STAT 216 Coursepack

Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

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.

Basics of Data and Sampling Methods

1.1 Vocabulary Review and Key Topics

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

1.1.1 Key topics

Module 1 introduces the foundations of data: observational units, types of variables, and how to collect sample data from a population of interest in a way that allows us to generalize our results back to the population.

1.1.2 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.
- **Sample size**: the number of observational units in a data set, denoted by n.
- **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 Q12 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

Count (y-axis, 0 to 180)
Sample means (x-axis, 2 to 8)
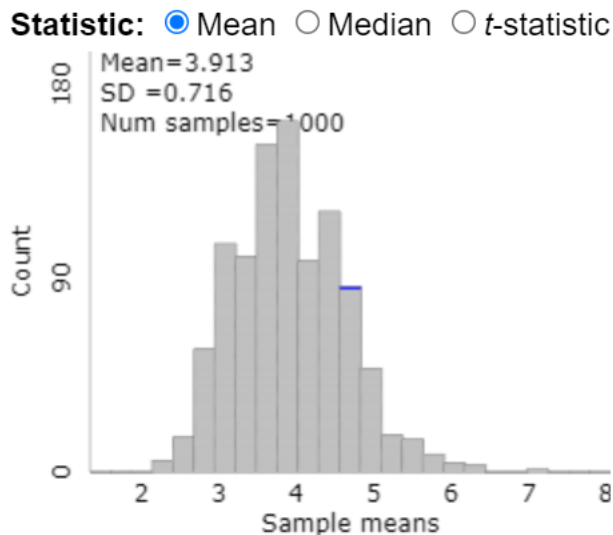
15. What is the center value (mean) of the distribution displayed above?

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 the 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

Mean=3.954
SD =0.411
Num samples=1000

(y-axis: Count, markings 0, 90, 18)

(x-axis: Sample means, markings 2.7, 3.6, 4.5, 5.4)
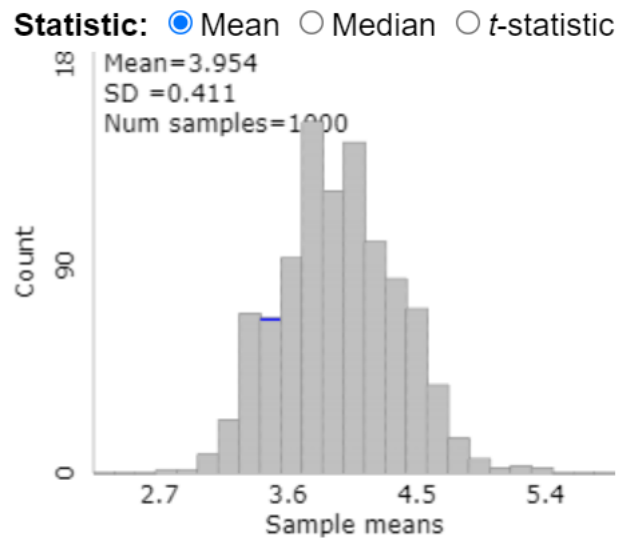
19. Compare the distribution displayed before question 15 to the one shown above.

    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.

Probability

2.1 Vocabulary Review and Key Topics

2.1.1 Key topics

Module 2 introduces the concept of probability as a long-run relative frequency and demonstrates how to use hypothetical two-way tables to set up a probability problem and solve for unconditional and conditional probabilities.

2.1.2 Vocabulary

- **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 23 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.
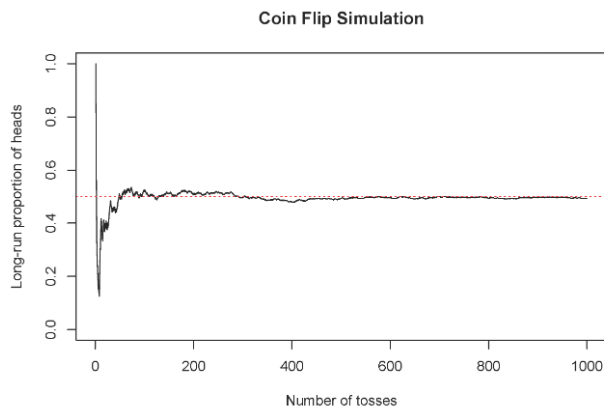
In today's activity we will discuss the probability of a single event, the probability of an "and" event, and the probability of a conditional event.

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
  - $P(A^C) = 1 - P(A)$
- $P(A \, and \, B)$ 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 survives.

   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 $P(B)$. 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.

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.

3.1.1 Key topics

Module 3 introduces the steps of the statistical investigation process. We conduct **exploratory data analysis** (summary statistics and plots) and simulation-based **inference** (hypothesis testing and confidence intervals) in the single categorical variable (one proportion) scenario.

- 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

Exploratory data analysis is step 3 of the statistical investigation process. We will then use simulation-based methods **to find evidence of an effect by finding a p-value** and **estimating how large the effect is by creating a confidence interval** in the one proportion (one categorical variable) scenario. These are 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.

3.1.2 Vocabulary

- **Summary measure**: a numerical quantity that summarizes 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.

- **Summary statistic (point estimate)**: the value of a numerical summary measure computed from *sample* data.

  - 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)

- For a single categorical variable, the category that we are counting the proportion of is generically called a "**success**", with categories not a success labeled "**failure**". Thus, a sample proportion is the "proportion of successes" in the sample: the total number of successes divided by the sample size (n).

Plotting one categorical variable

- **Frequency bar plot**: plots the count (frequency) of observational units in each level of a categorical variable. R code to create a frequency bar plot:

```
object %>% # 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 = "Don't forget to title your plot!",
   # Give your plot a title
   x = "x-axis label",   # Label the x axis
   y = "Frequency")  # Label the y axis
```

- **Relative frequency bar plot**: plots the proportion (relative frequency) of observational units in each level of a categorical variable. R code to create a relative frequency bar plot:

```
object %>% # 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 = "x-axis label",   # Label the x axis
   y = "Relative Frequency")  # Label the y axis
```

Inference

- **Sampling distribution** (of a statistic): the distribution of possible values of a statistic across repeated samples of the same size and under the same conditions.

  - We can create a *simulated* sampling distribution using simulation-based methods to simulate many samples, or we can mathematically model the sampling distribution (theory-based methods).

- **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 in notation for a single proportion**: In the hypotheses below, π_0 is the **null value**.

$$H_0 : \pi = \pi_0$$

$$H_A : \pi \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} \pi_0$$

- **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.
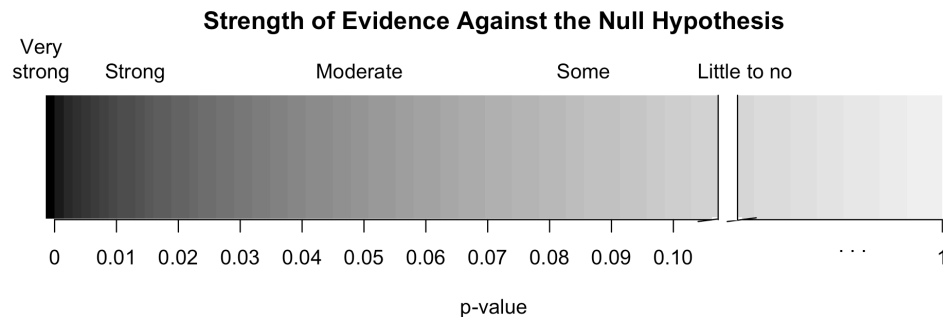
- **Conclusion** (to a hypothesis test): answers the research question. How much evidence is there in support of the alternative hypothesis?

  – 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; an interval of *plausible values* for the parameter.

  – 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

3.1.2.1 Simulation-based inference for a single proportion

- **Conditions necessary to use simulation-based methods for inference for a single categorical variable**:

  – **Independence**: observational units must be independent of one another; the outcome of one observational unit should have no influence on the outcome of another.

- **Null distribution**: a sampling distribution of simulated sample statistics created under the assumption that the null hypothesis is true

- **Simulation-based 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` (from the `catstats` package), is shown below.

```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?
```

- **Bootstrapping**: creating a simulated sample of the same size as the original sample by sampling with replacement from the original sample.

- **Simulation-based 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` (from the `catstats` package), is shown below.

```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 = 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 of the bootstrap distribution.
    - A 95% confidence interval will be found between the 2.5th and 97.5th percentiles of the bootstrap distribution.
    - A 99% confidence interval will be found between the 0.5th and 99.5th percentiles of the bootstrap distribution.

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_OneProp
- 4.2_OneProp
- Chapter9
- 14.1
- Chapter10
- 14.2

Summarizing categorical data - Video 4.1_OneProp

- 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:

Notation used for the sample proportion:

- Single categorical variable:

Categorical data can be reported in a _____table, which plots counts or a _____ frequency

table, which plots the proportion.

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.

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

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

```r
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_OneProp

- Types of plots for a single categorical variable

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

```r
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
```

Frequency Bar Plot of State of Origin for
Gallatin County Home Sales
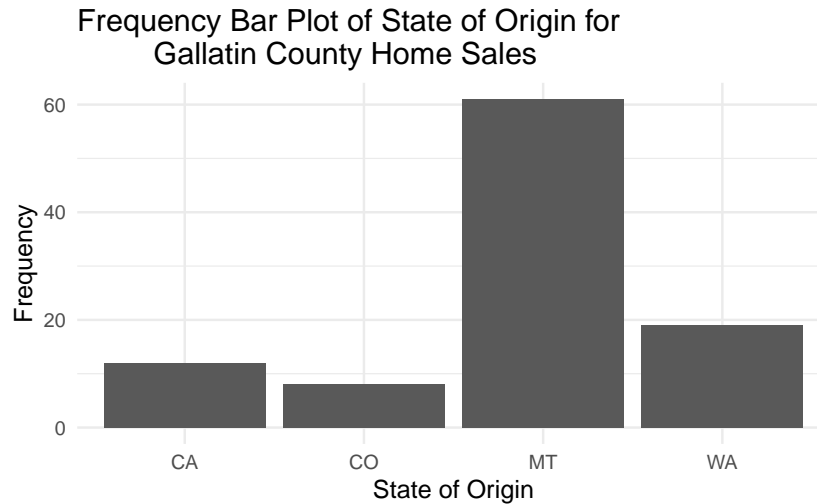
- What can we see from this plot?

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
```

Relative Frequency Bar Plot of State of Origin
for Gallatin County Home Sales
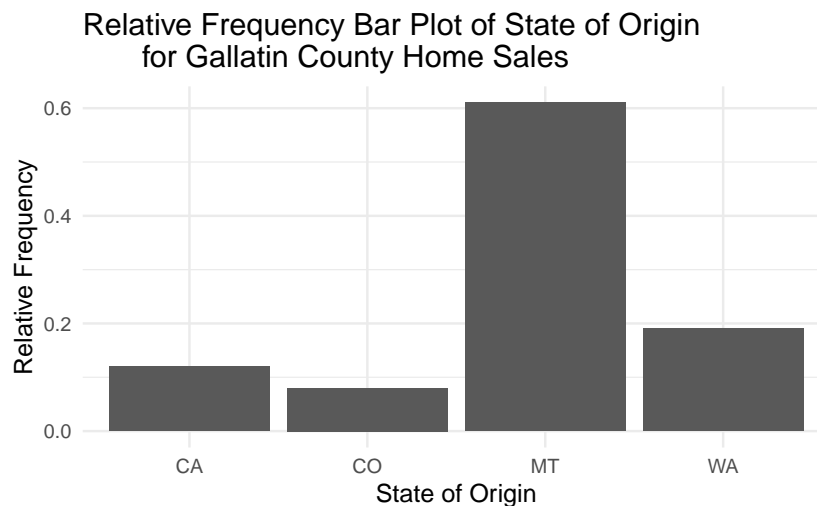
- Note: the x-axis is the _____ between the frequency bar plot and the relative frequency bar plot. However, the _____ differs. The scale for the frequency bar plot goes from _____ and the scale for the relative frequency bar plot is from _____.

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?

Strength of Evidence Against the Null Hypothesis

Very strong | Strong | Moderate | Some | Little to no

0   0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10   . . .   1
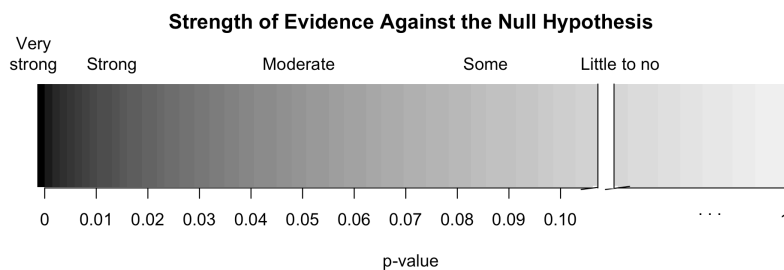
p-value

  - 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 10000 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 = 10000, # 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?
```


Simulated Proportion of Successes
Proportion of Samples = 69/10000 = 0.0069
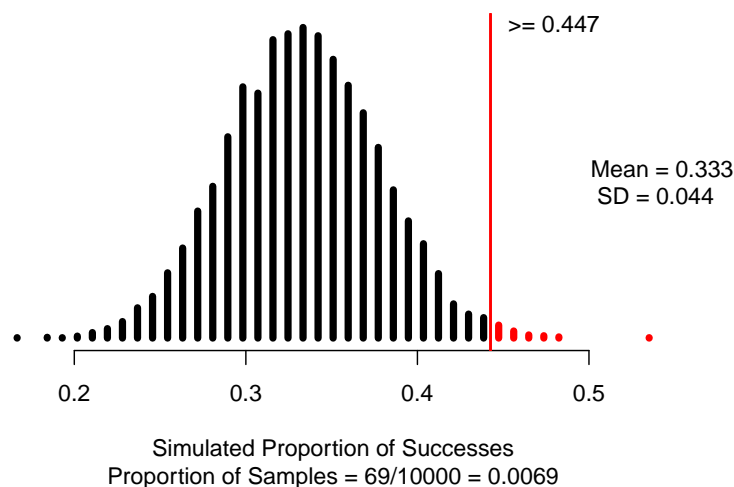
Explain why the null distribution is centered at the value of approximately 0.333:

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 $\pm$ 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 10000 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 = 10000, # Number of bootstrap samples to use
                    confidence_level = 0.90) # Confidence level as a decimal
```


Bootstrapped Proportions
Mean: 0.447, SD: 0.046, 90% CI: (0.368, 0.526)
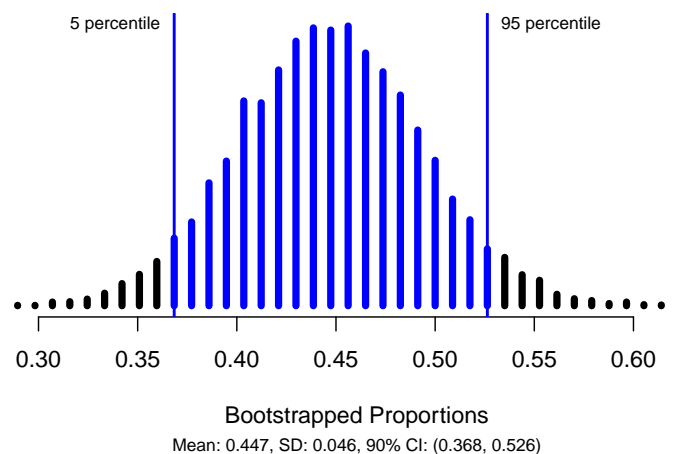
Confidence interval interpretation:

- 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 = 10000, # Number of bootstrap samples to use
                    confidence_level = 0.95) # Confidence level as a decimal
```


Bootstrapped Proportions
Mean: 0.447, SD: 0.046, 95% CI: (0.36, 0.535)

99% Confidence Interval:

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 114, # Sample size
                    number_successes = 51, # Observed number of successes
                    number_repetitions = 10000, # Number of bootstrap samples to use
                    confidence_level = 0.99) # Confidence level as a decimal
```


Bootstrapped Proportions
Mean: 0.447, SD: 0.046, 99% CI: (0.333, 0.57)

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.

