally likely to be chosen.

To use a random number generator to select a simple random sample, you first need a numbered list of all the words in the population, called a **sampling frame**. You can then generate 10 random numbers from the numbers 1 to 359 (the number of words in the population), and the chosen random numbers correspond to the chosen words in your sample.

10. Use the random number generator at https://istats.shinyapps.io/RandomNumbers/ to select a simple random sample from the population of all 359 words in the speech.

- Set "Choose Minimum" to 1 and "Choose Maximum" to 359 to represent the 359 words in the population (the sampling frame).
- Set "How many numbers do you want to generate?" to 10 and ensure the "No" option is selected under "Sample with Replacement?"
- Click "Generate".

Fill in the table below with the random numbers selected and use the Becenti.csv data file found on D2L to determine each number's corresponding word and word length (number of letters/digits in the word):

Observation	Number	Length
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

11. Calculate the mean word length in your selected sample in question 10. Is this value a parameter or a statistic?

12. Report your mean word length in the Google sheet. Your instructor will create a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive x-axis label. Report the mean and standard deviation of the data.

13. Where does the value 3.95, the true mean word length, fall in the distribution given? Near the center of the distribution? In the tails of the distribution? Circle this value on the provided distribution.

14. How does the plot from Q10 compare to the plot generated in Q1?

    Is the shape similar?

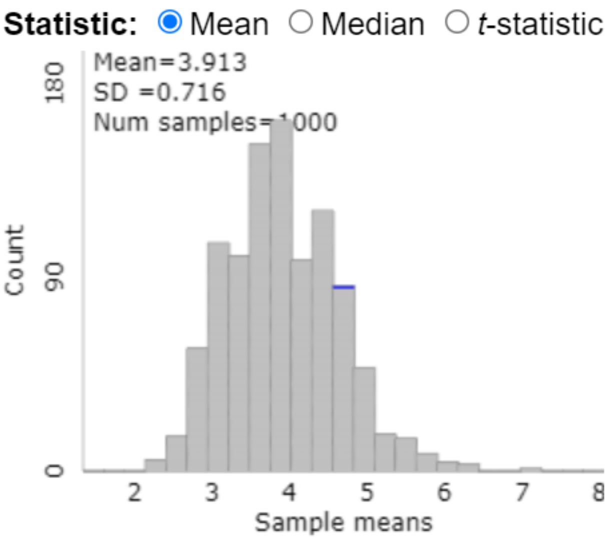
    Is the range (smallest to largest values) similar?

    Is the mean of the distribution similar?

    Why didn't everyone get the same sample mean?

One set of randomly generated sample mean word lengths from a single class may not be large enough to visualize the distribution results. Let's have a computer generate 1,000 sample mean word lengths for us.

The following plot illustrates a sampling distribution of 1000 samples of size 10 selected at random from the sample.



Statistic: ● Mean ○ Median ○ *t*-statistic

Mean=3.913
SD =0.716
Num samples=1000

15. What is the center value (mean) of the distribution displayed above?

16. Explain why the sampling method of using a random number generator to generate a sample is a "better" method than choosing 10 words "by eye".

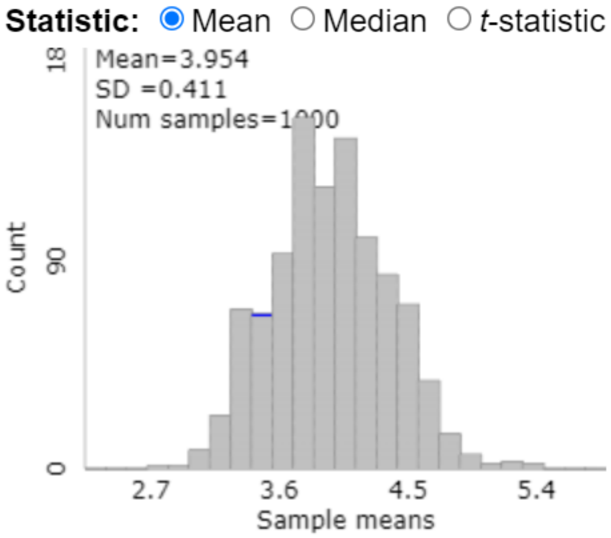
17. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from before Q15.

Effect of sample size

We will now consider the impact of sample size.

18. First, consider if each student had selected 30 words, instead of 10, by eye. Do you think this would make the plot from the previous activity centered on 3.95 (the true mean word length)? Explain your answer.

Now we will select 30 words instead of 10 words at random. The following plot illustrates a sampling distribution of 1000 samples of size 30 selected at random from the sample.



Statistic: ● Mean ○ Median ○ *t*-statistic

19. Compare the distribution displayed before question 15 to the one shown above.

    Is the shape similar?

    Is the range (smallest to largest values) similar?

    Is the mean of the distribution similar?

20. Compare the values of the standard deviation of the plots before question 15 and before question 19. Which plot shows the smallest standard deviation?

21. Using the evidence from your simulations, answer the following research questions:

    Does changing the sample size impact whether the sample estimates are unbiased? Explain your answer.

    Does changing the sample size impact the variability (spread) of sample estimates? Explain your answer

22. What is the purpose of random selection of a sample from the population?

1.5.6 Take-home messages

1. When we use a biased method of selection, we will over or underestimate the parameter.
2. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.
3. Random selection is an unbiased method of selection.
4. To determine if a sampling method is biased or unbiased, we compare the distribution of the estimates to the true value. We want our estimate to be on target or unbiased. When using unbiased methods of selection, the mean of the distribution matches or is very similar to our true parameter.
5. Random selection eliminates selection bias. However, random selection will not eliminate response or non-response bias.
6. The larger the sample size, the more similar (less variable) the statistics will be from different samples.
7. Sample size has no impact on whether a *sampling method* is biased or not. Taking a larger sample using a biased method will still result in a sample that is not representative of the population.

1.5.7 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

MODULE 2

Probability

2.1 Vocabulary Review and Key Topics

- **Probability** (of an event): the long-run proportion of times the event would occur if the random process were repeated indefinitely (under identical conditions)
- **Conditional probability** (of an event *given* another event): probability of an event calculated dependent on another event having occurred
- **Probability notation**:
  - $P(A)$: the probability of event A
    - This is the probability of a single event, *unconditional* probability calculated out of the overall population
  - $P(A^C)$: the probability of the **complement** of event A, or “A complement”
    - This is the probability of the opposite of event A, or “not A”
    - $P(A^C) = 1 - P(A)$
  - $P(A \text{ and } B)$: the probability of event A and B
    - The is the probability of an “and” event, *unconditional* probability calculated out of the overall population
  - $P(A|B)$: the probability of event A given (conditional on) event B
    - This is a *conditional* probability calculated out of the total population for which event B occurred

2.2 Video Notes: Probability

Read Chapters 22 in the course textbook. Use the following videos to complete the video notes for Module 12.

2.2.1 Course Videos

- Chapter23

Probability

Example: Two variables were collected on a random sample of people who had ever been married; whether a person had ever smoked and whether a person had ever been divorced. The data are displayed in the following table. This survey was based on a random sample in the United States in the early 1990s, so the data should be representative of the adult population who had ever been married at that time.

- Let event D be a person has gone through a divorce
- Let event S be a person smokes

	Has divorced	Has never divorced	Total
Smokes	238	247	485
Does not smoke	374	810	1184
Total	612	1057	1669

- What is the approximate probability that the person smoked?
- What is the approximate probability that the person had ever been divorced?
- Given that the person had been divorced, what is the probability that he or she smoked?
- Given that the person smoked, what is the probability that he or she had been divorced?
- Event: something that could occur, something we want to find the probability of
  - Getting a four when rolling a fair die
- Complement: opposite of the event
  - Getting any value but a four when rolling a fair die

- The probability of an event is the ____________________ proportion of times the event would occur if the _______________________ process were repeated indefinitely.
  - For example, the probability of getting a four when rolling a fair die is ________________.
- Unconditional probabilities
  - An _____________________________probability is calculated from the entire population not____ on the occurrence of another event.
  - Examples:
    - The probability of a single event
      - The probability a selected Stat 216 student is a computer science major.
    - An “And” probability
      - The probability a selected Stat 216 student is a computer science major and a freshman.
- Conditional probabilities
  - A ______________________________ probability is calculated _________________________________ on the occurrence of another event.
  - Examples:
    - The probability of event A given B
      - The probability a selected freshman Stat 216 student is a computer science major.
    - The probability of event B given A
      - The probability a selected computer science Stat 216 student is a freshman
- Let event D be a person has gone through a divorce
- Let event S be a person smokes

	Has divorced	Has never divorced	Total
Smokes	238	247	485
Does not smoke	374	810	1184
Total	612	1057	1669

Calculate and interpret each of the following:

- $P(S^C) =$

- $P(D^C|S^C) =$

Creating a hypothetical two-way table

Steps:

- Start with a large number like 100000.
- Then use the unconditional probabilities to fill in the row or column totals.
- Now use the conditional probabilities to begin filling in the interior cells.
- Use subtraction to find the remaining interior cells.
- Add the column values together for each row to find the row totals.
- Add the row values together for each column to find the column totals.

Example: An airline has noticed that 30% of passengers pre-pay for checked bags at the time the ticket is purchased. The no-show rate among customers that pre-pay for checked bags is 5%, compared to 15% among customers that do not pre-pay for checked bags.

- Let event B = customer pre-pays for checked bag
- Let event N = customer no shows

Start by identifying the probability notation for each value given.

- 0.30 =
- 0.05 =
- 0.15 =

	B	B^C	Total
N			
N^C			
Total			100,000

- What is the probability that a randomly selected customer who shows for the flight, pre-purchased checked bags?

Diagnostic tests

- Sensitivity:
- Specificity:
- Prevalence:

2.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. Calculate and interpret the following: $P(D^C|S^C) =$.
2. What is the probability notation for 0.15 in the airline example?

2.3 Activity 4: Probability Studies

2.3.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Construct two-way tables to evaluate conditional probabilities.

2.3.2 Terminology review

In today's activity, we will cover two-way tables and probability. Some terms covered in this activity are:

- Proportions
- Probability
- Conditional probability
- Two-way tables

To review these concepts, see Chapter 23 in the textbook.

2.3.3 Overview of probabiliy

The probability of an event is the long-run proportion of times the event would occur if the random process were repeated indefinitely (under identical conditions).

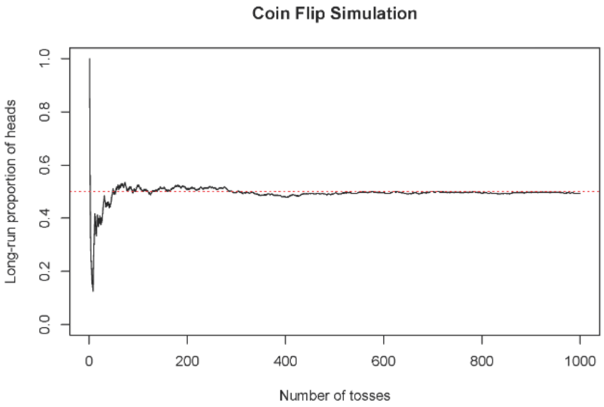
To calculate the probability of an event happening:

$$\text{probability} = \frac{\text{number of ways an event can happen}}{\text{total number of possible outcomes}}$$

For example, to calculate the probability of a coin flip landing on heads; there are only two outcomes (heads or tails) and only one possibility way to land on heads.

$$P(heads) = \frac{1}{2} = 0.5$$

The figure below shows the long-run proportion of times a simulated coin flip lands on heads on the y-axis, and the number of tosses on the x-axis. Notice how the long-run proportion starts converging to 0.5 as the number of tosses increases.



In today's activity we will discuss the probability of a single event, the probability of an "and" event, and the probability of a conditional event.

Probability notation

We will use the notation P(event) to represent the probability of an event and use letters to represent events. The following are notations for different probabilities where we are discussing event A and event B:

- $P(A)$ represents the probability of event A
- $P(A^C)$ represents the probability of the complement of event A
  - $P(A^C) = 1 - P(A)$
- $P(AandB)$ represents the probability of events A and B
- $P(A|B)$ represents the probability of event A given event B
- $P(B|A)$ represents the probability of event B given event A

2.3.3.1 Probability questions

For the beginning of this activity we will start with discussing the probabilities associated with drawing a card from a standard card deck. In a card deck there are:

- 52 cards
- Half are red, half are black
- Four suits: spades, hearts, diamonds, and clubs
- Each suit has 13 cards: cards 2$-$10, ace, jack, queen, and king
- Let A represent the event that a card is an ace
- Let B represent the event that a card is red

To find the probability of selecting an ace, first start with determining how many aces are possible (four) and how many cards will we select from (total of 52).

Find the probability of selecting a card that is not an ace. This is the complement of event A.

Find the probability of selecting a red ace. There are only two red aces and a total of 52 cards.

Find the probability of selecting an ace given that the card is red. There are two red aces but only $\frac{52}{2} = 26$ red cards

If a card drawn is an ace, what is the probability the card drawn is red. There are four aces but only two that are red.

2.3.4 Calculating probabilities from a two-way table

1. In 2014, the website FiveThirtyEight examined the works of Bob Ross to see what trends could be found. They determined that of all the paintings he created, 95% of them contained at least one "happy tree." Of those works with a happy tree, 43% contained at least one "almighty mountain." Of the paintings that did not have at least one happy tree, only 10% contained at least one almighty mountain.

Let A = Bob Ross painting contains a happy tree, and B = Bob Ross painting contains an almighty mountain

	A	A^C	Total
B	40850	500	41350
B^C	54150	4500	58650
Total	95000	5000	100000

a. What is the probability that a randomly selected Bob Ross painting contains both a "happy tree" and an "almighty mountain"? Use appropriate probability notation.

b. What is the probability that a selected Bob Ross painting without an "almighty mountain" contains a "happy tree." Use appropriate probability notation.

c. What is the probability that a selected Bob Ross painting does not contain a "happy tree" given it does not contain an "almighty mountain". Use appropriate probability notation.

2. A recent study of population decline of white-tailed deer in Wyoming due to chronic wasting disease (Edmunds 2016) (CWD) reported the prevalence of CWD to be 35.4%. The survival rate of CWD positive deer was 39.6% and the survival rate of CWD negative deer was 80.1%.

   Let A = the event a deer has CWD, and B = the event the deer survived.

   a. Identify what each numerical value given in the problem represents in probability notation.

      0.354 =

      0.396 =

      0.801 =

   b. Create a hypothetical two-way table to represent the situation.

	A	A^C	Total
B			
B^C			
Total			100,000

c. Find $P(A \text{ and } B)$. What does this probability represent in the context of the problem?

d. Find the probability that a deer that has CWD does not survive. What is the notation used for this probability?

e. What is the probability that a deer does not survive given they do not have CWD? What is the notation used for this probability?

2.3.5 Take home messages

1. Conditional probabilities are calculated dependent on a second variable. In probability notation, the variable following | is the variable on which we are conditioning. The denominator used to calculate the probability will be the total for the variable on which we are conditioning.

2. When creating a two-way table we typically want to put the explanatory variable on the columns of the table and the response variable on the rows.

3. To fill in the two-way table, always start with the unconditional variable in the total row or column and then use the conditional probabilities to fill in the interior cells.

2.3.6 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

2.4 Activity 5: What's the probability?

2.4.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Construct two-way tables to evaluate conditional probabilities.

2.4.2 Terminology review

In today's activity, we will cover two-way tables and probability. Some terms covered in this activity are:

- Proportions
- Probability
- Conditional probability
- Two-way tables

To review these concepts, see Chapter 23 in the textbook.

2.4.3 Probability

1. A dataset was collected on all NBA basketball players from inception of the league. The probability that an NBA player is above average height is 59.7%. Of NBA players that are above average height, 46.4% averaged at least four rebounds a game. The probability that an NBA player averages less than four rebounds in a game given they are below average height is 13.3%.

   Let A = player is above average height, and B = player averages at least four rebounds a game.

	A	A^C	Total
B	27700.8	34940.1	62640.9
B^C	31999.2	5359.9	37359.1
Total	59700	40300	100000

a. What is the probability that a randomly selected NBA player averages at least 4 rebounds a game? Use appropriate probability notation.

b. What is the probability that a randomly selected NBA player is both above average height and averages at least 4 rebounds a game. Use appropriate probability notation.

c. What is the probability that a randomly selected NBA player is not above average height given they do not average at least 4 rebounds a game. Use appropriate probability notation.

2. Since the early 1980s, the rapid antigen detection test (RADT) of group A *streptococci* has been used to detect strep throat. A recent study of the accuracy of this test shows that the **sensitivity**, the probability of a positive RADT given the person has strep throat, is 86% in children, while the **specificity**, the probability of a negative RADT given the person does not have strep throat, is 92% in children. The **prevalence**, the probability of having group A strep, is 37% in children. (Stewart et al. 2014)

   Let A = the event the child has strep throat, and B = the event the child has a positive RADT.

   a. Identify what each numerical value given in the problem represents in probability notation.

   0.86 =

   0.92 =

   0.37 =

   b. Create a hypothetical two-way table to represent the situation.

	A	A^C	Total
B			
B^C			
Total			100,000

c. Find $P(A \text{ and } B)$. What does this probability represent in the context of the problem?

d. Find the probability that a child with a positive RADT actually has strep throat. What is the notation used for this probability?

e. What is the probability that a child does not have strep given that they have a positive RADT? What is the notation used for this probability?

2.4.4 Take home messages

1. Conditional probabilities are calculated dependent on a second variable. In probability notation, the variable following | is the variable on which we are conditioning. The denominator used to calculate the probability will be the total for the variable on which we are conditioning.
2. When creating a two-way table we typically want to put the explanatory variable on the columns of the table and the response variable on the rows.
3. To fill in the two-way table, always start with the unconditional variable in the total row or column and then use the conditional probabilities to fill in the interior cells.

2.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

MODULE 3

Exploring Categorical Data: Exploratory Data Analysis and Inference using Simulation-based Methods

3.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a single categorical variable.

- **Summary statistic (point estimate)**: the value of a numerical summary measure computed from *sample* data
  - Summary measures covered in STAT 216 include: single proportion, difference in proportions, single mean, paired mean difference, difference in means, correlation, and slope of a regression line
  - For a single categorical variable, a proportion is calculated
  - To interpret in context include:
    - Summary measure (in context)
    - Value of the statistic
- **Parameter of interest**: a numerical summary measure of the entire *population* in which we are interested
  - The value of the parameter of interest is unknown (unless we have access to the entire population)
  - To write in context:
    - Population word (true, long-run, population)
    - Summary measure (depends on the type of data)
    - Context
      - Observational units
      - Variable(s)
- **Frequency bar plot**: plots the count (frequency) of observational units in each level of a categorical variable
- **Relative frequency bar plot**: plots the proportion (relative frequency) of observational units in each level of a categorical variable
- **Hypothesis testing**: a formal statistical technique for evaluating two competing possibilities about a population: the null hypothesis and alternative hypothesis
  - When we observe an effect in a sample, we would like to determine if this observed effect represents an actual effect in the population, or whether it was simply due to random chance.
  - A hypothesis test helps us answer the following question about the population: How strong is the *evidence* of an effect?
- **Null hypothesis**: typically represents a statement of “no difference”, “no effect”, or the status quo

  - The null hypothesis is what we assume is true when calculating the p-value. Thus, we can never have evidence *for* the null hypothesis—we cannot “accept” a null hypothesis—we can only find evidence *against* the null hypothesis if the observed data is very unlikely to have occurred under the assumption that the null hypothesis is true.

- **Alternative hypothesis**: represents an alternative claim under consideration and is often represented by a range of possible values for the parameter of interest.
  - The alternative hypothesis is determined by the research question.

Hypotheses:

$$H_0 : \pi = \pi_0$$

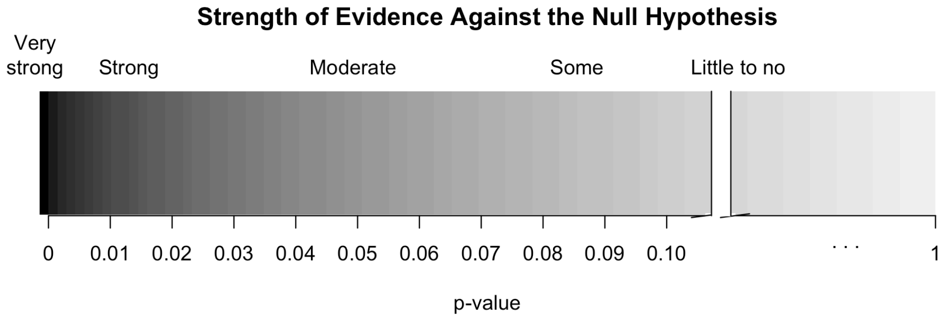
$$H_A : \pi \begin{Bmatrix} < \\ \neq \\ < \end{Bmatrix} \pi_0$$

- **Null Distribution**: a distribution of simulated sample statistics created under the assumption that the null hypothesis is true
- **Simulation methods to create the null distribution**: a process of using a computer program (e.g., R) to simulate many samples that we would expect based on the null hypothesis.

  R code to use simulation methods for one categorical variable to find the p-value, `one_proportion_test`, is shown below.

```
one_proportion_test(probability_success = xx, # Null hypothesis value
      sample_size = xx, # Enter sample size
      number_repetitions = 1000, # Enter number of simulations
      as_extreme_as = xx, # Observed statistic
      direction = "xx", # Specify direction of alternative hypothesis
      summary_measure = "proportion") # Reporting proportion or number of successes?
```

- **P-value**: the probability of the value of the observed sample statistic or a value more extreme, if the null hypothesis were true
  - To write in context include:
    - Statement about probability or proportion of samples
    - Statistic (summary measure and value)
    - Direction of the alternative
    - Null hypothesis (in context)
- **Strength of evidence**: the p-value indicates the amount of evidence there is against the null hypothesis. The smaller the p-value the more evidence there is against the null hypothesis.



- **Conclusion** (to a hypothesis test): answers the research question. How much evidence is there in support of the alternative hypothesis?
  - To write in context include:
    - Amount of evidence
    - Parameter of interest
    - Direction of the alternative hypothesis
- **Confidence interval**: an interval estimate for the parameter of interest
  - A confidence interval helps us answer the following question about the population: How *large* is the effect?
  - To write in context include:
    - How confident you are (e.g., 90%, 95%, 98%, 99%)
    - Parameter of interest
    - Calculated interval
- **Bootstrapping**: creating a simulated sample of the same size as the original sample by sampling with replacement from the original sample
- **Simulation methods to create the bootstrap distribution**: a process of using a computer program to simulate many bootstrapped samples.

  R code to use simulation methods for one categorical variable to find a confidence interval, `one_proportion_bootstrap_CI`, is shown below.

```
one_proportion_bootstrap_CI(sample_size = xx, # Sample size
                number_successes = xx, # Observed number of successes
                number_repetitions = 1000, # Number of bootstrap samples to use
                confidence_level = 0.95) # Confidence level as a decimal
```

- **Percentile method**: process to find the confidence interval from the bootstrap distribution
  - A 90% confidence interval will be found between the 5th and 95th percentiles
  - A 95% confidence interval will be found between the 2.5th and 97.5th percentiles
  - A 99% confidence interval will be found between the 0.5th and 99.5th percentiles

3.1.1 Key topics

Exploratory data analysis

At the end of this module, you should understand how to calculate a summary statistic and plot a single categorical variable.

- Notation for a sample proportion: $\hat{p}$
- Notation for a population proportion: π
- Types of plots for a single categorical variable:
  - Frequency bar plot
  - Relative frequency bar plot

Inference

Additionally, we will use simulation methods **to find evidence of an effect by finding a p-value** and **estimating how large the effect is by creating a confidence interval**.

This is steps 4 and 5 from the steps of the statistical investigation process.

Steps of the statistical investigation process

As we move through the semester we will work through the six steps of the statistical investigation process.

1. Ask a research question.
2. Design a study and collect data.
3. Summarize and visualize the data.
4. Use statistical analysis methods to draw inferences from the data.
5. Communicate the results and answer the research question.
6. Revisit and look forward.

3.2 Video Notes: Exploratory Data Analysis of Categorical Variables

Read Chapter 3, 4, 9, 10 and Sections 14.1 and 14.2 in the course textbook. Use the following videos to complete the video notes for Module 4.

3.2.1 Course Videos

- 4.1
- 4.2
- Chapter9
- 14.1
- Chapter10
- 14.2

Summarizing categorical data - Video 4.1

- A ____________________ is calculated on data from a sample
- The parameter of interest is what we want to know from the population.
- Includes:
  - Population word (true, long-run, population)
  - Summary measure (depends on the type of data)
  - Context
    - Observational units
    - Variable(s)

Categorical data can be numerically summarized by calculating a ______________________ from the data set.

Notation used for the population proportion:

- Single categorical variable:

- Two categorical variables:

  - Subscripts represent the ________________ variable groups

Notation used for the sample proportion:

- Single categorical variable:

- Two categorical variables

Categorical data can be reported in a _________________table, which plots counts or a ____________________ frequency table, which plots the proportion.

When we have two categorical variables we report the data in a _____________________ or two-way table with the ______________________ variable on the columns and the _________________ variable on the rows.

Example from the Video: Gallatin Valley is the fastest growing county in Montana. You'll often hear Bozeman residents complaining about the 'out-of-staters' moving in. A local real estate agent recorded data on a random sample of 100 home sales over the last year at her company and noted where the buyers were moving from as well as the age of the person or average age of a couple buying a home. The variable age was binned into two categories, "Under30" and "Over30." Additionally, the variable, state the buyers were moving from, was created as a binary variable, "Out" for a location out of state and "In" for a location in state.

The following code reads in the data set, `moving_to_mt` and names the object moving.

```
moving <- read.csv("data/moving_to_mt.csv")
```

The `R` function `glimpse` was used to give the following output.

```
glimpse(moving)
```

```
#> Rows: 100
#> Columns: 4
#> $ From      <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", ~
#> $ Age_Group <chr> "Under30", "Under30", "Under30", "Under30", "Under30", "Unde~
#> $ Age       <int> 25, 26, 27, 27, 29, 29, 35, 37, 49, 63, 65, 77, 22, 24, 24, ~
#> $ InOut     <chr> "Out", "Out", "Out", "Out", "Out", "Out", "Out", "Out", "Out~
```

- What are the observational units in this study?

- What type of variable is `Age`?

- What type of variable is `Age_Group`?

To further analyze the categorical variable, `From`, we can create either a frequency table:

```
#>   From  n
#> 1   CA 12
#> 2   CO  8
#> 3   MT 61
#> 4   WA 19
```

Or a relative frequency table:

```
#>   From  n freq
#> 1   CA 12 0.12
#> 2   CO  8 0.08
#> 3   MT 61 0.61
#> 4   WA 19 0.19
```

- How many home sales have buyers from WA?

- What proportion of sampled home sales have buyers from WA?

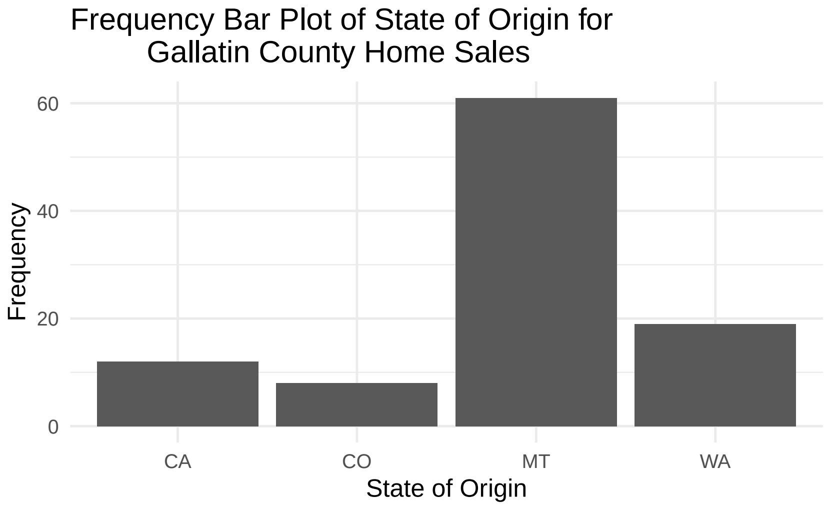
- What notation is used for the proportion of home sale buyers that that are from WA?

Displaying categorical variables - Video 4.2

- Types of plots for a single categorical variable

The following code in R will create a frequency bar plot of the variable, `From`.

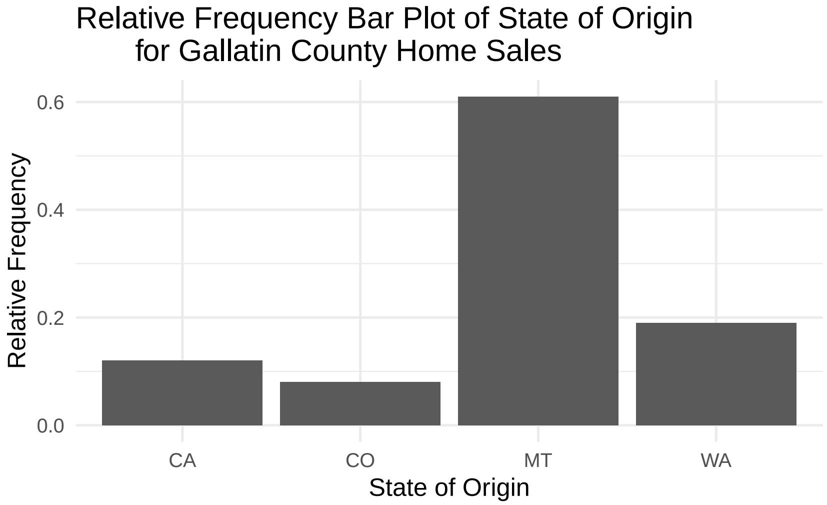
```
moving %>%
    ggplot(aes(x = From))+ #Enter the variable to plot
    geom_bar(stat = "count") +
    labs(title = "Frequency Bar Plot of State of Origin for
         Gallatin County Home Sales",
         #Title your plot (type of plot, observational units, variable)
       y = "Frequency", #y-axis label
       x = "State of Origin") #x-axis label
```



- What can we see from this plot?

Additionally, we can create a relative frequency bar plot.

```
moving %>%
  ggplot(aes(x = From))+ #Enter the variable to plot
  geom_bar(aes(y = after_stat(prop), group = 1)) +
  labs(title = "Relative Frequency Bar Plot of State of Origin
       for Gallatin County Home Sales",
       #Title your plot
       y = "Relative Frequency", #y-axis label
       x = "State of Origin") #x-axis label
```



- Note: the x-axis is the ______________________ between the frequency bar plot and the relative frequency bar plot. However, the _____________________ differs. The scale for the frequency bar plot goes from ____ and the scale for the relative frequency bar plot is from ____.