```r
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

```r
# 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 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

```r
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
```
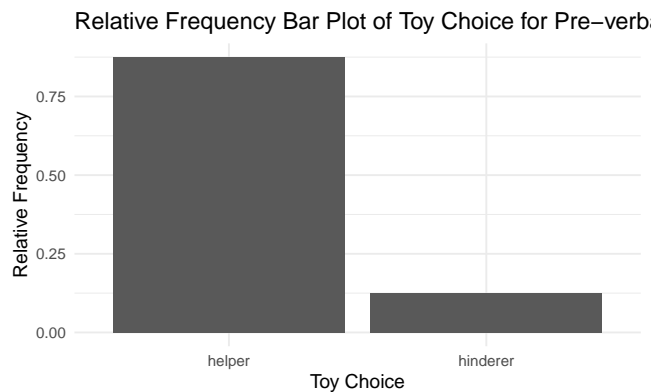

5. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

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 above.

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 10000 (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 11 closer to 0.5, the null value, or closer to the sample proportion, 0.875?

Report the number of coin flips you got as indicated by your instructor.

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 10000 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

In class today, we will revisit the study on infants as described below.

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 10000 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 10000 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 10000 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.)

Strength of Evidence Against the Null Hypothesis
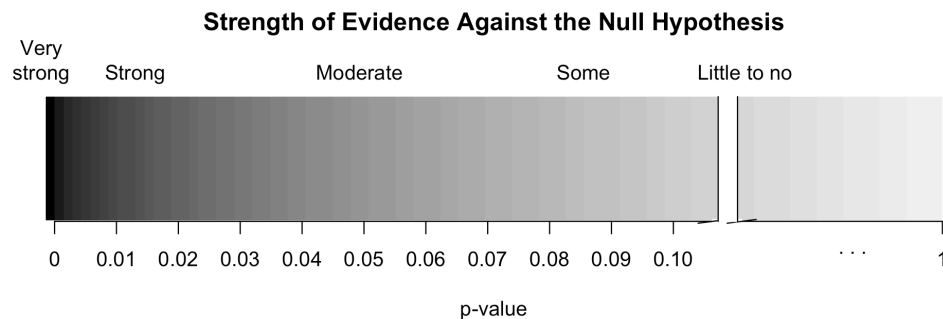
Interpret the p-value

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. The smaller the p-value of the test, the more evidence there is **against** the null hypothesis.

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 10000 simulations. Then highlight and run lines 1–9.

```
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 10000 simulations that follows. Circle the value that changed in the code.

```
one_proportion_bootstrap_CI(sample_size = 16, # Sample size
                    number_successes = 14, # Observed number of successes
                    number_repetitions = 10000, # Number of bootstrap samples to use
                    confidence_level = 0.90) # Confidence level as a decimal
```

5 percentile   95 percentile

0.5   0.6   0.7   0.8   0.9   1.0   1.1

Bootstrapped Proportions
Mean: 0.876, SD: 0.081, 90% CI: (0.75, 1)
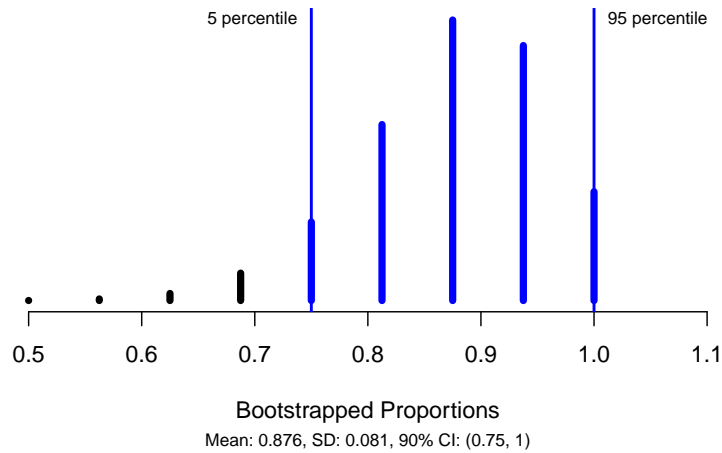
16. Report both the 95% confidence interval (question 8) and the 90% confidence interval (question 15). Is the 90% confidence interval narrower or wider than the 95% confidence interval?

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 10000 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

68

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.

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.

4.1.1 Key topics

Module 4 introduces theory-based inference methods (hypothesis testing and confidence intervals) for a single categorical variable. We also explore what "confidence level" means and which parts of a study impact the width of a confidence interval and the p-value.

- Theory-based methods should give the same results as simulation-based methods if the sample size is large enough. For a single categorical variable, the sample size is large enough if the success-failure condition is met.

- If repeated samples of the same size are taken from the population, 95% of samples will create a 95% confidence interval that contains the value of the parameter of interest.

4.1.2 Vocabulary

- **Theory-based methods**: when specific conditions are met, the distribution of sample statistics if we were to repeatedly sample from the population 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-based methods!)

  - **Success-failure condition**: we *expect* to see at least 10 successes and 10 failures in the sample, $n\pi \geq 10$ and $n(1 - \pi) \geq 10$. Since π is typically unknown, we consider this condition to be met if we observe at least 10 successes and 10 failures in our data set: $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$.

- **Standard normal distribution**: a theoretical distribution that is bell-shaped, centered on the mean of zero, and has a standard deviation of one, denoted in notation by $N(0, 1)$.

- **Standard error of a statistic**: an estimated standard deviation of the statistic as it would vary across repeated samples of the same size under the same conditions.

  - The standard error tells us about how far we would expect an observed sample statistic to fall from the true parameter value for which it is estimating, on average.

- **Standardized statistic**: calculation to standardize the sample statistic in order to compare the standardized value to the theoretical distribution.

  - Calculated by subtracting the null value from the sample statistic, then dividing by the standard error:
  $$\frac{\text{statistic} - \text{null value}}{\text{standard error}}$$

- – Measures the number of standard errors the sample statistic is above (if positive) or below (if negative) the null value.

- **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:

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

.

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

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

If the conditions for the sampling distribution of $\hat{p}$ to follow an approximate normal distribution are met, and if the true value of π is equal to the null value of π_0, the standardized sample proportion, Z, will have an approximate *standard* normal distribution.

- The theory-based **p-value** for hypothesis testing involving proportions can be found in R by using the `pnorm` function to find the probability of the observed standardized statistic or one more extreme (in the direction of H_A). This probability is the area under a *standard normal distribution* at or more extreme than the observed standardized statistic.

  - – Enter the value of the standardized statistic for `xx`.

  - – If a "greater than" alternative, change `lower.tail = TRUE` to `FALSE`.

  - – If a two-sided test, multiply by 2.

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

- **Margin of error**: half the width of the confidence interval. For a single proportion, the margin of error is:

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

where z^* is the **multiplier**, corresponding to the desired confidence level found from the standard normal distribution. For example, for a 95% confidence level, the middle 95% of the standard normal distribution falls between $-z^* = -1.96$ and $z^* = 1.96$.

- **Standard error of the sample proportion for a confidence interval** (not assuming the null is true):

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

- To find the endpoints of a confidence interval, add and subtract the margin of error to the sample statistic. The confidence interval for a population proportion is:

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

- R code to find the **multiplier** for the confidence interval using theory-based methods involving proportions.

  - – `qnorm` will give you the multiplier using the standard normal distribution.

  - – Enter the percentile for the given level of confidence (e.g., 0.975 for a 95% confidence level).

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

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

Read Chapter 11 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 _____.
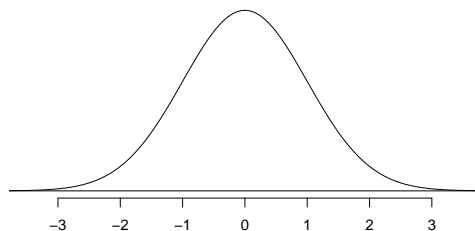
- $\hat{p} \sim N(\pi, \sqrt{\frac{\pi \times (1-\pi)}{n}})$

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

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

- 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.

```
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
```
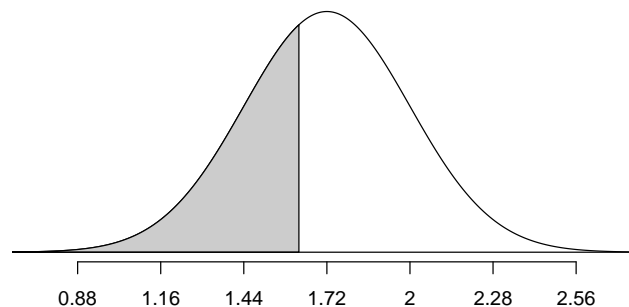

We can also reverse that order. Given a percentage, we can find the associated percentile, or quantile. Here we display calculating the value that cuts off the lower 0.75 proportion of male adult Caucasian heights using the qnorm() function.

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

| 0.88 | 1.16 | 1.44 | 1.72 | 2 | 2.28 | 2.56 |
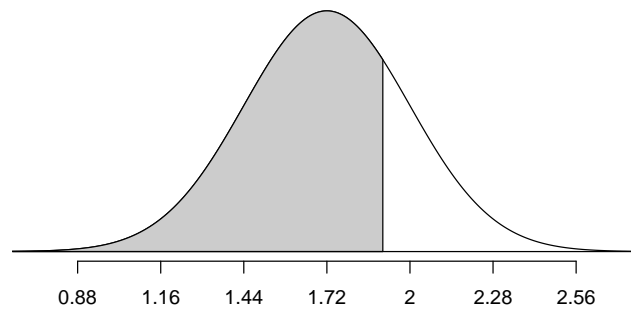
68-95-99.7 Rule

- 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)
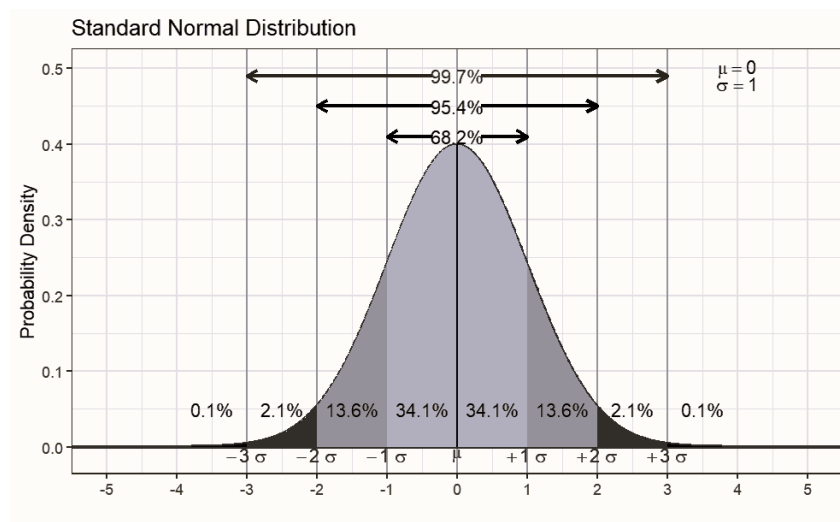
General steps of a hypothesis test

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.
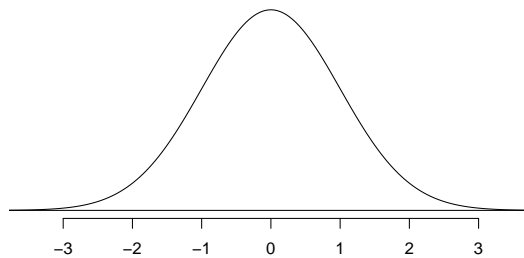
Interpret the standardized statistic

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
- $CI = \text{statistic} \pm \text{margin of error}$

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
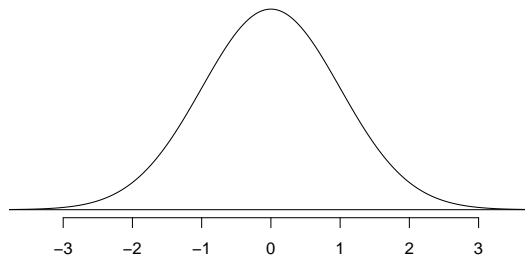
For a 95% confidence interval:

```r
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
```

_____ of Handedness of Male Profes:
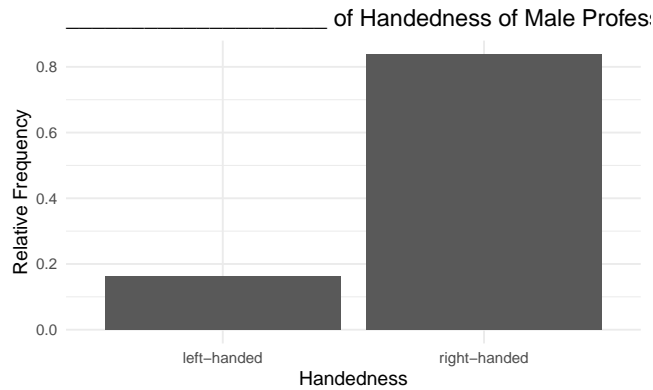
1. What type of plot was created of these data?

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.
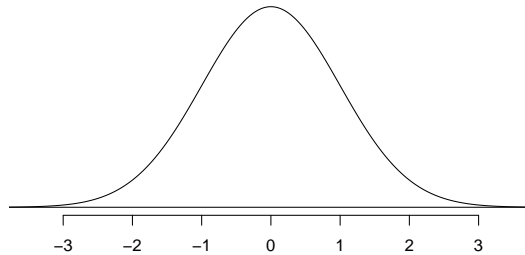
Figure 4.1: Standard Normal Curve

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. In the code below, 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 20–22

```
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? Would the p-value be larger or smaller?

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:

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

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.
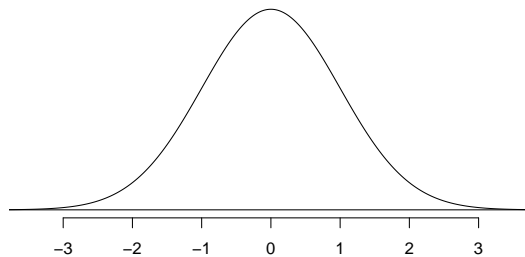
Figure 4.2: Standard Normal Curve

84

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:

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

19. Mark the Z-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.
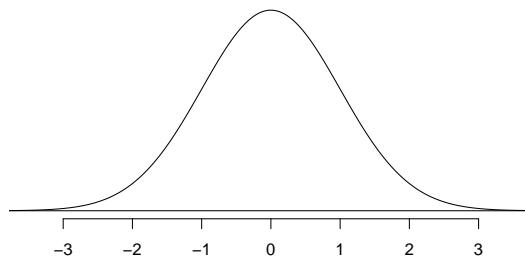
Figure 4.3: Standard Normal Curve

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

In today's activity, 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.