Hypothesis Testing - Video Chapter9

Purpose of a hypothesis test:

- Use data collected on a sample to give information about the population.
- Determines ___________________________ of _______________________________ of an effect

General steps of a hypothesis test

1. Write a research question and hypotheses.
2. Collect data and calculate a summary statistic.
3. Model a sampling distribution which assumes the null hypothesis is true.
4. Calculate a p-value.
5. Draw conclusions based on a p-value.

Hypothesis Testing/Justice System

- Two possible outcomes if the observed statistic is unusual:
  - Strong evidence against ___________________________ -> ______________________________
  - Not enough evidence against _______________________________ -> _________________________________
- Always written about the ___________________________ (population)

Null hypothesis

- Skeptical perspective, no difference, no effect, random chance

- What the researcher hopes is ________________________.

Notation:

Alternative hypothesis

- New perspective, a chance, a difference, an effect
- What the researcher hopes is __________________________.

Notation:

Simulation vs. Theory-based Methods

Simulation-based method

Creation of the null distribution

- Simulate many samples assuming

- Find the proportion of ______________________________ at least as extreme as the observed sample ___________________

- The null distribution estimates the sample to sample variability expected in the population

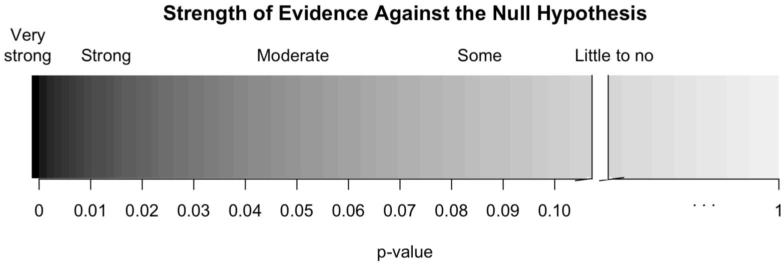
Theory-based method

- Use a mathematical model to determine a distribution under the null hypothesis
- Compare the observed sample statistic to the model to calculate a probability
- *Theory-based methods will be discussed in the next module*

P-value

- What does the p-value measure?
  - Probability of observing the sample ______________________________ or more ________________ assuming the ____________ hypothesis is ________________.

- How much evidence does the p-value provide against the null hypothesis?



- The ______________________________the p-value, the _______________________________ the evidence against the null hypothesis.

- Write a conclusion based on the p-value.

    - Answers the __________________________ question.

    - Amount of ____________________________ in support of the ____________________________ hypothesis.

- Decision: can we reject or fail to reject the null hypothesis?

  - Significance level: cut-off of “small” vs “large” p-value

    - p-value $\leq \alpha$

        - Strong enough evidence against the null hypothesis

        - Decision:

        - Results are ______________________________________ significant.

    - p-value $> \alpha$

        - Not enough evidence against the null hypothesis

        - Decision:

        - Results are not ___________________________________ significant.

One proportion test

- Reminder: review summary measures and plots discussed in the Week 3 material and Chapter 4 of the textbook.
- The summary measure for a single categorical variable is a _______________________.

Notation:

- Population proportion:
- Sample proportion:

Parameter of Interest:

- Include:
    - Reference of the population (true, long-run, population, all)
    - Summary measure
    - Context
        - Observational units/cases
        - Response variable (and explanatory variable if present)
            - If the response variable is categorical, define a ‘success’ in context

Hypothesis testing

Conditions:

- Independence:

Null hypothesis assumes "no effect", "no difference", "nothing interesting happening", etc.

Always of form: "parameter" = null value

H_0 :

H_A :

- Research question determines the direction of the alternative hypothesis.

Video 14.1 Example: A 2007 study published in the Behavioral Ecology and Sociobiology Journal was titled "Why do blue-eyed men prefer blue-eyed women?" (Laeng 2007) In this study, conducted in Norway, 114 volunteer heterosexual blue-eyed males rated the attractiveness of 120 pictures of females. The researchers recorded which eye-color (blue, green, or brown) was rated the highest, on average. In the sample, 51 of the volunteers rated the blue-eyed women the most attractive. Do blue-eyed heterosexual men tend to find blue-eyed women the most attractive?

Parameter of interest:

Write the null and alternative hypotheses for the blue-eyed study:

In notation:

H_0 :

H_A :

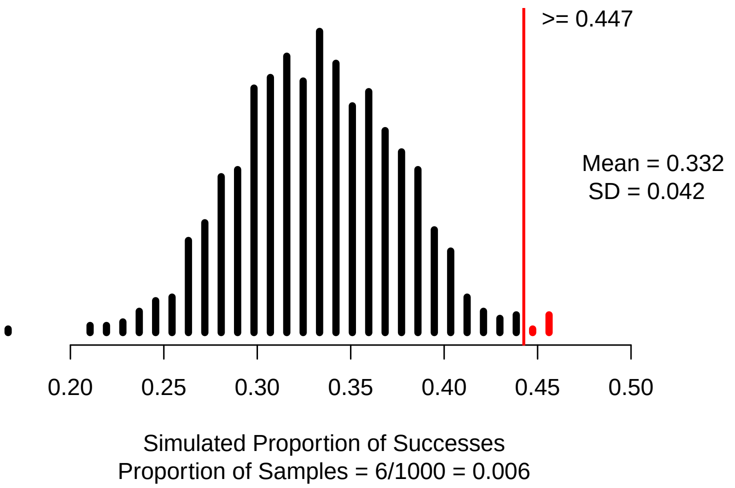
Statistic:

Is the independence condition met to analyze these data using a simulation-based approach?

Simulation-based method

- Simulate many samples assuming $H_0 : \pi = \pi_0$
  - Create a spinner with that represents the null value
  - Spin the spinner n times
  - Calculate and plot the simulated sample proportion from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\hat{p}$

```
set.seed(216)
one_proportion_test(probability_success = 0.333, # Null hypothesis value
          sample_size = 114, # Enter sample size
          number_repetitions = 1000, # Enter number of simulations
          as_extreme_as = 0.447, # Observed statistic
          direction = "greater", # Specify direction of alternative hypothesis
          summary_measure = "proportion") # Reporting proportion or number of successes?
```



Explain why the null distribution is centered at the value of approximately 0.333:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Generalization:

- Can the results of the study be generalized to the target population?

Confidence interval - Video Chapter10

statistic ± margin of error

Vocabulary:

- Point estimate:

- Margin of error:

Purpose of a confidence interval

- To give an ______________________________ ____________________________ for the parameter of interest

- Determines how _____________________ an effect is

Sampling distribution

- Ideally, we would take many samples of the same ________________ from the same population to create a sampling distribution

- But only have 1 sample, so we will _________________________ with _________________________ from the one sample.

- Need to estimate the sampling distribution to see the _________________________ in the sample

Simulation-based methods

Bootstrap distribution:

- Write the response variable values on cards
- Sample with replacement n times (bootstrapping)
- Calculate and plot the simulated difference in sample means from each simulation

- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

What is bootstrapping?

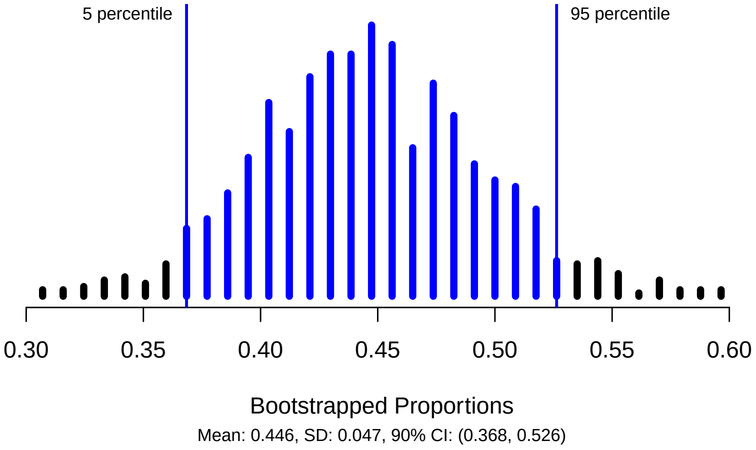
- Assume the "population" is many, many copies of the original sample.
- Randomly sample with replacement from the original sample n times.

Video 14.2

Let's revisit the blue-eyed male study to estimate the *proportion of ALL heterosexual blue-eyed males who tend to find blue-eyed women the most attractive* by creating a 90% confidence interval.

Bootstrap distribution:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                    number_successes = 51, # Observed number of successes
                    number_repetitions = 1000, # Number of bootstrap samples to use
                    confidence_level = 0.90) # Confidence level as a decimal
```



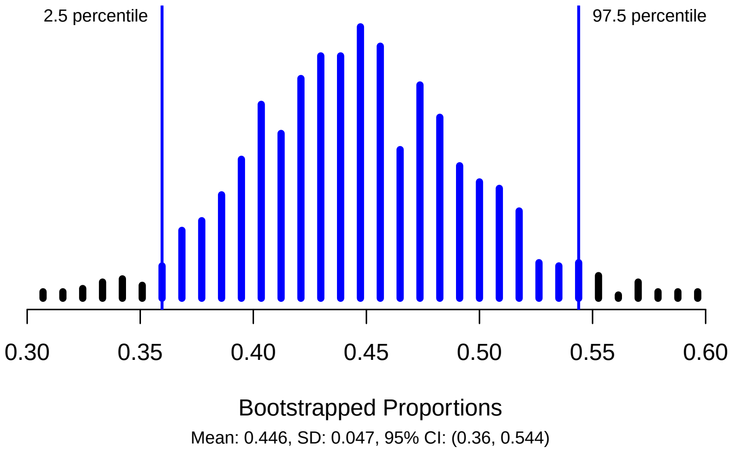
Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

How does changing the confidence level impact the width of the confidence interval?

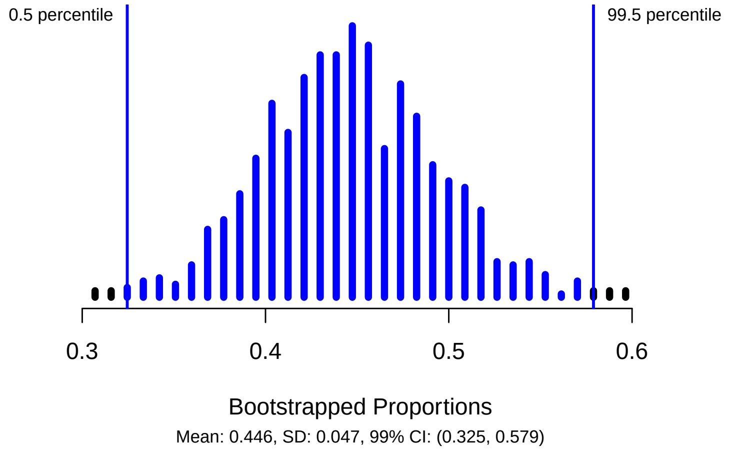
95% Confidence Interval:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                            number_successes = 51, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.95) # Confidence level as a decimal
```



99% Confidence Interval:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                            number_successes = 51, # Observed number of successes
                            number_repetitions = 1000, # Number of bootstrap samples to use
                            confidence_level = 0.99) # Confidence level as a decimal
```



3.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What is the summary measure calculated from a single categorical variable?

2. Write the alternative hypothesis for this study in notation? How was the direction of the alternative hypothesis determined?

3. Do the results of the confidence interval *match* the results based on the p-value?

3.3 Activity 6: Helper-Hinderer Part 1 — Simulation-based Hypothesis Test

3.3.1 Learning outcomes

- Identify the two possible explanations (one assuming the null hypothesis and one assuming the alternative hypothesis) for a relationship seen in sample data.
- Given a research question involving a single categorical variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for a single proportion.

3.3.2 Terminology review

In today's activity, we will work through a simulation-based hypothesis testing for a single categorical variable. Some terms covered in this activity are:

- Parameter of interest
- Null hypothesis
- Alternative hypothesis
- Simulation

To review these concepts, see Chapters 9 & 14 in your textbook.

3.3.3 Steps of the statistical investigation process

We will work through a five-step process to complete a hypothesis test for a single proportion, first introduced in the activity in week 1.

- **Ask a research question** that can be addressed by collecting data. What are the researchers trying to show?
- **Design a study and collect data**. This step involves selecting the people or objects to be studied and how to gather relevant data on them.
- **Summarize and visualize the data**. Calculate summary statistics and create graphical plots that best represent the research question.
- **Use statistical analysis methods to draw inferences from the data**. Choose a statistical inference method appropriate for the data and identify the p-value and/or confidence interval after checking assumptions. In this study, we will focus on using randomization to generate a simulated p-value.
- **Communicate the results and answer the research question**. Using the p-value and confidence interval from the analysis, determine whether the data provide statistical evidence against the null hypothesis. Write a conclusion that addresses the research question.

3.3.4 Helper-Hinderer

A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. As a class we will watch this short video to see how the experiment was run: https://youtu.be/anCaGBsBOxM. Researchers were hoping to assess: Are infants more likely to choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

In this study, the **observational units are the infants ages 6 to 10 months**. The **variable measured on each observational unit (infant) is whether they chose the helper or the hinderer toy**. This is a categorical variable so we will be assessing the proportion of infants ages 6 to 10 months that choose the helper toy. Choosing the helper toy in this study will be considered a success.

Ask a research question

1. Identify the research question for this study. What are the researchers hoping to show?

Design a study and collect data

Before using statistical inference methods, we must check that the cases are independent. The sample observations are independent if the outcome of one observation does not influence the outcome of another. One way this condition is met is if data come from a simple random sample of the target population.

2. Are the cases independent? Justify your answer.

R code

For almost all activities and labs it will be necessary to upload the provided R script file from D2L for that day. Your instructor will highlight a few steps in uploading files to and using RStudio.

The following are the steps to upload the necessary R script file for this activity:

- Download the Activity R script file from D2L.
- Click "Upload" in the "Files" tab in the bottom right window of RStudio. In the pop-up window, click "Choose File", and navigate to the folder where the Activity R script file is saved (most likely in your downloads folder). Click "Open"; then click "Ok".
- You should see the uploaded file appear in the list of files in the bottom right window. Click on the file name to open the file in the Editor window (upper left window).

Notice that the first threelines of code contain a prompt called `library`. Packages needed to run functions in R are stored in directories called libraries. When using the MSU RStudio server, all the packages needed for the class are already installed. We simply must tell R which packages we need for each R script file. We use the prompt `library` to load each **package** (or library) needed for each activity. Note, these `library` lines MUST be run each time you open a R script file in order for the functions in R to work.

- Highlight and run lines 1–3 to load the packages needed for this activity. Notice the use of the # symbol in the R script file. This symbol is not part of the R code. It is used by these authors to add comments to the R code and explain what each call is telling the program to do.

R will ignore everything after a # symbol when executing the code. Refer to the instructions following the # symbol to understand what you need to enter in the code.

```
library(tidyverse)
library(ggplot2)
library(catstats)
```

Throughout activities, we will often include the R code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into R using the `read.csv()` function. The line of code shown below (line 7 in the R script file) reads in the data set and names the data set `infants`.

Summarize and visualize the data

The following code reads in the data set and gives the number of infants in each level of the variable, whether the infant chose the helper or the hinderer.

- Highlight and run lines 7 and 8 to check that you get the same counts as shown below

```
 # Read in data set
infants <- read.csv("https://math.montana.edu/courses/s216/data/infantchoice.csv")
infants %>% count(choice)  # Count number in each choice category
```

```
#>     choice  n
#> 1   helper 14
#> 2 hinderer  2
```

The following formula is used to calculate the proportion of successes in the sample.

$$\hat{p} = \frac{\text{number of successes}}{\text{total number of observational units}}$$

3. Using the R output and the formula given, calculate the summary statistic (sample proportion) to represent the research question. Recall that `choosing the helper toy` is a considered a success. Use appropriate notation.

To visually display this data we can use either a frequency bar plot or a relative frequency bar plot.

- Enter the name of the variable name `choice` for `variable` in the R code to create the frequency bar plot.
- Note the name of the title is given in line 16 and includes the type of plot, observational units, and variable name
- Highlight and run lines 13–19 to create the plot

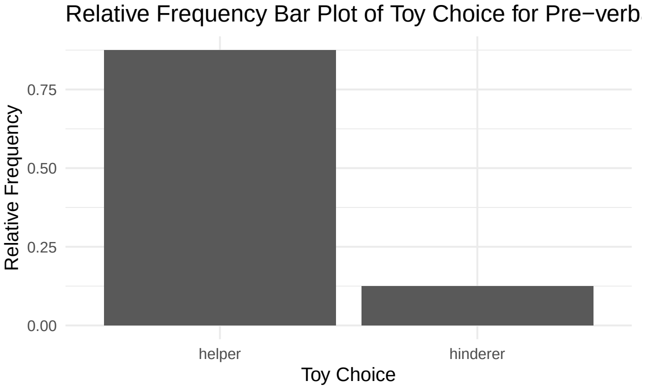
```
infants %>% # Data set piped into...
    ggplot(aes(x = variable)) +   # This specifies the variable
    geom_bar(stat = "count") +  # Tell it to make a bar plot
    labs(title = "Frequency Bar Plot of Toy Choice for Pre-verbal Infants",
       # Give your plot a title
       x = "Toy Choice",    # Label the x axis
       y = "Frequency")  # Label the y axis
```

4. Sketch the frequency bar plot created below.

We could also choose to display the data as a proportion in a **relative frequency** bar plot. To find the relative frequency, the count in each level of `choice` is divided by the sample size. This calculation is the sample proportion for each level of choice. Notice that in the following code we told R to create a bar plot with proportions.

- In the R script file, highlight and run lines 23–29 to create the relative frequency bar plot.

```
infants %>% # Data set piped into...
    ggplot(aes(x = choice)) +    # This specifies the variable
    geom_bar(aes(y = after_stat(prop), group = 1)) +  # Tell it to make a bar plot with proportions
    labs(title = "Relative Frequency Bar Plot of Toy Choice for Pre-verbal Infants",
       # Give your plot a title
       x = "Toy Choice",    # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



5. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

We cannot assess whether infants are more likely to choose the helper toy based on the statistic and plot alone. The next step is to analyze the data by using a hypothesis test to discover if there is evidence against the null hypothesis.

Use statistical analysis methods to draw inferences from the data

When performing a hypothesis test, we must first identify the null hypothesis. The null hypothesis is written about the parameter of interest, or the value that summarizes the variable in the population.

The parameter of interest is a statement about what we want to find about the population. The following must be included when writing the parameter of interest.

- Population word (true, long-run, population)
- Summary measure (depends on the type of data)
- Context
  - Observational units
  - Variable(s)

For this study, the parameter of interest, π, represents the **true or population proportion of infants ages 6–10 months who will choose the helper toy**.

If the children are just randomly choosing the toy, we would expect half (0.5) of the infants to choose the helper toy. This is the null value for our study.

6. Using the parameter of interest given above, write out the null hypothesis in words. That is, what do we assume to be true about the parameter of interest when we perform our simulation?

The notation used for a population proportion (or probability, or true proportion) is π. Since this summarizes a population, it is a parameter. When writing the **null hypothesis** in notation, we set the parameter equal to the null value, $H_0 : \pi = \pi_0$.

7. Write the null hypothesis in notation using the null value of 0.5 in place of π_0 in the equation given on the previous page.

The **alternative hypothesis** is the claim to be tested and the direction of the claim (less than, greater than, or not equal to) is based on the research question.

8. Based on the research question from question 1, are we testing that the parameter is greater than 0.5, less than 0.5 or different than 0.5?

9. Write out the alternative hypothesis in notation.

Remember that when utilizing a hypothesis test, we are evaluating two competing possibilities. For this study the **two possibilities** are either…

- The true proportion of infants who choose the helper is 0.5 and our results just occurred by random chance; or,
- The true proportion of infants who choose the helper is greater than 0.5 and our results reflect this.

Notice that these two competing possibilities represent the null and alternative hypotheses.

We will now simulate one sample of a **null distribution** of sample proportions. The null distribution is created under the assumption the null hypothesis is true. In this case, we assume the true proportion of infants who choose the helper is 0.5, so we will create 1000 (or more) different simulations of 16 infants under this assumption.

Let's think about how to use a coin to create one simulation of 16 infants under the assumption the null hypothesis is true. Let heads equal infant chose the helper toy and tails equal infant chose the hinderer toy.

10. How many times would you flip a coin to simulate the sample of infants?

11. Flip a coin 16 times recording the number of times the coin lands on heads. This represents one simulated sample of 16 infants randomly choosing the toy. Calculate the proportion of coin flips that resulted in heads.

12. Is the value from question 9 closer to 0.5, the null value, or closer to the sample proportion, 0.875?

Report the number of coin flips you got in the Google sheet on D2L.

13. Sketch the graph created by your instructor of the proportion of heads out of 16 coin flips.

14. Circle the observed statistic (value from question 3) on the distribution shown above. Where does this statistic fall in this distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

15. Is the observed statistic likely to happen or unlikely to happen if the true proportion of infants who choose the helper is 0.5? Explain your answer using the plot.

In the next class, we will continue to assess the strength of evidence against the null hypothesis by using a computer to simulate 1000 samples when we assume the null hypothesis is true.

3.3.5 Take-home messages

1. Two types of plots are used for plotting categorical variables: frequency bar plots, relative frequency bar plots.

2. In a hypothesis test we have two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis represents either a skeptical perspective or a perspective of no difference or no effect. The alternative hypothesis represents a new perspective such as the possibility that there has been a change or that there is a treatment effect in an experiment.

3. In a simulation-based test, we create a distribution of possible simulated statistics for our sample if the null hypothesis is true. Then we see if the calculated observed statistic from the data is likely or unlikely

to occur when compared to the null distribution.

4. To create one simulated sample on the null distribution for a sample proportion, spin a spinner with probability equal to π_0 (the null value), n times or draw with replacement n times from a deck of cards created to reflect π_0 as the probability of success. Calculate and plot the proportion of successes from the simulated sample.

3.3.6 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

3.4 Activity 7: Helper-Hinderer (continued)

3.4.1 Learning outcomes

- Describe and perform a simulation-based hypothesis test for a single proportion.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a single proportion.
- Explore what a p-value represents

3.4.2 Steps of the statistical investigation process

In today's activity we will continue with steps 4 and 5 in the statistical investigation process. We will continue to assess the Helper-Hinderer study from last class.

- **Ask a research question** that can be addressed by collecting data. What are the researchers trying to show?
- **Design a study and collect data**. This step involves selecting the people or objects to be studied and how to gather relevant data on them.
- **Summarize and visualize the data**. Calculate summary statistics and create graphical plots that best represent the research question.
- **Use statistical analysis methods to draw inferences from the data**. Choose a statistical inference method appropriate for the data and identify the p-value and/or confidence interval after checking assumptions. In this study, we will focus on using randomization to generate a simulated p-value.
- **Communicate the results and answer the research question**. Using the p-value and confidence interval from the analysis, determine whether the data provide statistical evidence against the null hypothesis. Write a conclusion that addresses the research question.

3.4.3 Helper-Hinderer

A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. As a class we will watch this short video to see how the experiment was run: https://youtu.be/anCaGBsBOxM. Researchers were hoping to assess: Are infants more likely to choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

1. Report the sample proportion (summary statistic) calculated in the previous activity.

2. Write the alternative hypothesis in words in context of the problem. Remember the direction we are testing is dependent on the research question.

Today, we will use the computer to simulate a null distribution of 1000 different samples of 16 infants, plotting the proportion who chose the helper in each sample, based on the assumption that the true proportion of infants who choose the helper is 0.5 (or that the null hypothesis is true).

To use the computer simulation, we will need to enter the

- assumed "probability of success" (π_0),
- "sample size" (the number of observational units or cases in the sample),
- "number of repetitions" (the number of samples to be generated - typically we use 10000),
- "as extreme as" (the observed statistic), and
- the "direction" (matches the direction of the alternative hypothesis).

3. What values should be entered for each of the following into the one proportion test to create 1000 simulations?

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

- Direction (`"greater"`, `"less"`, or `"two-sided"`):

We will use the `one_proportion_test()` function in `R` (in the `catstats` package) to simulate the null distribution of sample proportions and compute a p-value. Using the provided `R` script file, fill in the values/words for each `xx` with your answers from question 3 in the one proportion test to create a null distribution with 1000 simulations. Then highlight and run lines 1–16.

```r
one_proportion_test(probability_success = xx, # Null hypothesis value
          sample_size = xx, # Enter sample size
          number_repetitions = 10000, # Enter number of simulations
          as_extreme_as = xx, # Observed statistic
          direction = "xx", # Specify direction of alternative hypothesis
          summary_measure = "proportion") # Reporting proportion or number of successes?
```

4. Sketch the null distribution created from the `R` code here.

5. Around what value is the null distribution centered? Why does that make sense?

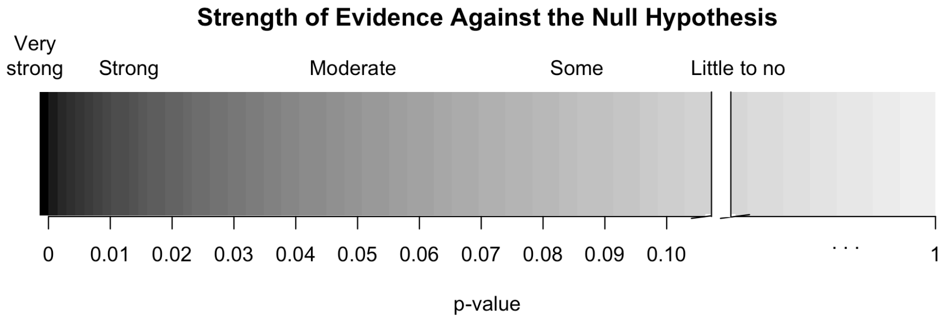
6. Circle the observed statistic (value from question 1) on the distribution you drew in question 4. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

7. Is the observed statistic likely to happen or unlikely to happen if the true proportion of infants who choose the helper is 0.5? Explain your answer using the plot.

8. Using the simulation, what is the proportion of simulated samples that generated a sample proportion at the observed statistic or greater, if the true proportion of infants who choose the helper is 0.5? *Hint*: Look under the simulation.

The value in question 8 is the **p-value**. The smaller the p-value, the more evidence we have against the null hypothesis.

9. Using the following guidelines for the strength of evidence, how much evidence do the data provide against the null hypothesis? (Circle one of the five descriptions.)



Interpret the p-value

The p-value measures the probability that we observe a sample proportion as extreme as what was seen in the data or more extreme (matching the direction of the Ha) IF the null hypothesis is true. This is a conditional probability, calculated dependent on the null hypothesis being true. Represented in probability notaton:

$$P(\text{statistic or more extreme}|\text{the null hypothesis is true})$$

10. What did we assume to create the null distribution? Write the null hypothesis is context.

11. What value did we compare to the null distribution to find the p-value? What is the value of the summary statistic (sample proportion)?

12. In what direction (greater than or less than) did we count from the statistic to find the number of simulations?

13. Fill in the blanks below to interpret the p-value.

We would observe a sample proportion of ______________________________

or (greater, less, more extreme) ______________________________

with a probability of ______________________________

IF we assume (H_0 in context) ______________________________.

____

Communicate the results and answer the research question

When we write a conclusion we answer the research question by stating how much evidence there is for the alternative hypothesis.

14. Write a conclusion in context of the study. How much evidence does the data provide in support of the alternative hypothesis?

3.4.4 Take-home messages

1. The null distribution is created based on the assumption the null hypothesis is true. We compare the sample statistic to the distribution to find the likelihood of observing this statistic.
2. The p-value measures the probability of observing the sample statistic or more extreme (in direction of the alternative hypothesis) is the null hypothesis is true.

3.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

3.5 Activity 8: Helper-Hinderer — Simulation-based Confidence Interval

3.5.1 Learning outcomes

- Use bootstrapping to find a confidence interval for a single proportion.
- Interpret a confidence interval for a single proportion.

3.5.2 Terminology review

In today's activity, we will introduce simulation-based confidence intervals for a single proportion. Some terms covered in this activity are:

- Parameter of interest
- Bootstrapping
- Confidence interval

To review these concepts, see Chapters 10 & 14 in your textbook.

3.5.3 Helper-Hinderer

In the last class, we found very strong evidence that the true proportion of infants who will choose the helper character is greater than 0.5. But what *is* the true proportion of infants who will choose the helper character? We will use this same study to estimate this parameter of interest by creating a confidence interval.

As a reminder: A study by Hamblin, Wynn, and Bloom reported in Nature (Hamblin, Wynn, and Bloom 2007) was intended to check young kids' feelings about helpful and non-helpful behavior. Non-verbal infants ages 6 to 10 months were shown short videos with different shapes either helping or hindering the climber. Researchers were hoping to assess: Are infants more likely to preferentially choose the helper toy over the hinderer toy? In the study, of the 16 infants age 6 to 10 months, 14 chose the *helper* toy and 2 chose the *hinderer* toy.

A **point estimate** (our observed statistic) provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible *range* of values for the parameter. This plausible range of values for the population parameter is called an **interval estimate** or **confidence interval**.

Activity intro

1. What is the value of the point estimate?

2. If we took another random sample of 16 infants, would we get the exact same point estimate? Explain why or why not.

In today's activity, we will use bootstrapping to find a 95% confidence interval for π, the parameter of interest.

3. In your own words, explain the bootstrapping process.

Use statistical analysis methods to draw inferences from the data

4. Write out the parameter of interest in words, in context of the study. *Hint: this is the same as in Activity 6 and 7.*

To create the null distribution we flipped a coin 16 times to simulate infants randomly choosing the helper toy with a probability of 50%.

5. Why can't we use a coin to simulate the bootstrap distribution.

To create the bootstrap distribution.

- First we would label the cards to represent the sample statistic: 14 helper and 2 hinderer.
- Sample with replacement 16 times

6. Using the cards provided by your instructor, create one bootstrap sample. Report your simulated sample proportion on the whiteboard.

To use the computer simulation to create a bootstrap distribution, we will need to enter the

- "sample size" (the number of observational units or cases in the sample),
- "number of successes" (the number of cases that choose the helper character),
- "number of repetitions" (the number of samples to be generated), and
- the "confidence level" (which level of confidence are we using to create the confidence interval).

7. What values should be entered for each of the following into the simulation to create the bootstrap distribution of sample proportions to find a 95% confidence interval?