```
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.
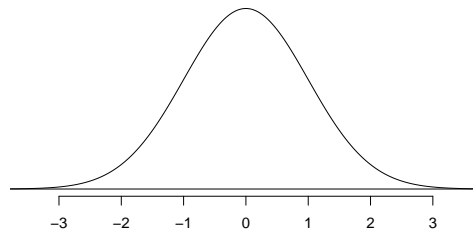
Figure 4.4: Standard Normal Curve

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**.

```
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 same 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` in the one proportion test to create a null distribution with 10000 simulations.
- Then highlight and run lines 21–26.

```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?
```

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 in the provided R script file to find the multiplier for a 90% confidence interval.
- Highlight and run line 31

```r
qnorm(percentile, 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):

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.

5.0.1 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.

5.0.2 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.

5.1   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.

5.1.1 Confidence intervals

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.

5.1.2 Probability

40. Calculate probabilities from a given table and give appropriate probability notation for both conditional and unconditional probabilities.

41. Create a two-way table using given probabilities.

42. Interpret a probability value in context of the problem.

5.2   Module 1 Review - Sampling Methods

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).

3. This study aims to explore whether Swiss university students feeling academic study pressure (whether the student had experienced academic failure) tend to use psychotropic drugs (whether the student had used psychotropic drugs during the student's time at university) as a coping mechanism. An invitation email was sent to all bachelor's and master's students at the University of Lausanne, totaling 15,400 individuals, with a link to access the online questionnaire containing 49 questions and 107 items. No reminder was sent out, and no incentive was given to complete the questionnaire. A total of 1,690 students initially participated in the study, but 424 questionnaires were too incomplete to be used for analysis and were excluded. Additionally, 67 questionnaires were removed because of significant missing sociodemographic information, resulting in 1,199 completed responses included in the final analysis. Is there an association between study pressure and use of psychotropic drugs among Swiss University students?

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).

d. Identify the type and roles of each variable in the study.

4. Researchers decided to investigate whether a cat's coat color is associated with aggressive cat behavior by creating a 20-minute survey. The survey was distributed by posting it to social media and through cat-related listservs (e.g., For the Love of Cats), inviting individuals to take the survey. A total of 1,365 surveys were completed by participants. The frequency of each of the following aggressive behavior categories was assessed: hiss, stalk/chase, bite, slap/scratch. Frequency of behaviors toward people were recorded on a 6-point scale: 0 = never, 1 = less than once every 6 months, 2 = once every 6 months, 3 = once per month, 4 = once per week, 5 = one or more times per day. Because there were four aggressive behavior categories, each with a frequency of 0 to 5 possible, each cat could score between 0 to 20 for human aggression. Is there an association between coat color and aggressive behavior among cats?

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).

d. Identify the type and roles of each variable in the study.

5.3   Module 2 Review - Probability

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.

5.4 Module 3 Review - Simulation Methods for a Single Proportion

```
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 10000 simulations?

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

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

```
one_proportion_test(probability_success = 0.2, #Null hypothesis value
                    sample_size = 1753, #Enter sample size
                    number_repetitions = 10000, #Enter number of simulations
                    as_extreme_as = 0.187, #observed statistic
                    direction = "two-sided", #specify direction of alternative hypothesis
                    summary_measure = "proportion") #Reporting proportion or number of successes?
```


Simulated Proportion of Successes
Proportion of Samples = 1706/10000 = 0.1706
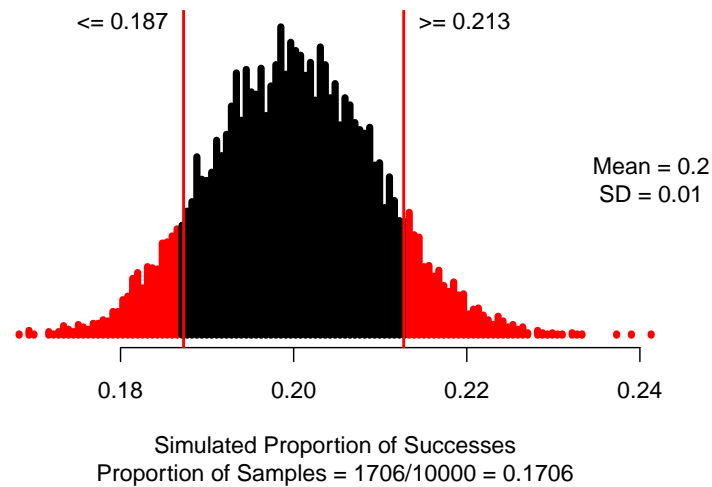
7. Interpret the p-value in context of the problem.

8. How much evidence does the data provide against the null hypothesis?
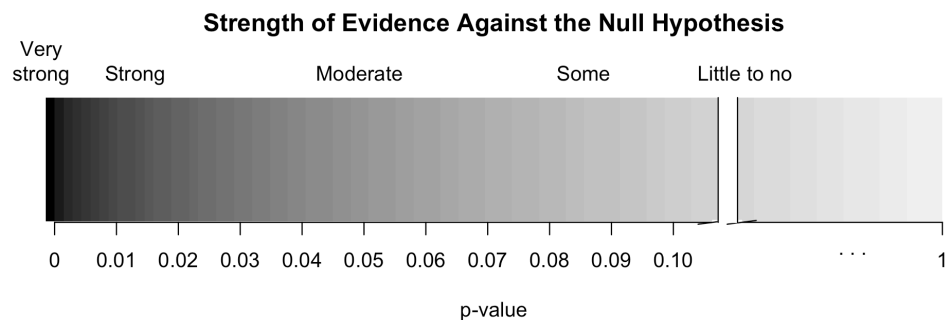
9. Write a conclusion to the study in context of the problem.

103

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):

```
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
```


Bootstrapped Proportions
Mean: 0.187, SD: 0.009, 95% CI: (0.169, 0.205)
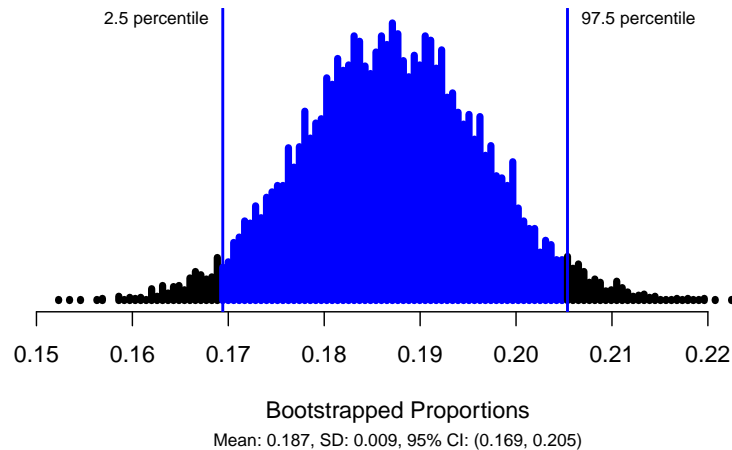
12. Explain how to use cards to create one bootstrap sample.

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

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

5.5 Module 4 Review - Theory-based Methods for a Single Proportion

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.
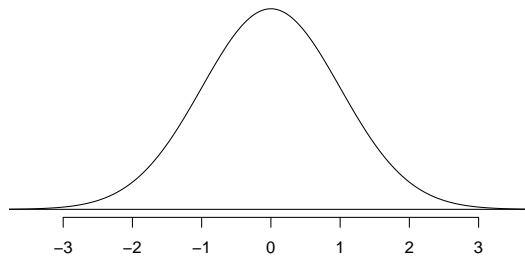
Figure 5.1: A standard normal curve.
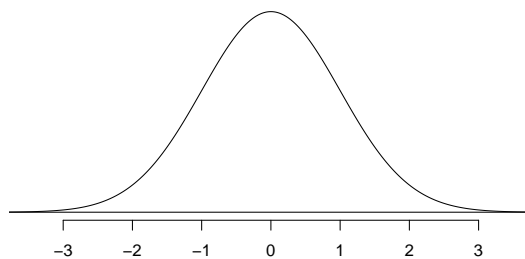
8. Interpret the standardized statistic in context of the problem.

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:

Simulated Proportion of Successes
Proportion of Samples = 0/10000 = 0
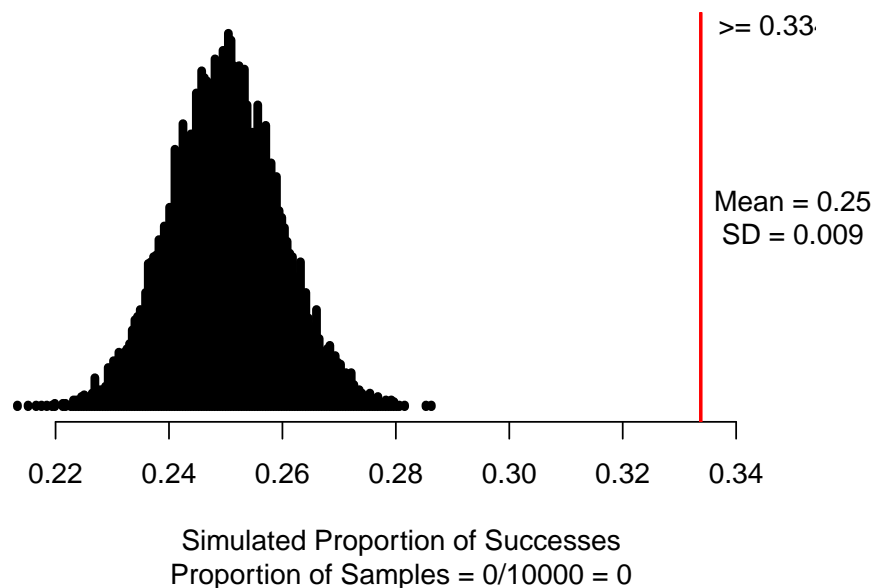
10. Interpret the p-value in context of the study.

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).
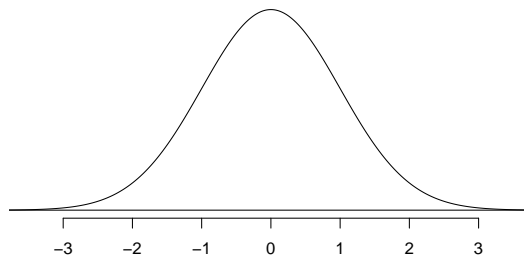
Figure 5.2: A standard normal curve.

```
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:

2.5 percentile   97.5 percentile

0.30   0.32   0.34   0.36

Bootstrapped Proportions
Mean: 0.334, SD: 0.01, 95% CI: (0.314, 0.354)
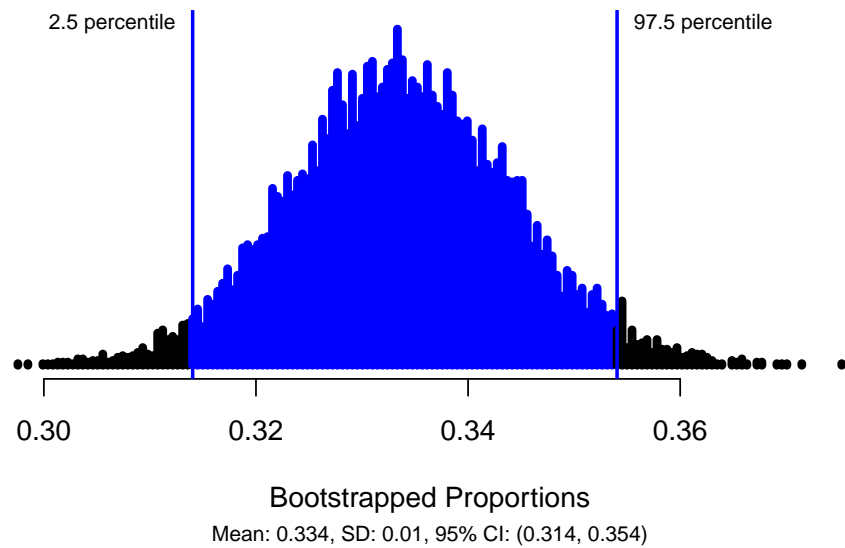
14. Interpret the 95% confidence interval in context of the problem.

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

Exploring Quantitative Data: Exploratory Data Analysis and Hypothesis Testing for a Single Quantitative Variable

6.1 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 quantitative variable.

6.1.1 Key topics

Module 6 will introduce exploratory data analysis and hypothesis testing using both simulation-based and theory-based methods for a single quantitative variable. The **summary measure** for one quantitative variable is the **mean**. Additionally, we can find the five number summary (min, Q1, median, Q3, max) as well as the sample standard deviation.

- Notation for a sample mean: $\bar{x}$

- Notation for a sample standard deviation: s

- Notation for a population mean: μ

- Notation for a population standard deviation: σ

- Types of plots for a single categorical variable:

    - Histogram

    - Boxplot

    - Dotplot

6.1.2 Vocabulary

Sample statistics for a single quantitative variable

- **Mean**, $\bar{x}$: the average

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \ldots, x_n$ are the data values and n is the sample size.

- **Median**: value at the 50th percentile; approximately 50% of data values are at or below the value of the median.

- **Quartile 1** (lower quartile), Q_1: value at the 25th percentile; approximately 25% of data values are at or below the value of Q_1.

- **Quartile 3** (upper quartile), Q_3: value at the 75th percentile; approximately 75% of data values are at or below the value of Q_3.

- **Sample standard deviation**, s: on average, each value in the data set is s units from the mean of the data set ($\bar{x}$). We will always calculate s using R, but it is calculated using the following formula:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}},$$

where $x_1, x_2, \ldots, x_n$ are the data values, $\bar{x}$ is the sample mean, and n is the sample size.

- **Interquartile range**: the range of the data between the two quartiles: $IQR = Q_3 - Q_1$.

- R code to find the summary statistics for a quantitative variable:

```
object %>% # Data set piped into...
    summarise(favstats(variable))
```

Plotting one quantitative variable

- **Histogram**: sorts a quantitative variable into bins of a certain width. R code to create a histogram:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) +    # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Boxplot**: plots the values of the five-number summary and shows any outliers in the data set. R code to create a boxplot:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) + # Name variable to plot
    geom_boxplot() + # Create boxplot
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Dotplot**: plots each value as a dot along the x-axis. R code to create a dotplot:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) + # Name variable to plot
    geom_dotplot() + # Create dotplot
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- Four characteristics of a distribution of a single quantitative variable:

    - Shape (symmetric, skewed left, or skewed right)

    - Center

    - Spread

    - Outliers?

Hypothesis testing for a single mean

- **Hypotheses in notation for a single mean**: In the hypotheses below, μ_0 is the **null value**.

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} \mu_0$$

Simulation-based hypothesis testing

- **Conditions necessary to use simulation-based methods for inference for a single quantitative variable**:

  - **Independence**: observational units must be independent of one another.

- **Simulation-based methods to create the null distribution**: R code to use for simulation-based methods for one quantitative variable to find the p-value, `one_mean_test` (from the `catstats` package), is shown below. Review the comments (instructions after the #) to see what each should be entered for each line of code.

```
one_mean_test(object$variable, #Enter the object name and variable
        null_value = xx, #Enter the null value for the study
        summary_measure = "mean",  #Can choose between mean or median
        shift = xx, #Difference between the null value and the sample mean
        as_extreme_as = xx, #Value of the summary statistic
        direction = "xx", #Specify direction of alternative hypothesis
        number_repetitions = 10000)
```

Theory-based hypothesis testing

- Theory-based methods should give the same results as simulation-based methods if conditions are met. For a single quantitative variable, conditions are met if either the data themselves follow a normal distribution or if the sample size is large enough. We call this the "normality condition."

- **Conditions for the sampling distribution of $\bar{x}$ 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!)

  - **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

    * $n < 30$: The distribution of the sample must be approximately normal with no outliers.

    * $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.

    * $n \geq 100$: Can assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

- **t-distribution**: a theoretical distribution that is bell-shaped with mean zero. Its degrees of freedom determine the variability of the distribution. For very large degrees of freedom, the t-distribution is close to a standard normal distribution. For a single quantitative variable, the degrees of freedom are calculated by subtracting one from the sample size: $n - 1$. A t-distribution with $n - 1$ degrees of freedom is denoted by: t_{n-1}.

- **Standard error of the sample mean**: measures the how far each possible sample mean is from the true mean, on average, and is calculated using the formula below:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation.

  - For inference involving means, the formula for the standard error will be the same for both hypothesis tests and confidence intervals (unlike inference involving proportions, where the standard error for a hypothesis test used the null value in the calculation).

- **Standardized sample mean**: standardized statistic for a single quantitative variable calculated using:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

If the conditions for the sampling distribution of $\bar{x}$ to follow an approximate normal distribution are met, and if the true value of μ is equal to the null value of μ_0, the standardized sample mean, T, will have an approximate t-distribution with $n-1$ degrees of freedom.

- The theory-based **p-value** for hypothesis testing involving means can be found in R by using the `pt` function to find the probability of the observed standardized statistic or one more extreme (in the direction of H_A). This probability is the area under a *t-distribution with the appropriate degrees of freedom* at or more extreme than the observed standardized statistic.

  - `pt` will give you a p-value using the t-distribution with a given degrees of freedom (enter for `yy`). For a single mean, `df` $= n-1$.

  - Enter the value of the standardized statistic for `xx`

  - If a "greater than" alternative, change `lower.tail = TRUE` to `FALSE`.

  - If a two-sided test, multiply by 2.

```
pt(xx, df = yy, lower.tail=TRUE)
```

6.2 Video Notes: Exploratory Data Analysis and Hypothesis Testing of Quantitative Variables

Read Chapters 5 and 17 in the course textbook. Use the following videos to complete the video notes for Module 6.

6.2.1 Course Videos

- QuantitativeData
- 5.5to5.6
- 5.7
- 17.2
- 17.3TheoryTests

Summarizing quantitative data - Videos 5.2to5.4 and 5.5to5.6

Types of plots

We will revisit the moving to Montana data set and plot the age of the buyers.

Dotplot:

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_dotplot() +
  labs(title = "Dotplot of Age of Buyers from Gallatin
       County Home Sales", #Title your plot
       x = "Age", #x-axis label
       y = "Proportion") #y-axis label
```


Dotplot of Age of Buyers from Gallatin County Home Sales
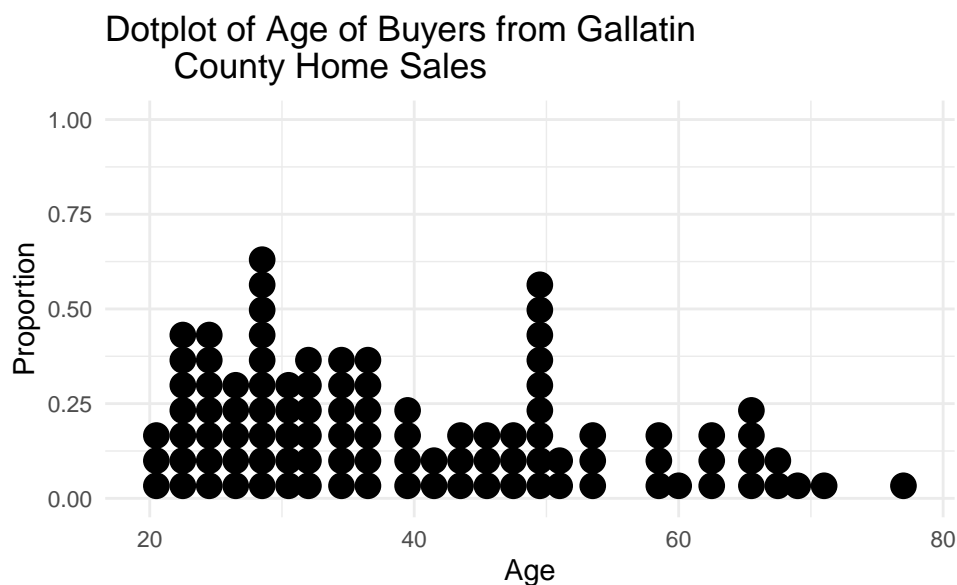
Histogram:

```
moving %>%
  ggplot(aes(x = Age))+
  geom_histogram(binwidth = 7) +
  labs(title = "Histogram of Age of Buyers from Gallatin
       County Home Sales",
       #Title your plot
       x = "Age",
       y = "Count")
```

Histogram of Age of Buyers from Gallatin
County Home Sales
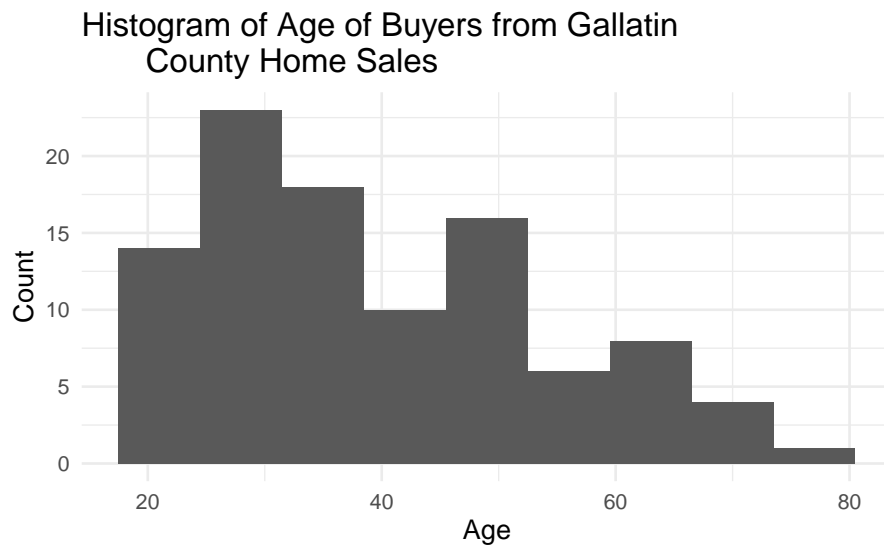
Quantitative data can be numerically summarized by finding:

Two measures of center:

- Mean: _____ of all the _____ in the data set.

  – Sum the values in the data set and divide the sum by the sample size

- Notation used for the population mean:

  – Single quantitative variable:

    - One categorical and one quantitative variable:

      - Subscripts represent the _____ variable groups

- Notation used for the sample mean:

    - Single quantitative variable:

    - One categorical and one quantitative variable:

- Median: Value at the _____ percentile

  - _____ % of values are at and _____ and at and _____ the value of the _____.

  - Middle value in a list of ordered values

Two measures of spread:

- Standard deviation: Average _____ each data point is from the _____ of the data set.

  - Notation used for the population standard deviation

  - Notation used for the sample standard deviation

- Interquartile range: middle 50% of data values

  Formula:

  Quartile 3 (Q3) - value at the 75th percentile

  - _____ % of values are at and _____ the value of Q3

  Quartile 1 (Q1) - value at the 25th percentile

  - _____ % of values are at and _____ the value of Q1

Boxplot (3rd type of plot for quantitative variables)

– Five number summary: minimum, Q1, median, Q3, maximum

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_boxplot() +
  labs(title = "Boxplot of Age of Buyers from Gallatin
       County Home Sales", #Title your plot
       x = "Age", #x-axis label
       y = "") #y-axis label
```


Boxplot of Age of Buyers from Gallatin County Home Sales
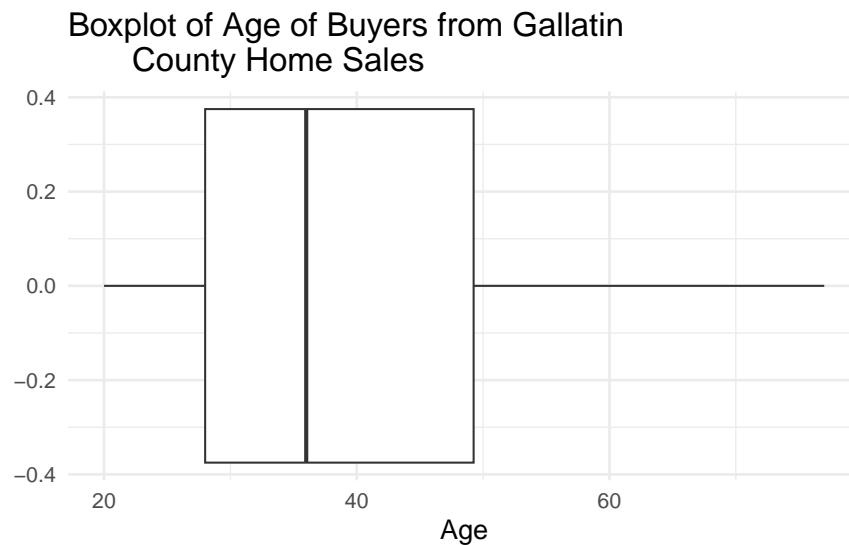
```
favstats(moving$Age)
```

```
#>  min Q1 median    Q3 max  mean        sd    n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Interpret the value of Q_3 for the age of buyers.

Interpret the value of s for the age of buyers.

Four characteristics of plots for quantitative variables

- Shape: overall pattern of the data

(a) Negatively skewed   **(b) Normal (no skew)**   **(c) Positively skewed**
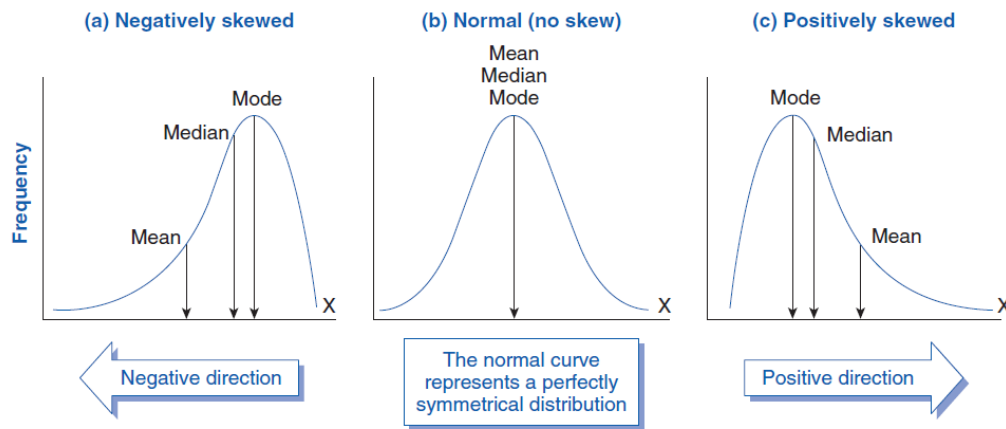
- What is the shape of the distribution of age of buyers for Gallatin County home sales?

- Center:

  Mean or Median

  - Report the measure of center for the boxplot of age of buyers for Gallatin County home sales.

- Spread (or variability):

  Standard deviation or IQR

  - Report the IQR for the distribution of age of buyers from Gallatin County home sales.

- Outliers?

  values $< Q_1 - 1.5 \times IQR$

  values $> Q_3 + 1.5 \times IQR$

  - Use these formulas to show that there are no outliers in the distribution of age of buyers from Gallatin County home sales.

Let's look at side-by-side boxplot of the variable age by state of origin moved from.