- Sample size:

- Number of successes:

- Number of repetitions:

- Confidence level (as a decimal):

We will use the `one_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample proportions and calculate a confidence interval. Using the provided R script file, fill in the values/words for each `xx` with your answers from question 5 in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations. Then highlight and run lines 1–9.

```r
one_proportion_bootstrap_CI(sample_size = xx, # Sample size
                    number_successes = xx, # Observed number of successes
                    number_repetitions = 10000, # Number of bootstrap samples to use
                    confidence_level = xx) # Confidence level as a decimal
```

8. Sketch the bootstrap distribution created below.

9. What is the value at the center of this bootstrap distribution? Why does this make sense?

10. Explain why the two vertical lines are at the 2.5th percentile and the 97.5th percentile.

11. Report the 95% bootstrapped confidence interval for π. Use interval notation: (lower value, upper value).

12. Interpret the 95% confidence interval in context.

Communicate the results and answer the research question

13. Is the value 0.5 (the null value) in the 95% confidence interval?

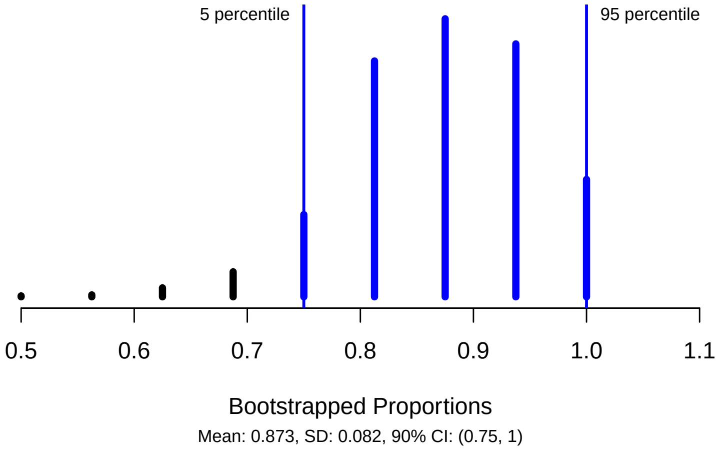
    Explain how this indicates that the p-value provides strong evidence against the null.

Effect of confidence level

14. Suppose instead of finding a 95% confidence interval, we found a 90% confidence interval. Would you expect the 90% confidence interval to be narrower or wider? Explain your answer.

15. The following R code produced the bootstrap distribution with 1000 simulations that follows. Circle the value that changed in the code.

```r
one_proportion_bootstrap_CI(sample_size = 16, # Sample size
                    number_successes = 14, # Observed number of successes
                    number_repetitions = 1000, # Number of bootstrap samples to use
                    confidence_level = 0.90) # Confidence level as a decimal
```



16. Report both the 95% confidence interval (question 9) and the 90% confidence interval (question 13). Is the 90% confidence interval narrower or wider than the 95% confidence interval?

17. Explain why the upper value of the confidence interval is truncated at 1.

18. Fill in the blanks below to write a paragraph summarizing the results of the study as if writing a press release.

Researchers were interested if infants observe social cues and would be more likely to choose the helper toy over the hinderer toy. In a sample of (sample size) ____________________infants, (number of successes) _______________________chose the helper toy. A simulation null distribution with 1000 simulations was created in RStudio. The p-value was found by calculating the proportion of simulations in the null distribution at the sample statistic of 0.875 and greater. This resulted in a p-value of (value of p-value)_______________________. We would observe a sample proportion of (value of the sample proportion) __________________________________ or (greater, less, more extreme) ________________________________ with a probability of (value of p-value)_____________________________
IF we assume (H_0 in context) ____.
Based on this p-value, there is (very strong/little to no) __________________________________ evidence that the (sample/true)________________________________ proportion of infants age 6 to 10 months who will choose the helper toy is (greater than, less than, not equal to) ________________________________ 0.5. In addition, a 95% confidence interval was found for the parameter of interest. We are 95% confident that the (true/sample)_______________________________________ proportion of infants age 6 to 10 months who will choose the helper toy is between (lower value)________________________ and (upper

value)______________________________. The results of this study can be generalized to (all infants age 6 to 10 months/infants similar to those in this study)____ as the researchers (did/did not)________________________________ select a random sample.

3.5.4 Take-home messages

1. The goal in a hypothesis test is to assess the strength of evidence for an effect, while the goal in creating a confidence interval is to determine how large the effect is. A **confidence interval** is a range of *plausible* values for the parameter of interest.
2. A confidence interval is built around the point estimate or observed calculated statistic from the sample. This means that the sample statistic is always the center of the confidence interval. A confidence interval includes a measure of sample to sample variability represented by the **margin of error**.
3. In simulation-based methods (bootstrapping), a simulated distribution of possible sample statistics is created showing the possible sample-to-sample variability. Then we find the middle X percent of the distribution around the sample statistic using the percentile method to give the range of values for the confidence interval. This shows us that we are $X\%$ confident that the parameter is within this range, where X represents the level of confidence.
4. When the null value is within the confidence interval, it is a plausible value for the parameter of interest; thus, we would find a larger p-value for a hypothesis test of that null value. Conversely, if the null value is NOT within the confidence interval, we would find a small p-value for the hypothesis test and strong evidence against this null hypothesis.
5. To create one simulated sample on the bootstrap distribution for a sample proportion, label n cards with the original responses. Draw with replacement n times. Calculate and plot the resampled proportion of successes.

3.5.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

MODULE 4

Inference for a Single Categorical Variable: Theory-based Methods

4.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a single categorical variable.

- **Theory-based methods**: when specific conditions are met, a data can be fit with a theoretical distribution
- **Conditions for the sampling distribution of $\hat{p}$ to follow an approximate normal distribution**:
  - **Independence**: The sample's observations are independent, e.g., are from a simple random sample. (*Remember*: This also must be true to use simulation methods!)
  - **Large enough sample size: Success-failure condition**: We *expect* to see at least 10 successes and 10 failures in the sample, $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
- **Standardized statistic**: calculation to standardize the sample statistic in order to compare the standardized value to the theoretical distribution
  - Measures the number of standard errors the sample statistic is from the null value.
- **Standard normal distribution**: a theoretical distribution that is symmetric centered on the mean of zero with a standard deviation of one

$$N(0,1)$$

- **Standardized sample proportion**: standardized statistic for a single categorical variable calculated using:

$$Z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})},$$

- **Standard error of the sample proportion assuming the null is true**: measures the how far each possible sample proportion is from the true proportion, on average and is calculated using the null value:

$$SE_0(\hat{p}) = \sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}$$

.

- The p-value can be found by using the pnorm function.
  - Enter the value of the standardized statistic for xx

```
pnorm(xx, lower.tail=TRUE)
```

- **Margin of error**: half the width of the confidence interval

$$ME = z^* \times SE(\hat{p})$$

- **Standard error of the sample proportion for a confidence interval**

$$SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

- To find the confidence interval add and subtract the margin of error to the sample statistic

$$\hat{p} \pm ME$$

- R code to find the multiplier for the confidence interval using theory-based methods.
  - qnorm will give you the multiplier using the standard normal distribution
  - Enter the percentile for the given level of confidence

```
qnorm(percentile, lower.tail=FALSE)
```

4.1.1 Key topics

- Theory-based methods should give the same results as simulation based methods if the sample size is large enough (success-failure condition is met).
- If repeat samples of the same size are taken from the population, 95% of samples will create a 95% confidence interval that contains the parameter of interest.

4.2 Video Notes: Inference for One Categorical Variable using Theory-based Methods

Read Chapters 11 and 13 and Sections 14.3 and 14.4 in the course textbook. Use the following videos to complete the video notes for Module 4.

4.2.1 Course Videos

- Chapter11
- 14.3TheoryTests
- 14.3TheoryIntervals

Theory-based methods

Central limit theorem - Video Chapter11

The Central Limit Theorem tells us that the ____________________ distribution of a sample proportion (and sample mean and sample differences) will be approximately ____________________ if the sample size is ____________________ ______________________.

The ____________________ of the distribution of sample proportions (sampling distribution) from thousands of samples will be bell-shaped/symmetric (Normal), if the sample size is large enough and the observations are ______________________.

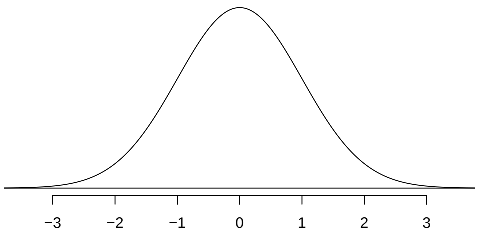
- $\hat{p} \sim N(\pi, \sqrt{\frac{\pi\times(1-\pi)}{n}})$

Conditions of the CLT:

- Independence (*also must be met to use simulation methods*): the response for one observational unit will not influence another observational unit
- Large enough sample size:

Normal distribution:

- Bell-shaped and ____________________
- Standard normal distribution: $N(0,1)$



Standardized statistic: Z - score

- $Z = \frac{\text{statistic}-\text{null value}}{\text{standard error of the statistic}}$

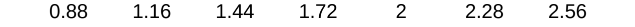
- Measures the ________________ of standard ___________________ the statistic is from the null value

Example(s): Heights of Caucasian American adult males are roughly Normally distributed with a mean of 1.72 m and a standard deviation of 0.28 m. Find and interpret the z-score for a man who is 5’4” (1.626 m) tall. Round your answer to three decimal places.

Heights of Caucasian American adult females are roughly Normally distributed with a mean of 1.59 meters and a standard deviation of 0.22 meters. Which is more unusual: a 5’4” (1.626 m) tall male or a 5’9” (1.753 m) tall female?

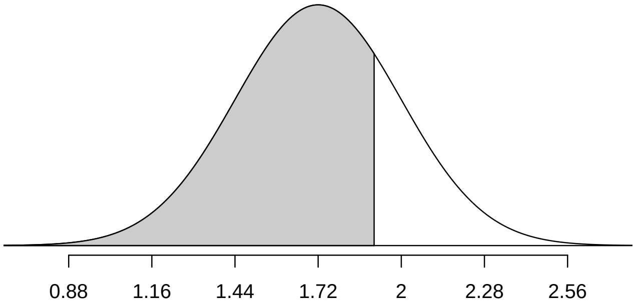
In a Normal curve, the area under the curve is equal to 1, representing a probability. Therefore the shaded area represents the probability of a man being under 1.626 meters tall.

```r
library(openintro)
normTail(m = 1.72, s = 0.28, L = 1.626)
pnorm(mean = 1.72, sd = 0.28, q = 1.626)
#> [1] 0.3685432
```



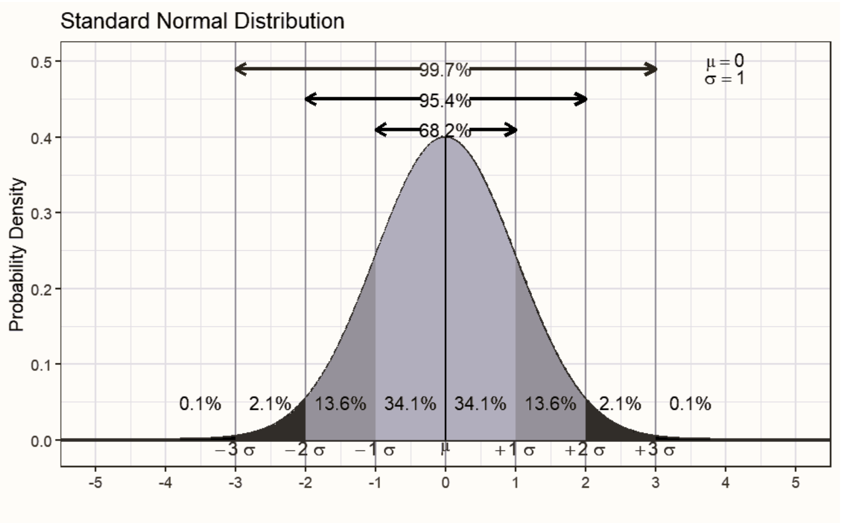
We can also reverse that order. Given a percentage, we can find the associated percentile, or quantile. Here we display calculating the value that cuts off the lower 0.75 proportion of male adult Caucasian heights using the qnorm() function.

```r
qnorm(mean = 1.72, sd = 0.28, p = 0.75)
#> [1] 1.908857
normTail(m = 1.72, s = 0.28, L = 1.909)
```



68-95-99.7 Rule

- 68% of Normal distribution within 1 SD of the mean (mean – SD, mean + SD)
- 95% within (mean – 2SD, mean + 2SD)
- 99.7% within (mean – 3SD, mean + 3SD)



General steps of a hypothesis test

1. Write a research question and hypotheses.
2. Collect data and calculate a summary statistic.
3. Model a sampling distribution which assumes the null hypothesis is true.
4. Calculate a p-value.
5. Draw conclusions based on a p-value.

Example in Video 14.3TheoryTests

Example: The American Red Cross reports that 10% of US residents eligible to donate blood actually do donate. A poll conducted on a representative of 200 Montana residents eligible to donate blood found that 33 had donated blood sometime in their life. Do Montana residents donate at a different rate than US population?

Hypotheses:

In notation:

H_0 :

H_A :

Parameter of interest:

Conditions for inference using theory-based methods:

- Independence:
  - The outcome of one observation does not influence the outcome of another.
  - Taking a random sample is one way to satisfy this condition.
- Large enough sample size:

Are the conditions met to analyze the blood donations data using theory-based methods?

To use theory-based methods to perform a hypothesis test:

- 1st: Calculate the standardized statistic
- 2nd: Find the area under the standard normal distribution at least as extreme as the standardized statistic

Equation for the standard error of the sample proportion assuming the null hypothesis is true:

- This value measures how far each possible sample statistic is from the null value, on average.

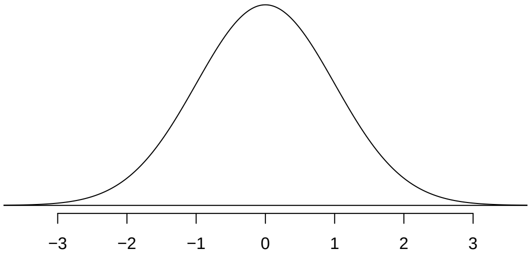
Equation for the standardized sample proportion:

- This value measures how many standard deviations the sample proportion is above/below the null value.

Calculate the standardized sample proportion of Montana residents that have donated blood sometime in their life.

- First calculate the standard error of the sample proportion assuming the null hypothesis is true

- Then calculate the Z score.



Interpret the standardized statistic

To find the p-value, find the area under the standard normal distribution at the standardized statistic and more extreme.

```
pnorm(3.064, lower.tail = FALSE)*2
```

```
#> [1] 0.002183989
```

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Decision at a significance level of 0.05 ($\alpha = 0.05$):

Generalization:

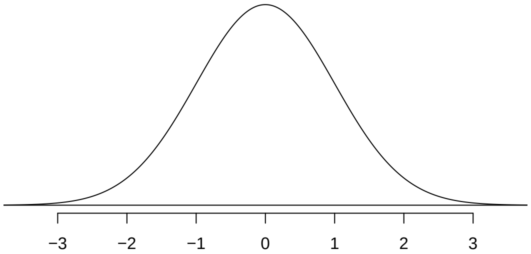
- Can the results of the study be generalized to the target population?