```
moving %>%  # Data set piped into...
  ggplot(aes(y = Age, x = From))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of Age by State of Origin
  of Buyers from Gallatin County Home Sales",  # Title
       x = "State of Origin",    # x-axis label
       y = "Age")  # y-axis label
```


Side by side box plot of Age by State of Origin
of Buyers from Gallatin County Home Sales
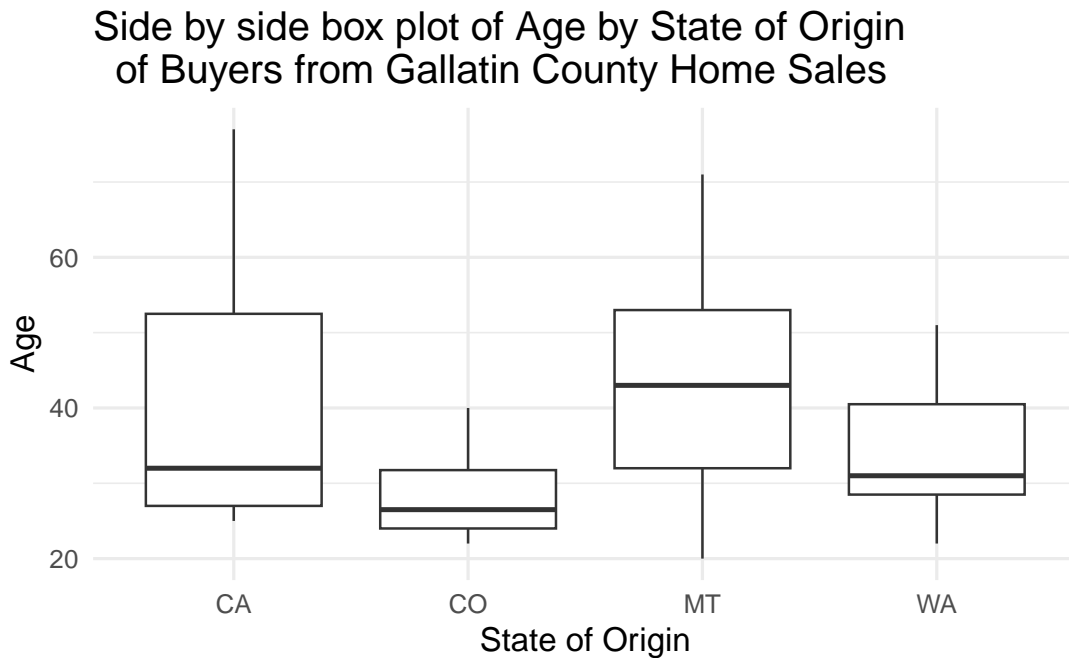
- Which state of origin had the oldest median age of buyers from sampled home sales?

- Which state of origin had the most variability in age of buyers from sampled home sales?

- Which state of origin had the most symmetric distribution of ages of buyers from sampled home sales?

- Which state of origin had outliers for the age of buyers from sampled home sales?

Robust statistics - Video 5.7

Let's review the summary statistics and histogram of age of buyers from sampled home sales.

Histogram of Age of Buyers from Gallatin County Home Sales
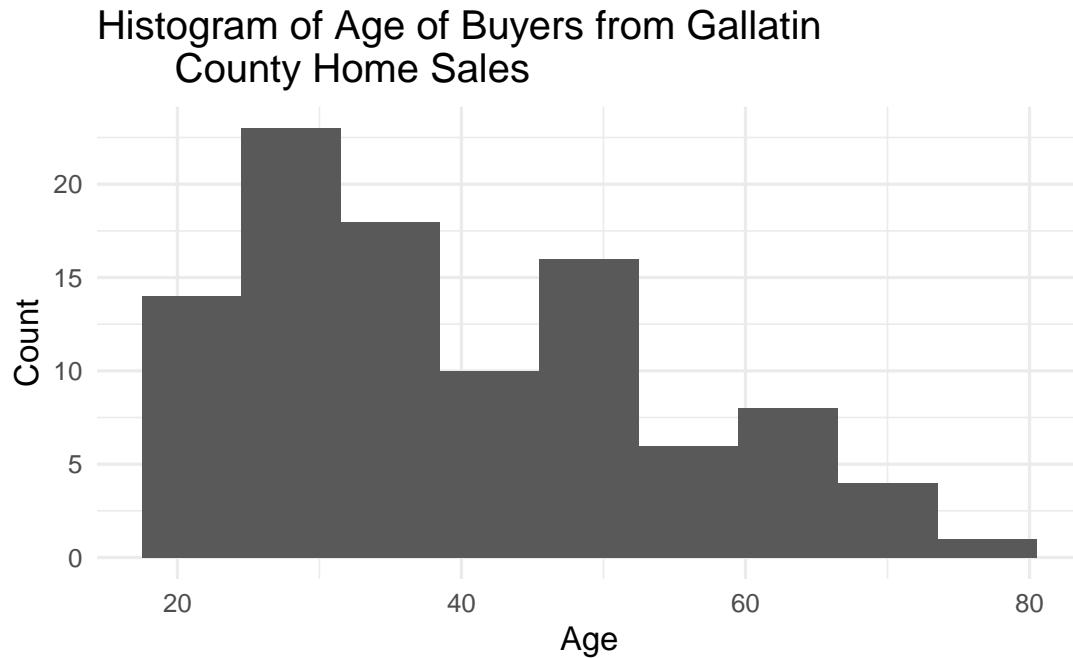
```
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Notice that the _____ has been pulled in the direction of the _____.

- The _____ is a robust measure of center.

- The _____ is a robust measure of spread.

- Robust means not _____ by outliers.

When the distribution is symmetric use the _____ as the measure of center and the _____ as the measure of spread.

When the distribution is skewed with outliers use the _____ as the measure of center and the _____ as the measure of spread.

6.2.2 Video notes single quantitative variable inference

Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

```r
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```r
pb %>%
    ggplot(aes(x = Weight)) +    # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Histogram of Male Polar Bear Weight", # Title for plot
        x = "Weight (kg)", # Label for x axis
        y = "Frequency") # Label for y axis

pb %>% # Data set piped into...
ggplot(aes(x = Weight)) +    # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot
        x = "Weight (kg)", # Label for x axis
        y = "Frequency") # Label for y axis
```

Summary Statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median     Q3    max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Hypothesis testing

- Hypotheses are always written about the _____. For a single mean we will use the notation _____.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

- Direction of the alternative depends on the _____ _____.

Simulation-based method

- Simulate many samples assuming $H_0 : \mu = \mu_0$
  - Shift the data by the difference between μ_0 and $\bar{x}$
  - Sample with replacement n times from the shifted data
  - Plot the simulated shifted sample mean from each simulation
  - Repeat 10000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}$

Example: Is there evidence that male polar bears weigh less than 370kg (previously recorded measure), on average? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

Hypotheses:

In notation:

H_0 :

H_A :

In words:

H_0 :

H_A :

Reminder of summary statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median    Q3   max    mean      sd n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Find the difference:

$\mu_0 - \bar{x} =$

```
set.seed(216)
one_mean_test(pb$Weight,    #Enter the object name and variable
              null_value = 370, #Enter null value for the study
              summary_measure = "mean",   #Can choose between mean or median
              shift = 45.4,    # Shift needed for bootstrap hypothesis test
              as_extreme_as = 324.6,   # Observed statistic
              direction = "less",   # Direction of alternative
              number_repetitions = 10000)  # Number of simulated samples for null distribution
```


Mean = 370.052
SD = 9.583

Simulated Mean
Count = 0/10000 = 0
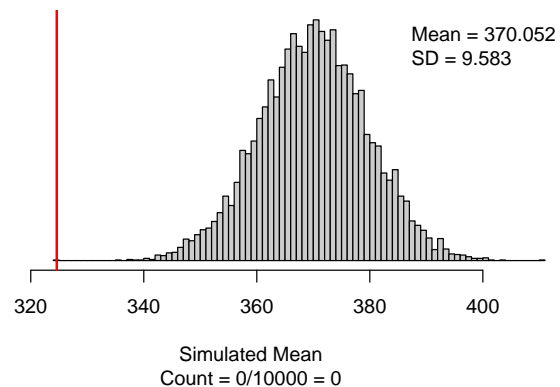
Interpretation of the p-value:

- 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

Theory-based method

Conditions for inference using theory-based methods:

- Independence:

- Large enough sample size:

t-distribution

In the theoretical approach, we use the Central Limit Theorem (CLT) to tell us that—under certain conditions—the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0, and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

- Estimate the population standard deviation, σ, with the _____ standard deviation, _____.

- For a single quantitative variable we use the _____ - distribution with _____ degrees of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n-1$ degrees of freedom.

t−distribution with 82 df

- Calculate the standardized statistic
- Find the area under the t-distribution with $n - 1$ df at least as extreme as the standardized statistic

Equation for the standard error of the sample mean:

Equation for the standardized sample mean:

Calculate the standardized sample mean weight of adult male polar bears:

t−distribution with 82 df
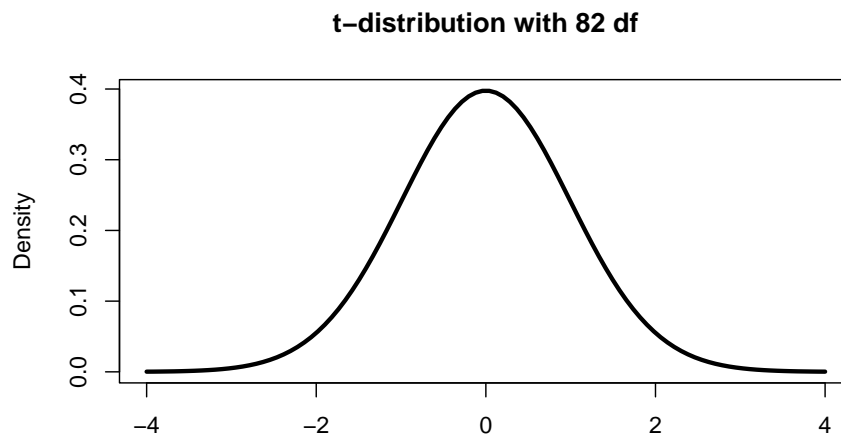
Interpret the standardized sample mean weight:

To find the theory-based p-value:

```
pt(-4.683, df=82, lower.tail=TRUE)
#> [1] 5.531605e-06
```

6.2.3 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 plots can be used to summarize quantitative data?

2. Which measure of center is robust to outliers?

3. How do we determine the direction of the alternative hypothesis?

6.3 Activity 11: Summarizing Quantitative Variables

6.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.

- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

6.3.2 Terminology review

In today's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median

- Two measures of spread (variability): standard deviation, interquartile range (IQR)

- Plots of quantitative variables: dotplots, boxplots, histograms

- Given a plot or set of plots, describe and compare the distribution(s) of quantitative variables (center, variability, shape, outliers).

To review these concepts, see Chapter 5 in the textbook.

6.3.3 The Integrated Postsecondary Education Data System (IPEDS)

These data were collected on a subset of higher education institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting

- United States only

- Title IV participating

- Not for profit

- 2-year or 4-year or above

- Has full-time first-time undergraduates

Some of the variables collected and their descriptions are below. Note that several variables have missing values for some institutions (denoted by "NA").

Variable	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Sector	whether public or private
LandGrant	Is this a land-grant institution (Yes/No)
Size	Institution size category based on total student enrolled for credit, Fall 2018: Under 1,000, $1,000\$-4,999, 5,000$-9,999, 10,000$-\$19,999, 20,000 and above
Cost_OutofState	Cost of attendance for full-time out-of-state undergraduate students
Cost_InState	Cost of attendance for full-time in-state undergraduate students
Retention	Retention rate is the percent of the undergraduate students that re-enroll in the next year
Graduation_Rate	6-year graduation rate for undergraduate students

Variable	Description
SATMath_75	75th percentile Math SAT score
ACT_75	75th percentile ACT score

Identifying variables in a data set

Look through the provided table of variable descriptions. The `UnitID` and `Name` are identifiers for each observational unit, *US degree-granting higher education institutions in 2018*.

1. Identify in the table which variables collected on the US institutions are categorical (C) and which variables are quantitative (Q).

Summarizing quantitative variables

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. The R output below provides the summary statistics for the variable `Graduation_Rate`. The summary statistics provided are the two measures of center (mean and median) and two measures of spread (standard deviation and the quartile values to calculate the IQR) for undergraduate 6-year graduation rate.

- Highlight and run lines 1–12 in the provided R script file to load the data set. Check that the summary statistics match the output given in the coursepack.

- Notice that the 2-year institutions were removed so the observational units for this study are **4-year US degree-granting higher education institutions in 2018.**

```
IPEDS <- read.csv("https://www.math.montana.edu/courses/s216/data/IPEDS_2018.csv")
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") # Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") # Filters the data set to remove Private 2-year
IPEDS %>%
    summarize(favstats(Graduation_Rate))
```

```
#>   min Q1 median Q3 max    mean      sd   n missing
#> 1   0 38     53 67 100 52.48749 20.63192 1918      49
```

2. Report the values for the two measures of center (mean and median).

3. Calculate the interquartile range ($IQR = Q_3 - Q_1$) of graduation rates.

4. Report the value of the standard deviation and interpret this value in context of the problem.

5. Interpret the value of Q_3 in context of the study.

Displaying a single quantitative variable

There are three type of plots used to plot a single quantitative variable: a dotplot, a histogram or a boxplot. A dotplot of graduation rates would plot a dot for the graduation rate for each 4-year US higher education institution.

First, let's create a histogram of the variable `Graduation_Rate`.

- Enter the name of the variable in line 19 for `variable` in the R script file.

- Replace the word title for the plot in line 21 between the quotations with a descriptive title. **A title should include: type of plot, variable or variables plotted, and observational units.**

- Highlight and run lines 18–23 to create the histogram.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = xx)) +    # Name variable to plot
  geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
  labs(title = "Don't forget to title the plot!", # Title for plot
       x = "Graduation Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

Notice that the **bin width** for the histogram is 10. For example the first bin consists of the number of institutions in the data set with a graduation rate of 0 to 10%. It is important to note that a graduation rate on the boundary of a bin will fall into the bin above it; for example, 20 would be counted in the bin 20–30.

6. Which range of Graduation Rates have the highest frequency?

Next we will create a boxplot of the variable `Graduation_Rate`.

- Enter the name of the variable in line 28 for `variable` in the R script file.

- Highlight and run lines 28–36 to create the boxplot.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +    # Name variable to plot
  geom_boxplot() +  # Create boxplot with specified binwidth
  labs(title = "Boxplot of Graduation Rates for \n 4-year Higher Education Institutions",
          # Title for plot
          # Note the \n starts a new line
       x = "Graduation_Rate", # Label for x axis
       y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

7. Sketch the boxplot created and identify the values of the 5-number summary (minimum value, Q_1, median, Q_3, maximum value) on the plot. Use the following formulas to find the invisible fence on both ends of the distribution. Draw a dotted line at the invisible fence to show how the outliers were detected (any values less than the lower fence or greater than the upper fence were flagged as outliers).

$$\text{Lower Fence:} \quad Q_1 - 1.5 \times IQR \qquad \text{Upper Fence:} \quad Q_3 + 1.5 \times IQR$$

When describing distributions of quantitative variables we discuss the **shape** (symmetric or skewed), the **center** (mean or median), **spread** (standard deviation or IQR), and if there are **outliers** present.

8. What is the shape of the distribution of graduation rates?

9. From which plot (histogram or boxplot) is it easier to determine the shape of the distribution?

10. From which plot is it easier to determine if there are outliers?

Robust statistics

Let's examine how the presence of outliers affects the different summary measures for center and spread. For this part of the activity, we will look at the retention rate variable (`Retention`) in the IPEDS data set.

```
IPEDS %>% # Data set piped into...
    summarise(favstats(Retention))
#>    min Q1 median Q3 max    mean       sd    n missing
#> 1    0 66     75 83 100 73.8525 15.14323 1817     150

IPEDS %>% # Data set piped into...
    ggplot(aes(x = Retention)) + # Name variable to plot
    geom_boxplot() + # Create boxplot
    labs(title = "Boxplot of Retention Rates for \n 4-year Higher Education Institutions",
         # Title for plot
        x = "Retention Rates (%)", # Label for x axis
        y = "Frequency") # Label for y axis
```
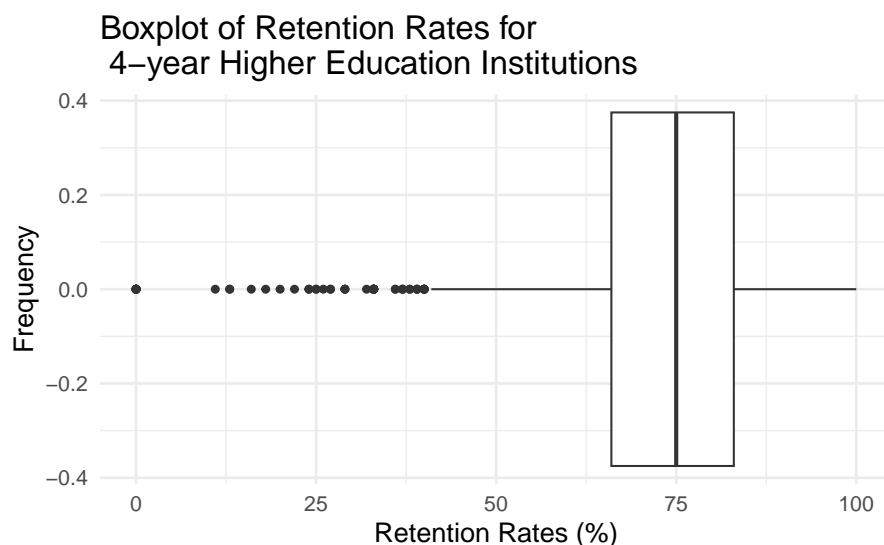