Confidence interval - 14.3TheoryIntervals

- Interval of ________________ values for the parameter of interest
- $CI =$ statistic ± margin of error

Theory-based method for a single categorical variable

- $CI = \hat{p} \pm (z^* \times SE(\hat{p}))$
- Multiplier (z^*) is the value at a certain ___________________ under the standard normal distribution



For a 95% confidence interval:

```
qnorm(0.975, lower.tail=TRUE)
```

```
#> [1] 1.959964
```

- When creating a confidence interval, we no longer assume the _____________________ hypothesis is true. Use _____________ to calculate the sample to sample variability, rather than π_0.

Equation for the standard error of the sample proportion *NOT* assuming the null is true:

Example: Estimate the true proportion of Montana residents that have donated blood at least once in their life.

Find a 95% confidence interval:

Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Interpreting confidence level

What does it mean to be 95% confident in a created confidence interval?

- Our goal is to only take one sample from the population to create a confidence interval.
- Based on the 68-95-99.7 rule, we know that approximately _________% of sample _____________________ will fall within _______________ from the parameter.
- If we create 95% confidence intervals, ____________% of samples will create a 95% _____________________ interval that will contain the ____________________ of interest.
- 95% of samples accurately _____________________ the parameter of interest
  - When we create one confidence interval, we are 95% ________________________ that we have a "good" sample that created a confidence interval that contains the ________________ of interest.

Interpret the confidence **level** for the blood donation study.

4.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What conditions must be met to use the Normal Distribution to approximate the sampling distribution of sampling proportions?

2. Should the conclusion include a population word like *true* or *long-run*? Explain your answer.

4.3 Activity 9: Handedness of Male Boxers

4.3.1 Learning outcomes

- Describe and perform a theory-based hypothesis test for a single proportion.
- Check the appropriate conditions to use a theory-based hypothesis test.
- Calculate and interpret the standardized sample proportion.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a single proportion.
- Use the normal distribution to find the p-value.

4.3.2 Terminology review

In this activity, we will introduce theory-based hypothesis tests for a single categorical variable. Some terms covered in this activity are:

- Parameter of interest
- Standardized statistic
- Normal distribution
- p-value

To review these concepts, see Chapter 11 & 14 in your textbook.

Activities from module 5 covered simulation-based methods for hypothesis tests involving a single categorical variable. This activity covers theory-based methods for testing a single categorical variable.

4.3.3 Handedness of male boxers

Left-handedness is a trait that is found in about 10% of the general population. Past studies have shown that left-handed men are over-represented among professional boxers (Richardson and Gilman 2019). The fighting claim states that left-handed men have an advantage in competition. In this random sample of 500 male professional boxers, we want to see if there is an over-prevalence of left-handed fighters. In the sample of 500 male boxers, 81 were left-handed.

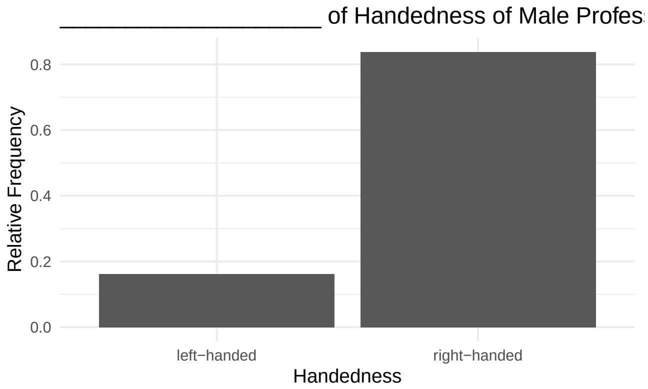
4.3.4 Summary statistics review

- Download the R file for today's activity from D2L
- Upload the file to the R server
- Run lines 1–15 to load the needed packages and the data set and create a plot of the data

```
 # Read in data set
boxers <- read.csv("https://math.montana.edu/courses/s216/data/Male_boxers_sample.csv")
boxers %>% count(Stance)  # Count number in each Stance category
```

```
#>         Stance   n
#> 1  left-handed  81
#> 2 right-handed 419
```

```
boxers %>% # Data set piped into...
    ggplot(aes(x = Stance)) +   # This specifies the variable
    geom_bar(aes(y = after_stat(prop), group = 1)) +  # Tell it to make a bar plot with proportions
    labs(title = "_____________________ of Handedness of Male Professional Boxers",
       # Give your plot a title
       x = "Handedness",   # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



1. What type of plot was created of these data?

Hypotheses and summary statistics

2. Write out the parameter of interest in words, in context of the study.

3. Write out the null hypothesis in words.

4. Write out the alternative hypothesis in notation.

5. Give the value of the summary statistic (sample proportion) for this study. Use proper notation.

Theory-based methods

The sampling distribution of a single proportion — how that proportion varies from sample to sample — can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sampling distribution of $\hat{p}$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent, e.g., are from a simple random sample. (*Remember*: This also must be true to use simulation methods!)
- **Success-failure condition**: We *expect* to see at least 10 successes and 10 failures in the sample, $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

6. Verify that the independence condition is satisfied.

7. Is the success-failure condition met to model the data with the normal distribution? Explain your answer in context of the problem.

To calculate the standardized statistic we use the general formula

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_0(\text{point estimate})}.$$

For a single categorical variable the standardized sample proportion is calculated using

$$Z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})},$$

where the standard error is calculated using the null value:

$$SE_0(\hat{p}) = \sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}$$

.

For this study, the null standard error of the sample proportion is calculated using the null value, 0.1.

$$SE_0(\hat{p}) = \sqrt{\frac{0.1 \times (1-0.1)}{500}} = 0.013$$

.

Each sample proportion of male boxers that are left-handed is 0.013 from the true proportion of male boxers that are left-handed, on average.

8. Label the standard normal distribution shown below with the null value as the center value (below the value of zero). Label the tick marks to the right of the null value by adding 1 standard error to the null value to represent 1 standard error, 2 standard errors, and 3 standard errors from the null. Repeat this process to the left of the null value by subtracting 1 standard error for each tick mark.

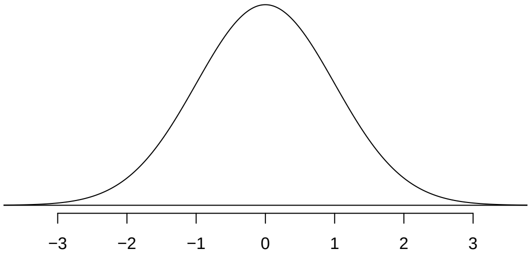


Figure 4.1: Standard Normal Curve

9. Using the null standard error of the sample proportion, calculate the standardized sample proportion (Z). Mark this value on the standard normal distribution above.

The standardized statistic is used as a ruler to measure how far the sample statistic is from the null value. Essentially, we are converting the sample proportion into a measure of standard errors to compare to the standard normal distribution.

The standardized statistic measures the *number of standard errors the sample statistic is from the null value.*

10. Interpret the standardized sample proportion from question 9 in context of the problem.

We will use the `pnorm()` function in `R` to find the p-value. The value for Z was entered into the code below to get the p-value. Check that this answer matches what you calculated in question 7. Notice that we used `lower.tail = FALSE` to find the p-value. `R` will calculate the p-value *greater* than the value of the standardized statistic.

Notes:

- Use `lower.tail = TRUE` when doing a left-sided test.
- Use `lower.tail = FALSE` when doing a right-sided test.
- To find a two-sided p-value, use a left-sided test for negative Z or a right-sided test for positive Z, then multiply the value found by 2 to get the p-value.
- Enter the value of the standardized statistic for xx
- Highlight and run lines 21–23

```r
pnorm(xx, # Enter value of standardized statistic
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
      lower.tail=FALSE) # Gives a p-value greater than the standardized statistic
```

11. Report the p-value obtained from the `R` output.

12. Write a conclusion based on the value of the p-value.

Impacts on the P-value

Suppose that we want to show that the true proportion of male boxers **differs** from that in the general population.

13. Write out the alternative hypothesis in notation for this new research question.

14. How would this impact the p-value?

15. Suppose instead of 500 male boxers the researchers only took a sample of 300 male boxers and found the same proportion ($\hat{p} = 0.162$) of male boxers that are left-handed. Since we are still assuming the same null value, 0.1, the standard error would be calculated as below:

$$SE_0(\hat{p}) = \sqrt{\frac{0.1(1-0.1)}{300}} = 0.017$$

.

The standardized statistic for this new sample is calculated below:

$$t = \frac{0.162 - 0.1}{0.017} = 3.64$$

16. Mark the value of the original standardized statistic from question 9 and the value of the standardized statistic from the smaller sample size on the standard normal distribution below.

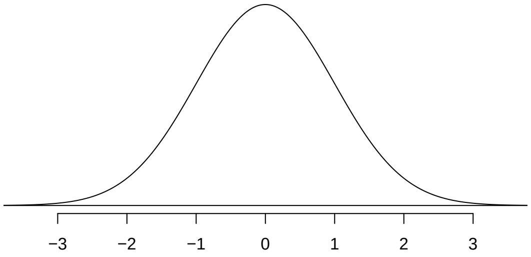


Figure 4.2: Standard Normal Curve

17. How does the decrease in sample size affect the p-value?

18. Suppose another sample of 500 male boxers was taken and 68 were found to be left-handed. Since we are still assuming the same null value, 0.1, the standard error would be calculated as before:

$$SE_0(\hat{p}) = \sqrt{\frac{0.1(1-0.1)}{500}} = 0.013$$

.

The standardized statistic for this new sample is calculated below:

$$t = \frac{0.136 - 0.1}{0.013} = 2.769$$

19. Mark the t-value of the original standardized statistic from question 9 and the value of the standardized statistic calculated with $\hat{p} = 0.136$ on the standard normal distribution below.

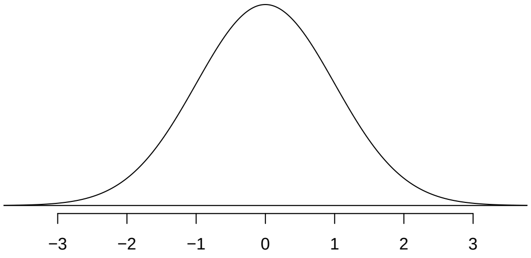


Figure 4.3: Standard Normal Curve

20. How does a statistic closer to the null value affect the p-value?

21. Summarize how each of the following affected the p-value:

a) Switching to a two-sided test.

b) Using a smaller sample size.

c) Using a sample statistic closer to the null value.

4.3.5 Take-home messages

1. Both simulation and theory-based methods can be used to find a p-value for a hypothesis test. In order to use theory-based methods we need to check that both the independence and the success-failure conditions are met.
2. The standardized statistic measures how many standard errors the statistic is from the null value. The larger the standardized statistic the more evidence there is against the null hypothesis.
3. The p-value for a two-sided test is approximately two times the value for a one-sided test. A two-sided test provides less evidence against the null hypothesis.
4. The larger the sample size, the smaller the sample to sample variability. This will result in a larger standardized statistic and more evidence against the null hypothesis.
5. The farther the statistic is from the null value, the larger the standardized statistic. This will result in a smaller p-value and more evidence against the null hypothesis.

4.3.6 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

4.4 Activity 10: Confidence interval and what confidence means

4.4.1 Learning outcomes

- Explore what confidence means
- Interpret the confidence level
- Explore impact of sample size, direction of the alternative hypothesis, and value of the sample statistic on the p-value.

4.4.2 Terminology review

In this activity, we will explore what being 95% confidence means. Some terms covered in this activity are:

- Parameter of interest
- Two-sided vs. one-sided tests
- Confidence level

4.4.3 Handedness of male boxers continued

We will use the male boxer study to look at what confidence means.

Left-handedness is a trait that is found in about 10% of the general population. Past studies have shown that left-handed men are over-represented among professional boxers (Richardson and Gilman 2019). The fighting claim states that left-handed men have an advantage in competition. In this random sample of 500 male professional boxers, we want to see if there is an over-prevalence of left-handed fighters. In the sample of 500 male boxers, 81 were left-handed.

```
 # Read in data set
boxers <- read.csv("https://math.montana.edu/courses/s216/data/Male_boxers_sample.csv")
boxers %>% count(Stance)  # Count number in each Stance category
```

```
#>         Stance   n
#> 1  left-handed  81
#> 2 right-handed 419
```

What does *confidence* mean?

In the interpretation of a 95% confidence interval, we say that we are 95% confident that the parameter is within the confidence interval. Why are we able to make that claim? What does it mean to say “we are 95% confident”?

1. In the last activity we found very strong evidence that the true proportion of male professional boxers that are left-handed is greater than 0.1. As a class, determine a plausible value for the true proportion of male boxers that are left-handed. *Note: we are making assumptions about the population here. This is not based on our calculated data, but we will use this applet to better understand what happens when we take many, many samples from this believed population.*

2. Go to this website, http://www.rossmanchance.com/ISIapplets.html and choose ‘Simulating Confidence Intervals’. In the input on the left-hand side of the screen enter the value from question 1 for π (the true value), 500 for n, and 100 for ‘Number of intervals’. Click ‘sample’.

- In the graph on the bottom right, click on a green dot. Write down the confidence interval for this sample given on the graph on the left. Does this confidence interval contain the true value chosen in question 1?

- Now click on a red dot. Write down the confidence interval for this sample. Does this confidence interval contain the true value chosen in question 1?

- How many intervals out of 100 contain π, the true value chosen in question 1? *Hint*: This is given to the left of the graph of green and red intervals.

3. Click on 'sample' nine more times. Write down the 'Running Total' for the proportion of intervals that contain π.

4. Change the confidence level to 90%. What happened to the width of the intervals?

5. Write down the `Running Total` for the proportion of intervals that contain π using a 90% confidence level.

6. Interpret the level of confidence. *Hint*: What proportion of samples would we expect to give a confidence interval that contains the parameter of interest?

Theory-based confidence interval

To calculate a theory-based 95% confidence interval for π, we will first find the **standard error** of $\hat{p}$ by plugging in the value of $\hat{p}$ for π in $SD(\hat{p})$:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

Note that we do not include a "0" subscript, since we are not assuming a null hypothesis.

7. Calculate the standard error of the sample proportion to find a 95% confidence interval.

We will calculate the margin of error and confidence interval in questions 10 and 11 of this activity. **The margin of error (ME)** is the value of the z^* multiplier times the standard error of the statistic.

$$ME = z^* \times SE(\hat{p})$$

The z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 95%, we find the Z values that encompass the middle 95% of the standard normal distribution. If 95% of the standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail.

The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, $95\% + 2.5\% = 97.5\%$ percentile).

- Enter the value of 0.975 for xx in the provided R script file.
- Highlight and run line 12. This will give the value of the multiplier for a 95% confidence interval.

```r
qnorm(xx, lower.tail = TRUE) # Multiplier for 95% confidence interval
```

8. Report the value of the multiplier needed to calculate the 95% confidence interval for the true proportion of male boxers that are left-handed.

9. Fill in the normal distribution shown below to show how R found the z^* multiplier.

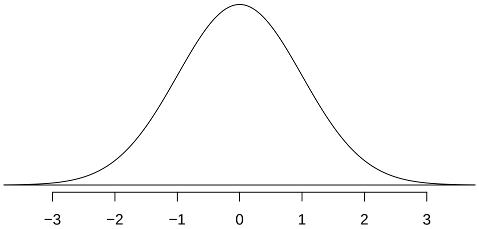


Figure 4.4: Standard Normal Curve

10. Calculate the margin of error for the 95% confidence interval.

To find the confidence interval, we will add and subtract the **margin of error** to the point estimate:

$$\text{point estimate} \pm \text{margin of error}$$

$$\hat{p} \pm z^* \times SE(\hat{p})$$

11. Calculate the 95% confidence interval for the parameter of interest.

12. Interpret the 95% confidence **interval** in the context of the problem.

4.4.4 Take-home messages

1. If repeat samples of the same size are selected from the population, approximately 95% of samples will create a 95% confidence interval that contains the parameter of interest.
2. The calculation of the confidence interval uses the standard error calculated using the sample proportion rather than the null value.

4.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

4.5 Module 3 and 4 Lab: Mixed Breed Dogs in the U.S.

4.5.1 Learning outcomes

- Determine whether simulation or theory-based methods of inference can be used.
- Analyze and interpret a study involving a single categorical variable.

4.5.2 Mixed Breed Dogs in the U.S.

The American Veterinary Medical Association estimated in 2010 that approximately 49% of dog owners in the U.S. own dogs that are classified as "mixed breed." As part of a larger 2022 international study (Banton 2022) about overall dog health, survey participants were asked, among other things, to report whether their dog was purebred or a mixed breed. Seven hundred and fifty (750) dog owners from the U.S. were recruited to complete an online survey via an email indicating they had been randomly selected by Qualtrics (an "experience management" company that specializes in surveys). Three hundred sixty-four (364) out of 675 respondents from the U.S. reported they owned a mixed breed dog. Is there evidence that, in the last decade, the proportion of dog owners in the U.S. that own a mixed breed dog has changed from the value reported in 2010?

Activity intro

- Download the R script file and the data file (US_dogs.csv) from D2L
- Upload both files to D2L and open the R script file
- Enter the name of the dataset for datasetname.csv.
- Highlight and run lines 1 - 6

1. What is the value of the point estimate?

2. Create a plot of the data using the R code. Make sure to include an appropriate title with type of plot, observational units, and variable. **Upload the plot to Gradescope**.

```r
dogs %>% # Data set piped into...
    ggplot(aes(x = variable)) +   # This specifies the variable
    geom_bar(aes(y = after_stat(prop), group = 1)) +  # Tell it to make a bar plot with proportions
    labs(title = "Don't forget to title your plot",
       # Give your plot a title
       x = "Breed of Dog",    # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```

Use statistical analysis methods to draw inferences from the data

3. **Write out the parameter of interest in words, in context of the study.**

4. Write out the null and alternative hypotheses in notation.

H_0 :

H_A :

5. **Will theory-based methods give the sample results as simulation based methods? Explain your answer.**

To use the computer simulation, we will need to enter the * assumed "probability of success" (π_0), * "sample size" (the number of observational units or cases in the sample), * "number of repetitions" (the number of samples to be generated), * "as extreme as" (the observed statistic), and * the "direction" (matches the direction of the alternative hypothesis).

We will use the `one_proportion_test()` function in `R` (in the `catstats` package) to simulate the null distribution of sample proportions and compute a p-value. Using the provided `R` script file, fill in the values/words for each `xx` with your answers from question 5 in the one proportion test to create a null distribution with 1000 simulations. Then highlight and run lines 21–26.

```
one_proportion_test(probability_success = xx, # Null hypothesis value
          sample_size = xx, # Enter sample size
          number_repetitions = 1000, # Enter number of simulations
          as_extreme_as = xx, # Observed statistic
          direction = "xx", # Specify direction of alternative hypothesis
          summary_measure = "proportion") # Reporting proportion or number of successes?
```

6. Report the p-value from the study.

The z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level.

- Enter the value of the appropriate percentile for xx in the provided R script file to find the multiplier for a 90% confidence interval.
- Highlight and run line 17

```
qnorm(xx. lower.tail = TRUE) # Multiplier for 90% confidence interval
```

7. **Calculate the margin of error for a 90% confidence interval.**

8. Calculate a 90% confidence interval.

Summarize the results of the study

9. Write a paragraph summarizing the results of the study. Be sure to describe:

- Summary statistic and interpretation
  - Summary measure (in context)
  - Value of the statistic
  - Order of subtraction when comparing two groups
- P-value and interpretation
  - Statement about probability or proportion of samples

  - Statistic (summary measure and value)
  - Direction of the alternative
  - Null hypothesis (in context)
- Confidence interval and interpretation
  - How confident you are (e.g., 90%, 95%, 98%, 99%)
  - Parameter of interest
  - Calculated interval
  - Order of subtraction when comparing two groups
- Conclusion (written to answer the research question)
  - Amount of evidence
  - Parameter of interest
  - Direction of the alternative hypothesis
- Scope of inference
  - To what group of observational units do the results apply (target population or observational units similar to the sample)?
  - What type of inference is appropriate (causal or non-causal)?

Upload a copy of your group's paragraph to Gradescope.

Paragraph (continued):

MODULE 5

Unit 1 Review

The following module contains both a list of key topics covered in Unit 1 as well as Module Review Worksheets that will be covered in Weekly Review Sessions.

5.1 Module Review

The following worksheets review each of the modules. These worksheets will be completed during Melinda's Study Sessions each week. Solutions will be posted on D2L in the Unit 1 Review folder after the study sessions.

5.2 Key Topics

Review the key topics for Unit 1 prior to the first exams. All of these topics will be covered in Modules 1 - 4.

5.3 Module 1 Review

1. Suppose that the proportion of all American adults that fit the medical definition of being obese is 0.23. A large medical clinic would like to determine if the proportion of their patients that are obese is higher than that of all American adults. The clinic takes a simple random sample of 30 of their patients and finds that 9 patients in the sample are obese.

a. What is the target population?

b. What are the observational units?

c. What variable is being studied?

d. Is the variable identified in part (c) categorical or quantitative?

2. Martha works in Macy's advertising department. She is interested in the shopping experience of all Macy's shoppers in the U.S. Every Saturday morning for a month she stands outside of the Bozeman Macy's asking people about their experience. One of the questions she uses is: "As a huge fan of Macy's, I believe Macy's has the best choices of clothing in Bozeman. Don't you agree?" Every person that was asked, responded.

a. Identify the target population.

b. Identify the sample.

c. Which of the three types of sampling bias (selection, non-response, response) may be present? Explain your choice(s).

5.4 Module 2 Review

1. Spelling errors in a text can either be non-word errors (teh instead of the) or word errors (lose instead of loose). It was found that non-word errors make up about 25% of all errors. A human proofreader will catch 92% of non-word errors and 75% of word errors.

Let N represent non-word errors and C represent that a human proofreader will catch the error.

a. Identify the following values with appropriate probability notation.

0.25

0.92

0.75

b. Fill in the table below to represent the situation:

	N	N^C	Total
C			
C^C			
Total			100000

c. Using your table calculate the probability that a randomly selected error caught by a human proofreader is a non-word error. Use appropriate probability notation.

d. Find the probability a selected error is a non-word error and was not caught by a human proofreader. Use appropriate probability notation.

e. Find the value of $P(N|C)$. What does this probability mean?

2. A private college report contains these statistics:

- 70% of incoming freshmen attended public schools
- 75% of public-school students who enroll as freshmen eventually graduate
- 90% of other freshmen eventually graduate

Let A represent the event that a freshman attended public school and B the event that a freshman eventually graduates.

a. Identify the following values with appropriate probability notation.

```
 = 0.70

= 0.75

= 0.90
```

b. Fill in the table below to represent the situation:

	A	A^C	Total
B			
B^C			
Total			100000

c. Calculate the probability a selected freshman attended public school given they did not graduate. Use appropriate probability notation.

d. Calculate the probability a selected freshman does not graduate. Use appropriate probability notation.

e. Of the population of freshman that attended public school, what is the probability they do not graduate. Use appropriate probability notation.

f. Find the value of $P(A \text{and} B^C)$. Write this probability in context of the problem.

5.5 Module 3 Review - Simulation Methods

```
hearing <- read.csv("data/hearing_loss.csv")
```

A recent study examined hearing loss data for 1753 U.S. teenagers. In this sample, 328 were found to have some level of hearing loss. News of this study spread quickly, with many news articles blaming the prevalence of hearing loss on the higher use of ear buds by teens. At MSNBC.com (8/17/2010), Carla Johnson summarized the study with the headline: "1 in 5 U.S. teens has hearing loss, study says." Is this an appropriate or a misleading headline?

1. Write the parameter of interest in context of the study.

2. Write the null hypothesis in words and notation in context of the problem.

3. Based on the research questions, choose the direction for the alternative hypothesis.

4. Write the alternative hypothesis in words and notation in context of the problem.

5. Calculate the summary statistic. Use proper notation.

6. What values should be entered for each of the following into the one proportion test to create 1000 simulations?

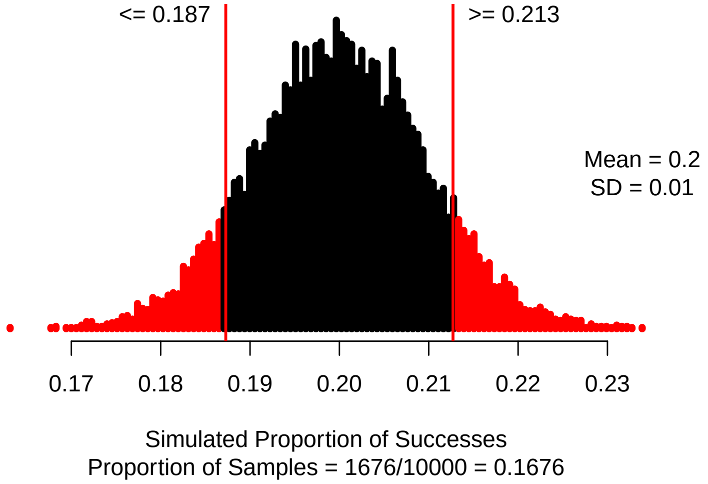
- Probability of success:

- Sample size:

- Number of repetitions:

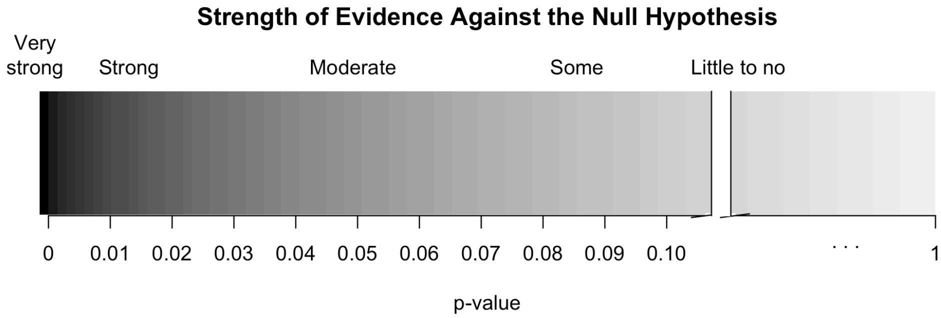
- As extreme as:

- Direction ("greater", "less", or "two-sided"):



7. Interpret the p-value in context of the problem.

8. How much evidence does the data provide against the null hypothesis?



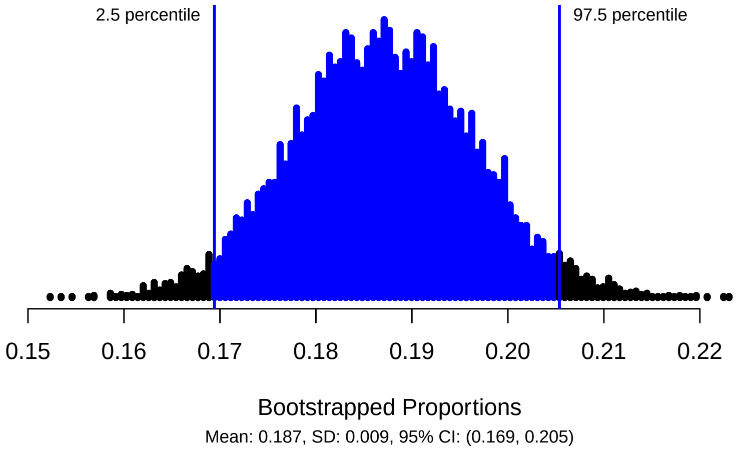
9. Write a conclusion to the study in context of the problem.

10. Would a 95% confidence interval contain the null value of 0.2? Explain.

11. What values should be entered for each of the following into the simulation to create the bootstrap distribution of sample proportions to find a 95% confidence interval?

- Sample size:
- Number of successes:
- Number of repetitions:
- Confidence level (as a decimal):

```r
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 1753, # Sample size
                            number_successes = 328, # Observed number of successes
                            number_repetitions = 10000, # Number of bootstrap samples to use
                            confidence_level = 0.95) # Confidence level as a decimal
```



12. Explain how to use cards to create one bootstrap sample.

13. Report the 95% confidence interval in interval notation.

14. Interpret the 95% confidence interval in context of the problem.

5.6 Module 4 Review

Statistician Jessica Utts has conducted an extensive analysis of Ganzfeld studies that have investigated psychic functioning. Ganzfeld studies involve a "sender" and a "receiver." Two people are placed in separate rooms. The sender looks at a "target" image on a television screen and attempts to transmit information about the target to the receiver. The receiver is then shown four possible choices or targets, one of which is the correct target and the other three are "decoys." The receiver must choose the one he or she thinks best matches the description transmitted by the sender. If the correct target is chosen by the receiver, the session is a "hit." Otherwise, it is a miss. Utts reported that her analysis considered a total of 2,124 sessions and found a total of 709 "hits" (Utts, 2010). Is there evidence of psychic ability?

1. Write the parameter of interest in context of the study.

2. Calculate the point estimate. Use proper notation.

3. Write the null hypothesis in words.

4. Write the alternative hypothesis in notation.

A single proportion can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sample distribution of $\hat{p}$.

- Independence: The sample's observations are independent, e.g., are from a simple random sample
- Large enough sample size:
    - Success-Failure Condition: There are at least 10 successes and 10 failures in the sample

$$n \times \hat{p} \geq 10$$

and

$$n \times (1 - \hat{p}) \geq 10$$

5. Are the conditions met to model the data with the Normal distribution?

Standardized sample proportion.

The standardized statistic for theory-based methods for one proportion is:

$$Z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})}$$

Where

$$SE_0(\hat{p}) = \sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}$$

6. Calculate the null standard error of the sample proportion

7. Calculate the standardized statistic for the sample proportion.

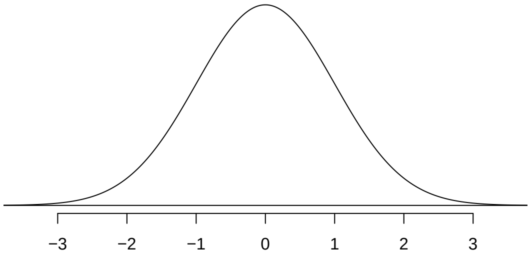
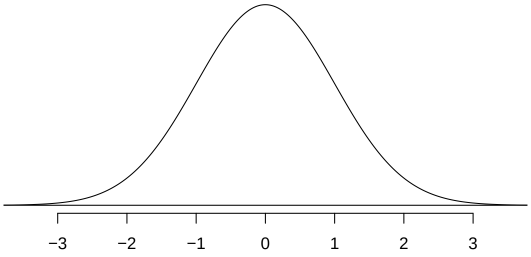


Figure 5.1: A standard normal curve.



8. Interpret the standardized statistic in context of the problem.

We will use the `pnorm()` function in `R` to find the p-value. The value of the standardized statistic calculated in question 8 is entered into the `R` code. We used `lower.tail = FALSE` to find the p-value so that `R` will calculate the p-value *greater* than the value of the standardized statistic.