11. Report the values for the two measures of center for these data.

12. Report the values for the two measures of spread for these data.

To show the effect of outliers on the measures of center and spread, the smallest values of retention rate in the data set were increased by 30%. This variable is called `Retention_Inc`.

```
IPEDS %>% # Data set piped into...
    summarise(favstats(Retention_Inc))
#>    min Q1 median Q3 max     mean       sd    n missing
#> 1   30 66     75 83 100 74.49642 13.41255 1817     150


IPEDS %>% # Data set piped into...
    ggplot(aes(x = Retention_Inc)) + # Name variable to plot
    geom_boxplot() + # Create histogram
    labs(title = "Boxplot of Increased Retention Rates for \n 4-year Higher Education Institutions",
        # Title for plot
        x = "Retention Rates (%)", # Label for x axis
        y = "Frequency") # Label for y axis
```
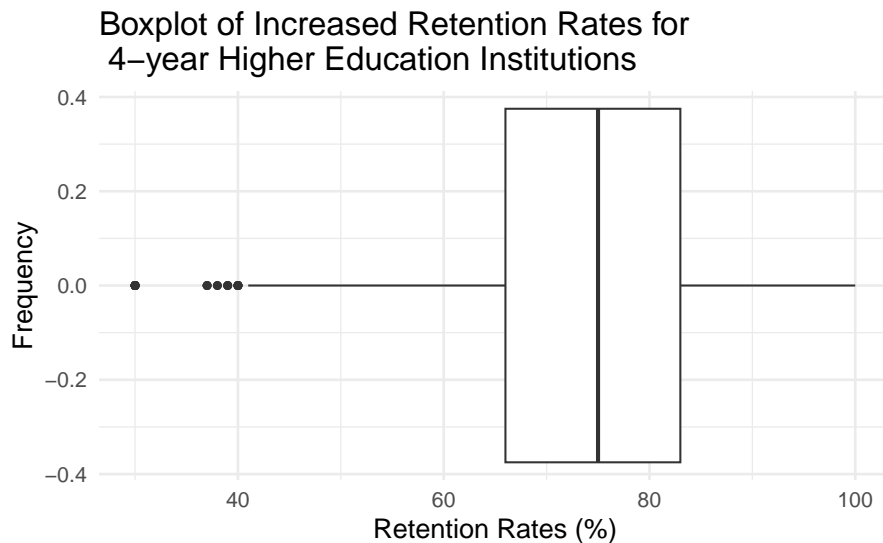

13. Report the values for the two measures of center for this new data set.

14. Report the values for the two measures of spread for this new data set.

15. Which measure of center is robust to outliers? Explain your answer.

16. Which measure of spread is robust to outliers? Explain your answer.

6.3.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.

2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Whiskers extend to the lowest value and highest value that are *not* considered outliers. Values in the data set that are less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.

3. Data should be summarized numerically and displayed graphically to give us information about the study.

4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. In this course, we only consider two measures of center (mean and the median), and two measures of spread (standard deviation and the interquartile range, $IQR = Q_3 - Q_1$).

6.3.5 Additional notes

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

6.4 Activity 12: Hypothesis Testing of a Single Quantitative Variable

6.4.1 Learning outcomes

- Given a research question involving one quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Investigate the process of creating a null distribution for one quantitative variable.

- Find, evaluate, and interpret a p-value from the null distribution.

6.4.2 Terminology review

In today's activity, we will use simulation-based and theory-based methods to analyze a single quantitative variable. Some terms covered in this activity are:

- Null hypothesis

- Alternative hypothesis

- Null distribution

- t-distribution

- p-value

To review these concepts, see Chapters 9 and 17 in the textbook.

6.4.3 College student sleep habits

According to an article in *Sleep* (Watson 2015), experts recommend adults (>18 years old) get at least 7 hours of sleep per night. A professor at MSU is interested in the sleep habits of MSU students. The professor obtained a representative sample of MSU students and asked each student to report the amount of sleep they get on a typically night. Is there evidence that MSU students get less than the recommended 7 hours of sleep per night, on average?

Summarizing quantitative variables

- Download the R script file and data file for this activity

- Upload both files to the RStudio server and open the R script file

- Enter the name of the dataset for datasetname.csv

- Highlight and run lines 1–8 to load the data

```
sleep <- read.csv("datasetname.csv")
```

Ask a research question

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

2. Write the null hypothesis in words in context of the study.

3. Write the alternative hypothesis in notation.

Summarize and visualize the data

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable.

- Enter the variable name, `SleepHours`, for `variable` in line 13
- Highlight and run lines 12–13

```
sleep %>%
    summarize(favstats(variable))
```

4. About how far is each number of hours of sleep for a Stat 216 student from the mean number of hours of sleep, on average?

Create a boxplot of the variable `SleepHours`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Enter a title in line 21 for the plot between the quotations.
- Highlight and run lines 18–25.

```
sleep %>% # Data set piped into...
    ggplot(aes(x = variable)) +    # Name variable to plot
    geom_boxplot() +  # Create boxplot with specified binwidth
    labs(title = "Don't forget to title your plot!", # Title for plot
        x = "Amount of sleep (hrs)", # Label for x axis
        y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

5. Describe the distribution of number of hours of sleep using the four characteristics of boxplots.

Simulation methods

To simulate the null distribution of sample means we will use a bootstrapping method. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping, we will need to *shift* each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the mean of the shifted data is μ_0 (rather than the mean of the original data, $\bar{x}$), and that the simulated null distribution will be centered at the null value.

6. Calculate the difference $\mu_0 - \bar{x}$. Based on the sign of this difference, will we need to shift the data up or down?

- Open the data set (`sleep_college`) in Excel.

- Create a new column labeled Shift.

- In the column, Shift, add the shifted value (answer to Q6) to each value in the `SleepHours` column.

- Save the file and upload again to the RStudio server.

- Find the `favstats` of the variable, `Shift`.

- Highlight and run lines 30–32

```
sleep <- read.csv("sleep_college.csv")
sleep %>%
    summarize(favstats(Shift))
```

7. Report the mean of the `Shift` variable. Why does it make sense that this value is the same as the null value?

8. Report the standard deviation of the `Shift` variable. How does this compare to the standard deviation for the variable `SleepHours`? Explain why these values are the same.

9. What inputs should be entered for each of the following to create the simulated null distribution?

- Null value (What is the null value for the study?):

- Summary measure ("mean" or "median"):

- Shift (difference between $\mu_0 - \bar{x}$):

- As extreme as (enter the value for the observed sample mean):

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

- Number of repetitions:

The `one_mean_test` will be used to find the p-value for the simulation test. Following the instructions below to complete the code.

- Enter your answers for question 9 in place of the `xx`'s to produce the null distribution with 10000 simulations.

- Highlight and run lines 36–42.

```r
one_mean_test(sleep$SleepHours,#Enter the object name and variable
              null_value = xx,
              summary_measure = "xx",  #Can choose between mean or median
              shift = xx, #Difference between the null value and the sample mean
              as_extreme_as = xx, #Value of the summary statistic
              direction = "xx", #Specify direction of alternative hypothesis
              number_repetitions = 10000)
```

10. Interpret the p-value of the test in context of the problem.

11. Write a conclusion to the test in context of the problem. Include the appropriate population to which the study design allows us to generalize.

6.4.4   Take-home messages

1. We use bootstrapping—sampling with replacement—from the shifted data to generate a null distribution of simulated sample means. In order to ensure that the null distribution is centered at the null value, μ_0, we shift the data by adding $\mu_0 - \bar{x}$ to each value in the original data set. Note that if this value of the shift is negative, we are shifting the data down; if it is positive, we shift the data up.

2. The mean of the shifted data will equal the null value, μ_0, but the standard deviation of the shifted data will be the same as the standard deviation of the original data.

3. As in the one proportion scenario, we calculate the p-value for a simulation-based hypothesis test for a single mean by finding the proportion of simulated sample means that are as or more extreme (in the direction of H_A) as the observed sample mean, $\bar{x}$.

6.4.5   Additional notes

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

6.5 Activity 13: Body Temperature

6.5.1 Learning outcomes

- Given a research question involving a quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a theory-based hypothesis test for a single mean.

- Interpret and evaluate a p-value for a theory-based hypothesis test for a single mean.

6.5.2 Terminology review

In today's activity, we will analyze quantitative data using theory-based methods. Some terms covered in this activity are:

- Normality

- t-distribution

- Degrees of freedom

- T-score

To review these concepts, see Chapters 11 and 17 in the textbook.

6.5.3 Body Temperature

It has long been reported that the mean body temperature of adults is 98.6°F. There have been a few articles that challenge this assertion. (LUETKEMEIER 2017) In 2018, a sample of 52 Stat 216 undergraduates were asked to report their body temperature. Is there evidence that body temperatures of adults differ from the known temperature of 98.6°F?

Ask a research question

1. Write out the null hypothesis in proper notation for this study.

2. Write out the alternative hypothesis in words for this study.

In general, the sampling distribution for a sample mean, $\bar{x}$, based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a Normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}$ 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!)

- **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

- $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

- $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

- $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Like we saw in Chapter **5**, we will not know the values of the parameters and must use the sample data to estimate them. Unlike with proportions, in which we only needed to estimate the population proportion, π, quantitative sample data must be used to estimate both a population mean μ and a population standard deviation σ. This additional uncertainty will require us to use a theoretical distribution that is just a bit wider than the standard Normal distribution. Enter the *t*-**distribution**!

As you can seen from Figure 6.1, the *t*-distributions (dashed and dotted lines) are centered at 0 just like a standard Normal distribution (solid line), but are slightly wider. The variability of a *t*-distribution depends on its degrees of freedom, which is calculated from the sample size of a study. (For a single sample of n observations or paired differences, the degrees of freedom is equal to $n - 1$.) Recall from previous classes that larger sample sizes tend to result in narrower sampling distributions. We see that here as well. The larger the sample size, the larger the degrees of freedom, the narrower the *t*-distribution. (In fact, a *t*-distribution with infinite degrees of freedom actually IS the standard Normal distribution!)
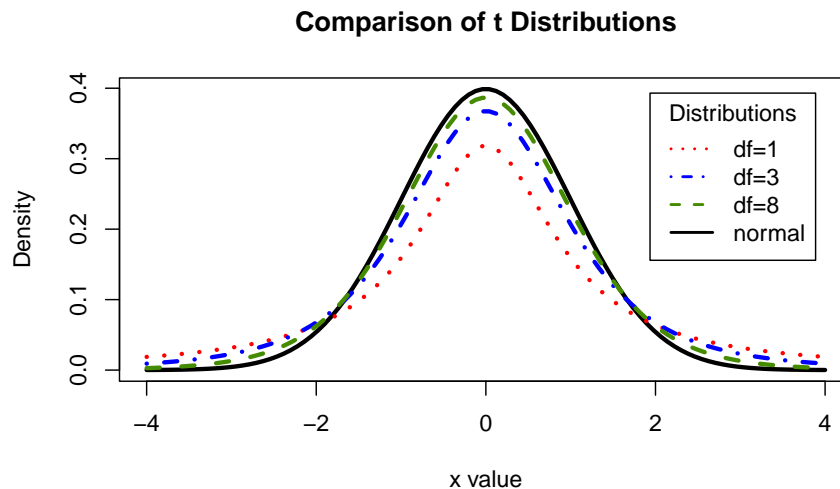
Figure 6.1: Comparison of the standard Normal vs t-distribution with various degrees of freedom

Summarize and visualize the data

The following code is used to create a boxplot of the data.

- Download the R script file upload to the R studio server.

- Open the R script file and highlight and run lines 1–14.

```
bodytemp <- read.csv("https://math.montana.edu/courses/s216/data/normal_temperature.csv")
bodytemp %>%
```

```
ggplot(aes(x = Temp))+
geom_boxplot()+
labs(title="Boxplot of Body Temperatures for Stat 216 Students",
     x = "body temperature (*F)") +
    theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) # Removes y-axis ticks
```

Boxplot of Body Temperatures for Stat 216 Students

body temperature (*F)
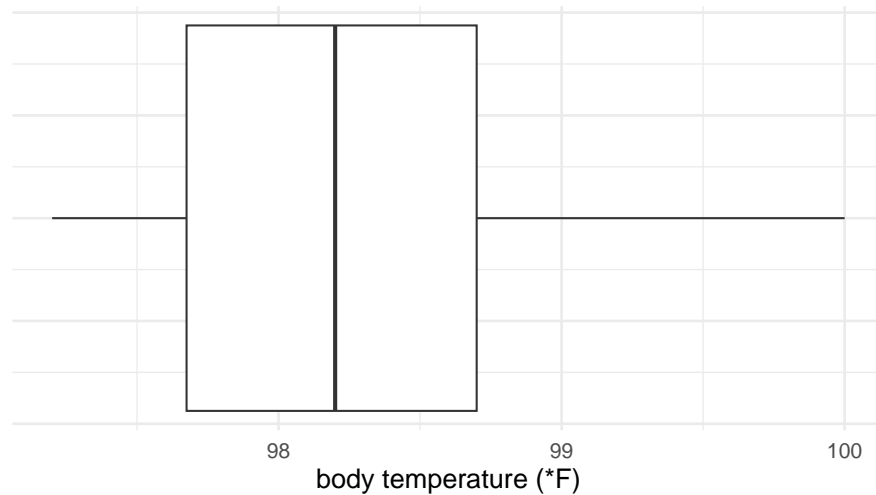
- Highlight and run lines 17 - 18 to get the summary statistics for the variable `Temp`.

```
bodytemp %>%
  summarise(favstats(Temp))
```

```
#>    min    Q1 median   Q3 max    mean        sd  n missing
#> 1 97.2 97.675   98.2 98.7 100 98.28462 0.6823789 52       0
```

Check theoretical conditions

3. Report the sample size of the study. Give appropriate notation.

4. Report the sample mean of the study. Give appropriate notation.

5. How do you know the independence condition is met for these data?

6. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the mean we will use the following formula:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

where the standard error of the sample mean is:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

7. Calculate the standard error of the sample mean.

8. Interpret the standard error in context of the study.

9. Calculate the standardized mean.

10. We model a single mean with a t-distribution with $n - 1$ degrees of freedom. Calculate the degrees of freedom for this study and use it to fill in the blank in the title of the t-distribution displayed below.

11. Mark the value of the standardized statistic on the t-distribution and illustrate how the p-value is found.

t Distribution with __ df

140
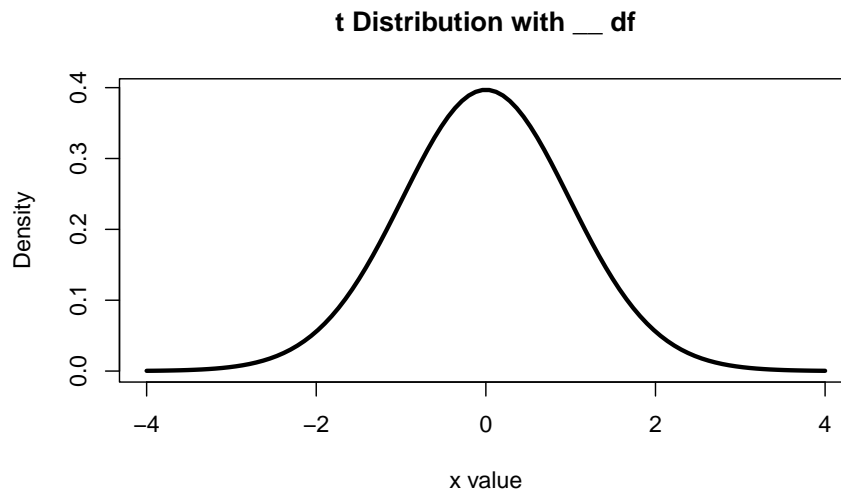
To find the p-value for the theory-based test in R:

- Enter the value for the standardized statistic for `xx` in the `pt` function.

- Enter the degrees of freedom for `yy` in the `pt` function.

- Highlight and run line 24.

```
pt(xx, df=yy, lower.tail=FALSE)
```

12. What does this p-value mean, in the context of the study? Hint: it is the probability of what…assuming what?

13. Write a conclusion to the test in context of the study.

14. Can we generalize the results of the study to all adults? Explain your answer.

6.5.4   Take-home messages

1. In order to use theory-based methods for a quantitative variable, the independent observational units and normality conditions must be met.

2. In order to find a theory-based p-value, we use R to calculate the area under a t-distribution with $n-1$ degrees of freedom (df) that is at or more extreme than the observed T-score. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.

3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

6.5.5   Additional notes

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

Confidence Intervals for a Single Quantitative Variable

7.1 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 quantitative variable.

7.1.1 Key topics

Module 7 will cover creating confidence intervals using both simulation-based and theory-based methods. Additionally, we learn about types of errors and power in hypothesis testing.

Simulation-based confidence interval

- Review the conditions necessary for simulation-based inference in Module 6.

- R code to find the simulation-based confidence interval using the `onemean_CI` function from the `catstats` package.

```
one_mean_CI(object$variable, #Enter the name of the variable
        summary_measure = "mean", #choose the mean or median
        number_repetitions = 10000,  # Number of simulations
        confidence_level = xx)
```

- Interpretation of the confidence interval is very similar as for a single proportion only the context and summary measure has changed.

    - To write in context include:

        * How confident you are (e.g., 90%, 95%, 98%, 99%)

        * Parameter of interest

        * Calculated interval

Theory-based confidence interval

- Review the conditions necessary for theory-based inference in Module 6.

- **Margin of error**: half the width of the confidence interval. For a single mean, the margin of error is:

$$ME = t^* \times SE(\bar{x})$$

where t^* is the **multiplier**, corresponding to the desired confidence level found from a t-distribution with $n - 1$ degrees of freedom and

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

- To find the endpoints of a confidence interval, add and subtract the margin of error to the sample statistic. The confidence interval for a population mean is:

$$\bar{x} \pm ME$$

- R code to find the **multiplier** for a confidence interval using theory-based methods involving means.

  - `qt` will give you the multiplier using a *t*-distribution with a given degrees of freedom (enter for `yy`). For a single mean, `df` $= n - 1$.

  - Enter the percentile for the given level of confidence (e.g., 0.975 for a 95% confidence level).

```
qt(percentile, df = yy, lower.tail=FALSE)
```

Vocabulary

- **Significance level (α)**: a given cut-off value that we compare the p-value to determine a decision of a test.

- **Decisions**:

  - If the p-value is less than the significance level, we make the decision to *reject the null hypothesis.*

  - If the p-value is greater than the significance level, we make the decision to *fail to reject the null hypothesis.*

- **Type 1 Error**: concluding there is evidence to reject the null hypothesis, when the null is actually true.

  - The probability of making a Type 1 error when the null is actually true is equal to the significance level, α.

- **Type 2 Error**: concluding there is no evidence to reject the null hypothesis, when the null is actually false.

- **Power**: probability of concluding there is evidence to reject the null hypothesis, when the null is actually false.

  - When the null is actually false, the event "reject the null hypothesis" is the *complement* of the event "fail to reject the null hypothesis." Thus, power is equal to 1 minus the probability of a Type 2 error.

7.2 Video Notes: Theory-based Inference for a single quantitative variable

Read Chapters 5 and 17 in the course textbook. Use the following videos to complete the video notes for Module 7.

7.2.1 Course Videos

- 17.1
- 17.3TheoryIntervals

7.2.2 Single quantitative variable

- Reminder: review summary measures and plots discussed in the Module 6 material and Chapter 5 of the textbook.
- The summary measure for a single quantitative variable is the _____.

Notation:

- Population mean:

- Population standard deviation:

- Sample mean:

- Sample standard deviation:

- Sample size:

Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

```
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```
pb %>%
    ggplot(aes(x = Weight)) +    # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Histogram of Male Polar Bear Weight", # Title for plot
        x = "Weight (kg)", # Label for x axis
        y = "Frequency") # Label for y axis

pb %>% # Data set piped into...
ggplot(aes(x = Weight)) +    # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot
        x = "Weight (kg)", # Label for x axis
        y = "Frequency") # Label for y axis
```

Histogram of Male Polar Bear Weight

Boxplot of Male Polar Bear Weight

Summary Statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median    Q3   max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Confidence interval

Simulation-based method

- Label cards with the values from the data set

- Sample with replacement (bootstrap) from the original sample n times

- Plot the simulated sample mean on the bootstrap distribution

- Repeat at least 10000 times (simulations)

- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

- ie. 95% CI = (2.5th percentile, 97.5th percentile)

Conditions for inference for a single mean:

- Independence:

145

```
set.seed(216)
one_mean_CI(pb$Weight,
  summary_measure = "mean",
  number_repetitions = 10000,
  confidence_level = 0.95)
```


Bootstrap Mean
95% CI: (305.436, 343.361)
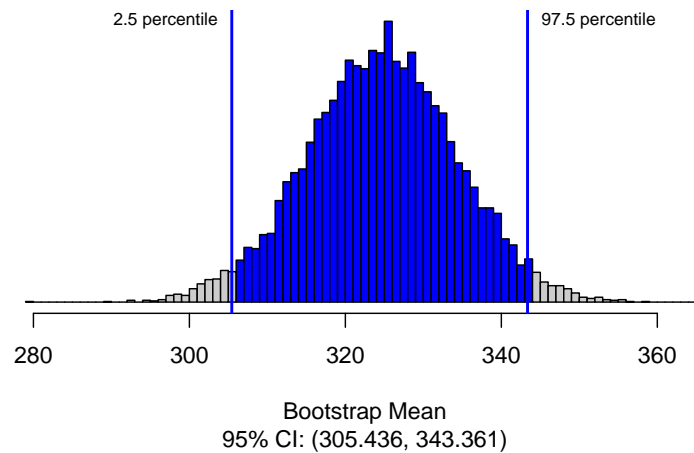
The confidence interval estimates the _____ of _____.

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

Theory-based method

- Calculate the interval centered at the sample statistic

  statistic $\pm$ margin of error

Conditions for inference using theory-based methods:

- Independence:

- Large enough sample size:

t-distribution

In the theoretical approach, we use the CLT to tell us that the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0 and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

- Estimate the population standard deviation, σ, with the _____ standard

  deviation, _____.

- For a single quantitative variable we use the _____ - distribution with _____ degrees

  of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n-1$ degrees of freedom.

t–distribution with 82 df

To find the t^* multiplier for a 95% confidence interval:

```
qt(0.975, df = 82)
#> [1] 1.989319
```

Calculation of the confidence interval for the true mean weight of polar bears from the Southern Beaufort Sea:

7.2.3 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. Are the conditions met to analyze the polar bear data using theory-based methods?

2. Interpret the confidence interval found with simulation methods.

7.3 Activity 14: Danceability of Songs

7.3.1 Learning outcomes

- Use simulation-based methods to find a confidence interval for a single mean.
- Use theory-based methods to find a confidence interval for a single mean.
- Interpret a confidence interval for a single mean.
- Use a confidence interval to determine the conclusion of a hypothesis test.

7.3.2 Terminology review

In today's activity, we will estimate the parameter of interest using simulation-based and theory-based methods. Some terms covered in this activity are:

- Bootstrap distribution
- t-distribution
- Degrees of freedom
- T-score

To review these concepts, see Chapter 17 in the textbook.

7.3.3 Danceability

Spotify created a list of the top songs around the world for the past 10 years and several different audio features of those songs. One of the variables measured on these songs is "danceability." Danceability measures how easy it is to dance to a song; the higher the point value the easier it is to dance to the song. Estimate the average danceability of top songs from Spotify.

- Download the R script file for this activity from D2L and upload to the RStudio server.
- Open the R script file, highlight and run the lines to load libraries and the code below.

```
#>   min Q1 median Q3 max      mean       sd   n missing
#> 1    0 57     66 73  97 64.37977 13.37872 603       0
```

```
songs %>% # Data set piped into...
    ggplot(aes(x = Danceability)) +    # Name variable to plot
    geom_boxplot() +   # Create boxplot with specified binwidth
    labs(title = "Boxplot of Danceability Score for Top Spotify Songs", # Title for plot
         x = "danceability score (points)", # Label for x axis
         y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

Boxplot of Danceability Score for Top Spotify Songs

danceability score (points)
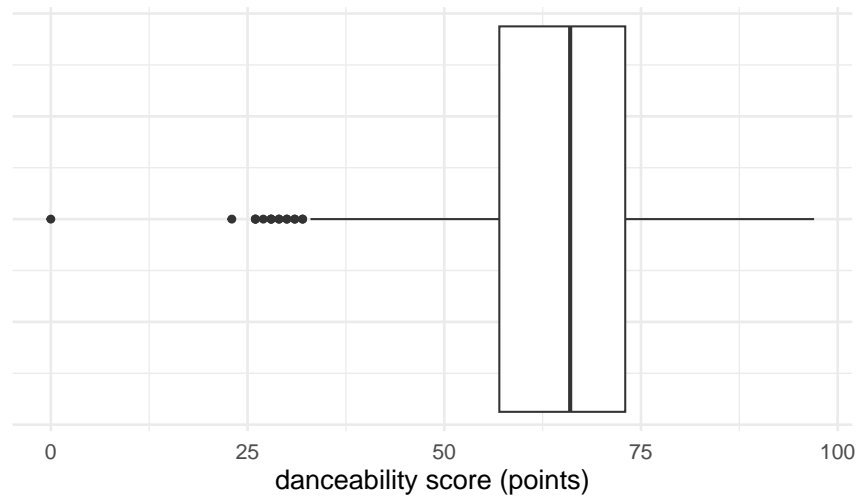
Summarizing quantitative variables

1. Describe the distribution of danceability of top songs over the past 10 years on Spotify.

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

Simulation methods to create a confidence interval

Unlike creation of the null distribution, the bootstrap distribution we use for creating a confidence interval is found by sampling with replacement from the original sample. To create one dot on the bootstrap distribution:

- Write the original values for the variable on n cards; one card for each observational unit.
- Sample with replacement from the cards n times.
- Plot the mean from each resampled sample on the bootstrap distribution.

Use the provided R script file to find a 95% confidence interval.

- Enter the name of the variable for `variable`.
- Enter the appropriate confidence level for `xx`.
- Highlight and run lines 22–25.

```
one_mean_CI(songs$variable, #Enter the name of the variable
            summary_measure = "mean", #choose the mean or median
            number_repetitions = 10000,  # Number of simulations
            confidence_level = xx)
```

3. Report the 95% confidence interval for the parameter of interest.

Theory-based methods to create a confidence interval

- **Conditions for the sampling distribution of $\bar{x}$ 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!)
  - **Normality Condition**: Either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:
    * $n < 30$: The distribution of the sample must be approximately normal with no outliers.
    * $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.
    * $n \geq 100$: Can assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Next we will calculate a theory-based confidence interval. To calculate a theory-based confidence interval for the a single mean, use the following formula:

$$\bar{x} \pm t^* \times SE(\bar{x}).$$

We will need to find the t^* multiplier using the function `qt()`.

- Enter the appropriate percentile in the R code to find the multiplier for a 95% confidence interval.

- Enter the degrees of freedom for `yy`. *The degrees of freedom for a single mean is $n - 1$.*

- Highlight and run line 31.

```
qt(percentile, df = yy, lower.tail=TRUE)
```

4. Mark on the t-distribution found below the values of $\pm t^*$. Draw a line at each multiplier and write the percentiles used to find each.

t Distribution with 602 df

151
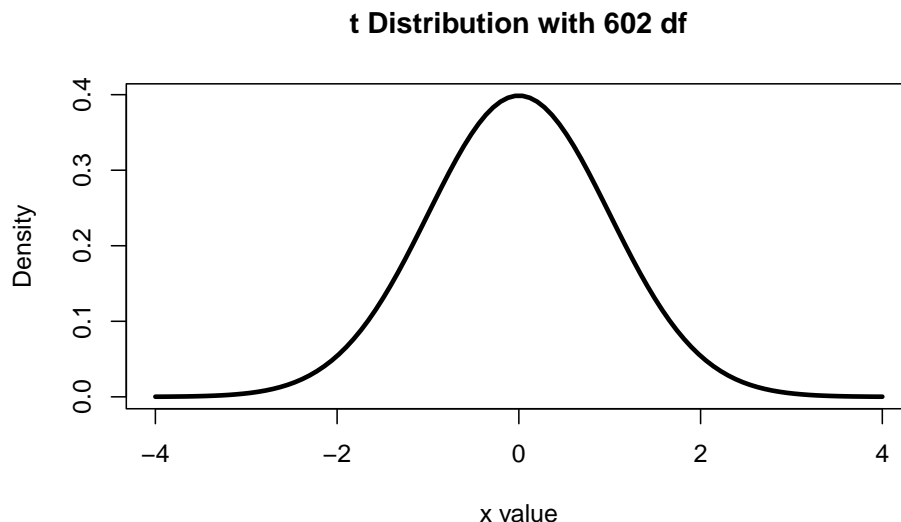
5. Calculate the margin of error using theory-based methods.

6. Calculate the confidence interval for the true mean using theory-based methods.

7. Interpret the confidence interval in context of the study.

8. Explain why the confidence interval with theory-based methods is similar to the confidence interval found using the bootstrap distribution.

7.3.4 Take-home messages

1. In order to use theory-based methods for a single mean, the independent observational units and normality conditions must be met.

2. The simulation-based confidence interval and theory-based confidence interval should be similar if the normality condition is met.

3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

7.3.5 Additional notes

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

7.4 Activity 15: Errors and Power

7.4.1 Learning outcomes

- Explain Type I and Type 2 errors in the context of a study.

- Explain the power of a test in the context of a study.

- Understand how changes in sample size, significance level, and the difference between the null value and the parameter value impact the power of a test.

- Understand how significance level impacts the probability of a Type 1 error.

- Understand the relationship between the probability of a Type 2 error and power.

- Be able to distinguish between practical importance and statistical significance.