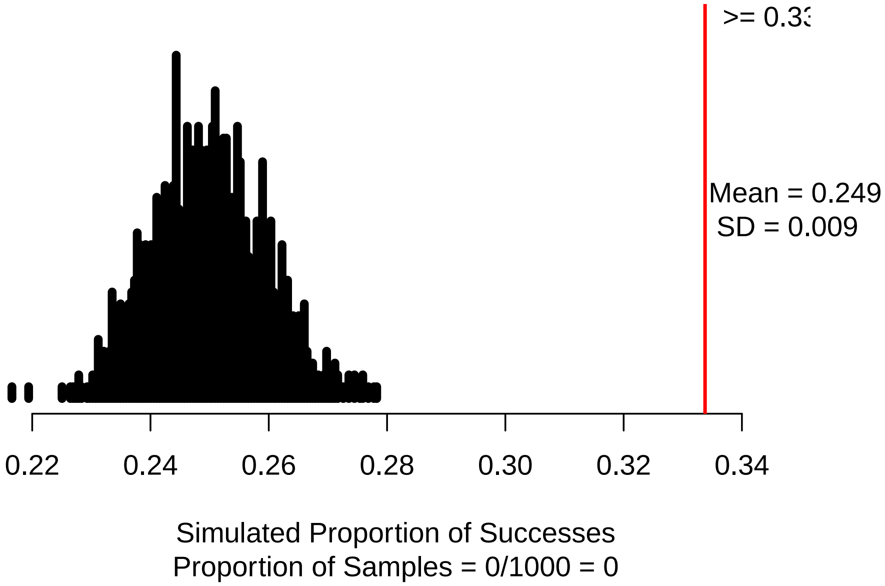
Notes:

- Use `lower.tail = TRUE` when doing a left-sided test.
- Use `lower.tail = FALSE` when doing a right-sided test.
- To find a two-sided p-value, use a left-sided test for negative Z or a right-sided test for positive Z, then multiply the value found by 2 to get the p-value.

```
pnorm(9.333, # Enter value of standardized statistic
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
      lower.tail=FALSE) # Gives a p-value greater than the standardized statistic
#> [1] 5.145792e-21
```

9. Report the value of the p-value.

Simulation Method:



10. Interpret the p-value in context of the study.

Next we will use theory-based methods to estimate the parameter of interest.

To calculate a theory-based 95% confidence interval for π, we will first find the **standard error** of $\hat{p}$ by plugging in the value of $\hat{p}$ for π in $SD(\hat{p})$:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}.$$

Note that we do not include a "0" subscript, since we are not assuming a null hypothesis.

11. Calculate the standard error of the sample proportion to find a 95% confidence interval.

To find the confidence interval, we will add and subtract the **margin of error** to the point estimate:

$$\text{point estimate} \pm \text{margin of error}$$

$$\hat{p} \pm z^* SE(\hat{p})$$

The z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 95%, we find the Z values that encompass the middle 95% of the standard normal distribution. If 95% of the standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail. The `qnorm()` function in `R` will tell us the z^* value for the desired percentile (in this case, 95% + 2.5% = 97.5% percentile).

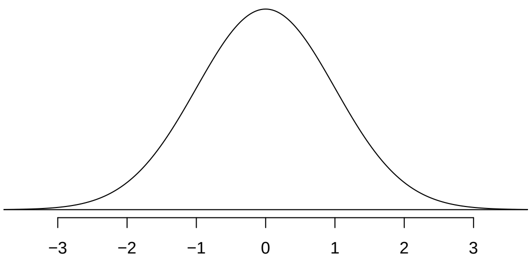


Figure 5.2: A standard normal curve.

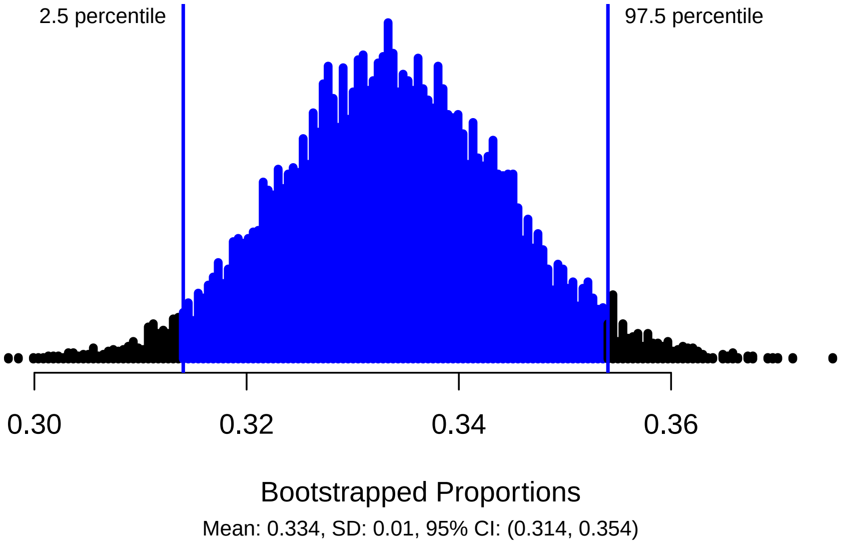
```
qnorm(0.975) # Multiplier for 95% confidence interval
```

```
#> [1] 1.959964
```

12. Calculate the margin of error for a 95% confidence interval for the true proportion of sessions that will result in a hit.

13. Calculate the 95% confidence interval for the true proportion of sessions that will result in a hit.

Simulation Methods:



14. Interpret the 95% confidence interval in context of the problem.

15. Write a conclusion based on the p-value and the 95% confidence interval.

5.7 Unit 1 Review

5.8 Key Topics Exam 1

Descriptive statistics and study design:

1. Identify the observational units.
2. Identify the types of variables (categorical or quantitative).
3. Identify the explanatory variable (if present) and the response variable (roles of variables).
4. Identify the appropriate type of graph and summary measure.
5. Identify if a given value is a statistic or a parameter. Identify the appropriate notation.
6. Identify the study design (observational study or randomized experiment).
7. Identify the sampling method and potential types of sampling bias (non-response, response, selection).
8. Identify and interpret the summary statistic
9. Identify the target population
10. Identify the types of sampling bias (response, non-response, selection, none)
11. Identify the type(s) of graph(s) that could be used to plot the given variable(s).

Hypothesis testing:

12. Write the parameter of interest in context of the problem.
13. State the null and alternative hypotheses in both words and notation
14. Verify the validity condition is met to use simulation-based methods to find a p-value.
15. Verify the validity conditions are met to use theory-based methods to find a p-value from the theoretical distribution.
16. In a simulation-based hypothesis test, describe how to create one dot on a dotplot of the null distribution using coins, cards, or spinners.
17. Explain where the null distribution is centered and why.
18. Describe and illustrate how R calculates the p-value for a simulation-based test.
19. Describe and illustrate how R calculates the p-value for a theory-based test.
20. Type of theoretical distribution (standard normal distribution or t-distribution with appropriate degrees of freedom) used to model the standardized statistic in a theory-based hypothesis test.
21. Calculate and interpret the standard error of the statistic under the null using the correct formula on the Golden ticket.
22. Calculate and interpret the appropriate standardized statistic using the correct formula on the Golden ticket.
23. Interpret the p-value in context of the study: it is the probability of ______, assuming ______.
24. Evaluate the p-value for strength of evidence against the null: how much evidence does the p-value provide against the null?
25. Write a conclusion about the research question based on the p-value.
26. Describe which features of the study impact the p-value and how.

Confidence interval:

27. Describe how to simulate one bootstrapped sample using cards.
28. Explain where the bootstrap distribution is centered and why.
29. Find an appropriate percentile confidence interval using a bootstrap distribution from R output.
30. Verify the validity condition is met to use simulation-based methods to find the confidence interval.
31. Verify the validity conditions are met to use theory-based methods to calculate a confidence interval.
32. Describe and illustrate how the bootstrap distribution is used to find the confidence interval for a given confidence level.
33. Describe and illustrate how the standard normal distribution or t-distribution is used to find the multiplier for a given confidence level.
34. Calculate and interpret the standard error of the statistic (not assuming the null hypothesis) using the correct formula on the Golden ticket
35. Calculate the appropriate margin of error and confidence interval using theory-based methods.
36. Interpret the confidence interval in context of the study.
37. Based on the interval, what decision can you make about the null hypothesis? Does the confidence interval agree with the results of the hypothesis test? Justify your answer.
38. Interpret the confidence level in context of the study. What does “confidence” mean?
39. Describe which features of the study have an effect on the width of the confidence interval and how.

References

“Average Driving Distance and Fairway Accuracy.” 2008. https://www.pga.com/ and https://www.lpga.com/.

Banton, et al, S. 2022. “Jog with Your Dog: Dog Owner Exercise Routines Predict Dog Exercise Routines and Perception of Ideal Body Weight.” *PLoS ONE* 17(8).

Bhavsar, et al, A. 2022. “Increased Risk of Herpes Zoster in Adults 50 Years Old Diagnosed with COVID-19 in the United States.” *Open Forum Infectious Diseases* 9(5).

Bulmer, M. n.d. “Islands in Schools Project.” https://sites.google.com/site/islandsinschoolsprojectwebsite/home.

“Bureau of Transportation Statistics.” 2019. https://www.bts.gov/.

“Child Health and Development Studies.” n.d. https://www.chdstudies.org/.

Darley, J. M., and C. D. Batson. 1973. “”From Jerusalem to Jericho”: A Study of Situational and Dispositional Variables in Helping Behavior.” *Journal of Personality and Social Psychology* 27: 100–108.

Davis, Smith, A. K. 2020. “A Poor Substitute for the Real Thing: Captive-Reared Monarch Butterflies Are Weaker, Paler and Have Less Elongated Wings Than Wild Migrants.” *Biology Letters* 16.

Du Toit, et al, G. 2015. “Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy.” *New England Journal of Medicine* 372.

Edmunds, et al, D. 2016. “Chronic Wasting Disease Drives Population Decline of White-Tailed Deer.” *PLoS ONE* 11(8).

Education Statistics, National Center for. 2018. “IPEDS.” https://nces.ed.gov/ipeds/.

“Great Britain Married Couples: Great Britain Office of Population Census and Surveys.” n.d. https://discovery.nationalarchives.gov.uk/details/r/C13351.

Group, TODAY Study. 2012. “A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.” *New England Journal of Medicine* 366: 2247–56.

Hamblin, J. K., K. Wynn, and P. Bloom. 2007. “Social Evaluation by Preverbal Infants.” *Nature* 450 (6288): 557–59.

Hirschfelder, A., and P. F. Molin. 2018. “I Is for Ignoble: Stereotyping Native Americans.” Retrieved from https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm.

Hutchison, R. L., and M. A. Hirthler. 2013. “Upper Extremity Injuies in Homer’s Iliad.” *Journal of Hand Surgery (American Volume)* 38: 1790–93.

“IMDb Movies Extensive Dataset.” 2016. https://kaggle.com/stefanoleone992/imdb-extensive-dataset.

Kalra, et al., Dl. 2022. “Trustworthiness of Indian Youtubers.” Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4426566.

Keating, D., N. Ahmed, F. Nirappil, Stanley-Becker I., and L. Bernstein. 2021. “Coronavirus Infections Dropping Where People Are Vaccinated, Rising Where They Are Not, Post Analysis Finds.” *Washington Post*. https://www.washingtonpost.com/health/2021/06/14/covid-cases-vaccination-rates/.

Laeng, Mathisen, B. 2007. “Why Do Blue-Eyed Men Prefer Women with the Same Eye Color?” *Behavioral Ecology and Sociobiology* 61(3).

Levin, D. T. 2000. “Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit.” *Journal of Experimental Psychology* 129(4).

Madden, et al, J. 2020. “Ready Student One: Exploring the Predictors of Student Learning in Virtual Reality.” *PLoS ONE* 15(3).

Miller, G. A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.” *Psychological Review* 63(2).

Moquin, W., and C. Van Doren. 1973. “Great Documents in American Indian History.” Praeger.

“More Americans Are Joining the ’Cashless’ Economy.” 2022. https://www.pewresearch.org/short-reads/2022/10/05/more-americans-are-joining-the-cashless-economy/.

National Weather Service Corporate Image Web Team. n.d. “National Weather Service – NWS Billings.” https://w2.weather.gov/climate/xmacis.php?wfo=byz.

O’Brien, Lynch, H. D. 2019. “Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms.” *Integrative Organismal Biology* 1.

“Ocean Temperature and Salinity Study.” n.d. https://calcofi.org/.

“Older People Who Get Covid Are at Increased Risk of Getting Shingles.” 2022. https://www.washingtonpost.com/health/2022/04/19/shingles-and-covid-over-50/.

“Physician’s Health Study.” n.d. https://phs.bwh.harvard.edu/.

Porath, Erez, C. 2017. “Does Rudeness Really Matter? The Effects of Rudeness on Task Performance and Helpfulness.” *Academy of Management Journal* 50.

Quinn, G. E., C. H. Shin, M. G. Maguire, and R. A. Stone. 1999. “Myopia and Ambient Lighting at Night.” *Nature* 399 (6732): 113–14. https://doi.org/10.1038/20094.

Ramachandran, V. 2007. “3 Clues to Understanding Your Brain.” https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.

“Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status.” 2021. CDC. https://covid.cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination.

Richardson, T., and R. T. Gilman. 2019. “Left-Handedness Is Associated with Greater Fighting Success in Humans.” *Scientific Reports* 9 (1): 15402. https://doi.org/10.1038/s41598-019-51975-3.

Stephens, R., and O. Robertson. 2020. “Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel ”Swear” Words.” *Frontiers in Psychology* 11: 643–62.

Stewart, E. H., B. Davis, B. L. Clemans-Taylor, B. Littenberg, C. A. Estrada, and R. M. Centor. 2014. “Rapid Antigen Group a Streptococcus Test to Diagnose Pharyngitis: A Systematic Review and Meta-Analysis” 9 (11). https://doi.org/10.1371/journal.pone.0111727.

Stroop, J. R. 1935. “Studies of Interference in Serial Verbal Reactions.” *Journal of Experimental Psychology* 18: 643–62.

Subach, et al, A. 2022. “Foraging Behaviour, Habitat Use and Population Size of the Desert Horned Viper in the Negev Desert.” *Soc.Open Sci* 9.

Sulheim, S., A. Ekeland, I. Holme, and R. Bahr. 2017. “Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders: Changes After an Interval of One Decade” 51 (1): 44–50. https://doi.org/10.1136/bjsports-2015-095798.

“Titanic.” n.d. http://www.encyclopedia-titanica.org.

“US COVID-19 Vaccine Tracker: See Your State’s Progress.” 2021. Mayo Clinic. https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker.

US Environmental Protection Agency. n.d. “Air Data – Daily Air Quality Tracker.” https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker.

Wahlstrom, et al, K. 2014. “Examining the Impact of Later School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study.” *Center for Applied Research and Educational Improvement.*

Weiss, R. D. 1988. “Relapse to Cocaine Abuse After Initiating Desipramine Treatment.” *JAMA* 260(17).

“Welcome to the Navajo Nation Government: Official Site of the Navajo Nation.” 2011.Retrieved from https://www.navajo-nsn.gov/.

Wilson, Woodruff, J. P. 2016. “Vertebral Adaptations to Large Body Size in Theropod Dinosaurs.” *PLoS ONE* 11(7).