7.4.2 Terminology review

In this activity, we will examine the possible errors that can be made based on the decision in a hypothesis test as well as factors influencing the power of the test. Some terms covered in this activity are:

- Significance level

- Type 1 error

- Type 2 error

- Power

To review these concepts, see Chapter 12 in the textbook.

7.4.3 College textbook cost

A college student spends, on average, $280 on textbooks per year. Many universities have started using open-source resources to help defray the cost of textbooks. One such university is hoping to show they have successfully reduced costs by $100 per year, on average.

1. Write the parameter of interest (μ) in words, in the context of this problem.

2. Use proper notation to write the null and alternative hypotheses the university would need to test in order to check their claim.

After determining hypotheses and prior to collecting data, researchers should set a **significance level** for a hypothesis test. The significance level, represented by α and most commonly 0.01, 0.05, or 0.10, is a cut-off for determining whether a p-value is small or not. The *smaller* the p-value, the *stronger* the evidence against the null hypothesis, so a p-value that is smaller than or equal to the significance level is strong enough evidence to *reject the null hypothesis*. Similarly, the *larger* the p-value, the *weaker* the evidence against the null hypothesis, so a p-value that is larger than the significance level does not provide enough evidence against the null hypothesis and the researcher would *fail to reject the null hypothesis*. Rejecting the null hypothesis or failing to reject the null hypothesis are the two **decisions** that can be made based on the data collected.

As you have already learned in this course, sample size of a study is extremely important. Often times, researchers will conduct what is called a power analysis to determine the appropriate sample size based on the goals of

their research, including a desired **power** of their test. Power is the probability of correctly rejecting the null hypothesis, or the probability of the data providing strong evidence against the null hypothesis *when the null hypothesis is false.*

The remainder of this activity will be spent investigating how different factors influence the power of a test, after which you will complete a power analysis for this university.

- Navigate to https://istats.shinyapps.io/power/.

- Choose the tab "Population Mean".

- Use the scale under "Null Hypothesis value μ_0" to change the value to your null value from question 2. *Note we will convert this to a scale in hundreds of dollars (e.g., 1 = \$100). In other words, use the null value of 2.8.

- Change the "Alternative Hypothesis" to the direction you wrote in question 2.

- Leave all boxes un-checked.

- Set the "True value of μ" to 2.8 as well.

- Do not change the scales for "Sample size n" or "Type I Error α" or "Population Std. Dev. σ".

The red distribution you see is the scaled-Normal distribution representing the null distribution for this hypothesis test, if the sample size was $n = 30$ and the significance level was $\alpha = 0.05$. This means the red distribution is showing the distribution of possible sample mean amounts spent on textbooks per year (in hundreds of dollars) for a sample of 30 college students ($\bar{x}$) if we assume the null hypothesis is true.

3. Based off this distribution and your alternative hypothesis, give one possible sample mean which you think would lead to rejecting the null hypothesis. Explain how you decided on your value.

4. Check the box for "Show Critical Value(s) and Rejection Region(s)". You will now see a vertical line on the plot indicating the *maximum* sample mean which would lead to reject the null hypothesis. That is, any sample means below this value would lead us to reject the null hypothesis; any sample means below this value would lead us to fail to reject the null hypothesis. What is this value?

5. Notice that there are some sample means under the red line (when the null hypothesis is true) which would lead us to reject the null hypothesis. Give the range of sample means which would lead to rejecting the null hypothesis when the null hypothesis is true? What is the statistical name for this mistake?

Check the "Type I Error" box under **Display**. This should verify (or correct) your answer to question 5! The area shaded in red represents the probability of making a **Type 1 Error** in our hypothesis test. Recall that a Type 1 error is when we reject the null hypothesis even though the null hypothesis is true. To reject the null hypothesis, the p-value, which was found assuming the null hypothesis is true, must be less than or equal to the significance level. Therefore the significance level is the probability of rejecting the null hypothesis when the null hypothesis is true, so the significance level IS the probability of making a Type 1 error in a hypothesis test!

6. Based on the current applet settings, what percent of the null distribution is shaded red (i.e., what is the probability of making a Type 1 error)?

Let's say this university believes their program can reduce the cost of textbooks for college students by $100 per year. In the applet, set the scale under "True value of μ" to 1.8.

7. Where is the blue distribution centered?

The blue distribution that appears represents what the university believes, that $180 (not $280) is the true mean textbook cost for college students at this university. This blue distribution represents the idea that the **null hypothesis is false**.

8. Consider the definition of power provided earlier in this activity. Do you believe the power of the test will be an area within the blue distribution or red distribution? How do you know? What about the probability of making a Type 2 error?

Check the "Type II Error" and "Power" boxes under **Display**. This should verify (or correct) your answers to question 8! The area shaded in blue represents the probability of making a **Type 2 Error** in our hypothesis test (failing to reject the null hypothesis even though the null hypothesis is false). The area shaded in green represents the power of the test. Notice that the Type 1 and Type 2 error rates and the power of the test are provided above the distribution.

9. Complete the following equation: Power + Type 2 Error Rate = _____. Explain why that equation makes sense. *Hint: Consider on what power and Type 2 error are conditional.*

Now let's investigate how changes in different factors influence the power of a test.

10. Using the same sample size and significance level, change the "True value of μ" to see the effect on power.

True value of p	2.0	1.5	1.0	0.05
Power				

11. What is changing about the simulated distributions pictured as you change the "True value of μ"?

12. How does increasing the distance between the null and believed true mean affect the power of the test?

13. Using the same significance level, set the "True value of μ" back to 1.8 and change the sample size to see its effect on power.

Sample Size	20	40	50	60	80
Power					

14. What is changing about the simulated distributions pictured as you change the sample size?

15. How does increasing the sample size affect the power of the test?

16. Using the same "True value of μ", set the sample size to 30 and change the "Type I Error α" to see the effect on power.

Type I Error α	0.01	0.03	0.05	0.10	0.15
Power					

17. What is changing about the simulated distributions pictured as you change the significance level?

18. How does increasing the significance level affect the power of the test?

19. Complete the power analysis for this university: The university believes they can reduce the cost of textbooks for their students by $100. They want to limit the probability of a type 1 error to 10% and the probability of a type 2 error to 15%. What is the minimum number of students the university will need to collect data on in order to meet these goals? Use the applet to answer this question.

20. Based on the goals outlined in question 19, which mistake below is the university more concerned about? In other words, which of the following two errors were the researchers trying to minimize. Explain your answer.

- Not being able to show their textbook cost is lower, on average, when their textbook cost really is lower.

- Advertising their textbook cost is lower, on average, even though it is not.

7.4.4 Take-home messages

1. There is a possibility of Type 1 error when we make the decision to reject the null hypothesis. Type 1 error: reject the null hypothesis when the null hypothesis is true. The probability of a Type 1 error when the null hypothesis is true is equal to the significance level, α.

2. There is a possibility of Type 2 error when we make the decision to fail to reject the null hypothesis. Type 2 error: fail to reject the null hypothesis when the null hypothesis is false.

3. Power of a test is the probability we reject the null when the null hypothesis is false. Power is equal to 1 minus the probability of a Type 2 error.

4. Changing the following will *increase* the power of the test:

   - *Increase* the sample size

   - *Increase* the significance level

   - *Increase* the distance between the null value and the parameter value (note that we don't have control over this!)

7.4.5 Additional notes

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

7.5   Module 6 and 7 Lab: Arsenic

7.5.1   Learning outcomes

- Given a research question involving one quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Investigate the process of creating a null distribution for one quantitative variable.

- Find, evaluate, and interpret a p-value from the null distribution.

- Use simulation-based methods to find a confidence interval for a single mean.

- Interpret a confidence interval for a single mean.

- Use a confidence interval to determine the conclusion of a hypothesis test.

7.5.2   Arsenic

Scientists have devised a new way to measure a person's level of arsenic poisoning by examining toenail clippings. Scientists measured the arsenic levels (in parts per million or ppm) in toenail clippings from 19 randomly selected individuals with private wells in New Hampshire. An arsenic level greater than 0.150 ppm is considered hazardous. Is there evidence the ground water in New Hampshire has hazardous levels of arsenic concentration (as seen in the arsenic levels of New Hampshire residents)? How high is the arsenic concentration for New Hampshire residents with a private well?

1. What does μ represent in the context of this study?

2. Notice that there are two research questions for this study. Identify which research question is best answered by finding a confidence interval and which is best answered by completing a hypothesis test?

3. Write out the null hypothesis in proper notation for this study.

4. What sign ($<$, $>$, or $\neq$) would you use in the alternative hypothesis for this study? Explain your choice.

- Upload and open the R script file for Week 12 lab.

- Upload and import the csv file, `arsenic`.

- Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 11.

- Enter the name of the variable in lines 15

- Write a title for the plot between the quotations and an x-axis label

- Highlight and run lines 1–21 to load the data and create a plot of the data.

- **Upload a screenshot of your plot to Gradescope**.

```
water <- read.csv("datasetname.csv")
water %>%
    summarise(favstats(variable))
water %>% # Data set piped into...
    ggplot(aes(x = variable)) +     # Name variable to plot
    geom_boxplot() +  # Create boxplot with specified binwidth
    labs(title = "Don't forget to title the plot!", # Title for plot
         x = "Enter an x-axis label! Don't forget the units!", # Label for x axis
         y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

5. Based on the plot, does there appear to be some support in favor of the alternative hypothesis? How do you know?

6. Interpret the value of Q_3 in context of the study.

7. What is the value of $\bar{x}$? What is the sample size?

8. **How far, on average, is each arsenic level from the mean arsenic level? What is the appropriate notation for this value?**

Use statistical inferential methods to draw inferences from the data

9. Using the provided graphs and summary statistics, determine if both theory-based methods and simulation-based methods could be used to analyze the data. Explain your reasoning.

Hypothesis test

Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that the average arsenic levels are not hazardous.

We will use the `one_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of sample means and compute a p-value.

10. Simulate a null distribution and compute the p-value, using the R script file for this lab.

```
one_mean_test(water$level_arsenic,    #Enter the name of the variable
              null_value = 0.150, #Enter the name of the null value
              summary_measure = "mean", #Choose mean or median to test
              shift = -0.122,  # Shift needed for bootstrap hypothesis test
              as_extreme_as = 0.272,  # Observed statistic
              direction = "greater",  # Direction of alternative
              number_repetitions = 10000)  # Number of simulated samples for null distribution
```

Sketch the null distribution created using the `one_mean_test` code.

Communicate the results and answer the research question

11. **Report the p-value. Based off of this p-value and a 1% significance level, what decision would you make about the null hypothesis? What potential error might you be making based on that decision?**

12. Do you expect the 98% confidence interval to contain the null value of zero? Explain.

Confidence interval

We will use the `one_mean_CI()` function in R (in the `catstats` package) to simulate a bootstrap distribution of sample means and calculate a confidence interval.

13. Using bootstrapping and the provided R script file, find a 98% confidence interval. Fill in the missing values/numbers in the `one_mean_CI()` function to create the 98% confidence interval.

```
one_mean_CI(data = water$level_variable, # Enter vector of differences
            summary_measure = "mean",   # Not needed when entering vector of differences
            number_repetitions = 10000, # Number of bootstrap samples for CI
            confidence_level = xx)  # Confidence level in decimal form
```

Report the 98% confidence interval in interval notation.

14. Write a paragraph summarizing the results of the study. **Upload a copy of your group's paragraph to Gradescope.** Be sure to describe:

- Summary statistic and interpretation
  - Summary measure (in context)
  - Value of the statistic

- 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

- 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)?

Paragraph (continued):

Exploratory Data Analysis and Simulation-based Hypothesis Testing for Two Categorical Variables

8.1 Vocabulary Review and Key Topics

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

8.1.1 Key topics

Module 8 will introduce exploratory data analysis and simulation-based hypothesis testing for two categorical variables. The **summary measure** for two categorical variables is a **difference in proportions**. We can also calculate a **relative risk** (ratio of proportions).

- Notation for a difference in sample proportions: $\hat{p}_1 - \hat{p}_2$, where 1 represents the 1st group of the explanatory variable and 2 represents the 2nd group

- Notation for a sample relative risk: $\dfrac{\hat{p}_1}{\hat{p}_2}$

- Notation for a difference in population proportions: $\pi_1 - \pi_2$

- Notation for a population relative risk: $\dfrac{\pi_1}{\pi_2}$

Types of plots for two categorical variables:

- Segmented bar plot

- Mosaic plot

We also explore study design and confounding variables.

8.1.2 Vocabulary

Sample statistics for two categorical variables

- **Difference in proportion**: $\hat{p}_1 - \hat{p}_2$

- **Relative risk**: the ratio of the conditional proportions:

$$\text{Relative Risk} = \frac{\hat{p}_1}{\hat{p}_2}$$

  - Interpretation of relative risk (RR): The risk of success in group 1 is RR times the risk of success in group 2.

- **Percent increase/decrease in risk**: an alternate way of interpreting the relative risk by first converting it into a percent increase or decrease in risk:

$$(RR - 1) \times 100\%$$

– If the quantity above is negative, the risk of success in group 1 is a decrease in risk compared to group 2; if positive, an increase.

– Interpretation of percent increase/decrease in risk: The risk of success in group 1 is xx% higher/lower than the risk of success in group 2.

Plotting two categorical variables

- **Segmented bar plot**: plots the conditional proportion of the response outcomes in each explanatory variable group. R code to create a segmented bar plot:

```
object %>%
    ggplot(aes(x = explanatory, fill = response))+ #Enter the variables to plot
    geom_bar(stat = "count", position = "fill") + #Creates a segmented bar plot
    labs(title = "Don't forget to title a plot!", #Make sure to title your plot
        y = "y-axis label", #y-axis label
        x = "x-axis label")  #x-axis label
```

– The plot shows no association between the variables if the height of each segment is approximately the same in each group.

- **Mosaic plot**: similar to the segmented bar plot but the sample size is reflected by the width of the bars. R code to create a mosaic plot:

```
object %>% # Data set piped into...
    ggplot() +   # This specifies the variables
    geom_mosaic(aes(x=product(explanatory), fill = response)) +  #Creates a mosaic plot
    labs(title = "Don't forget to title a plot!",  # Make sure to title your plot
        y = "y-axis label", #y-axis label
        x = "x-axis label")  #x-axis label
```

Hypotheses

- **Hypotheses in notation for a difference in proportions**: In the hypotheses below, the **null value** is equal to zero.

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{or} \quad H_0 : \pi_1 = \pi_2$$

$$H_A : \pi_1 - \pi_2 \begin{Bmatrix} < \\ \neq \\ < \end{Bmatrix} 0 \quad \text{or} \quad H_A : \pi_1 \begin{Bmatrix} < \\ \neq \\ < \end{Bmatrix} \pi_2$$

Simulation-based hypothesis testing for a difference in proportions

- **Conditions necessary to use simulation-based methods for inference for a two categorical variables**:

– **Independence**: observational units must be independent of one another both within and between groups.

- **Simulation-based methods to create the null distribution**: R code to use simulation-based methods for two categorical variables to find the p-value, `two_proportion_test` (from the `catstats` package), is shown below.

```
two_proportion_test(formula = response~explanatory, # response ~ explanatory
    data = object, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
```

```
    number_repetitions = 10000, # Always use a minimum of 10000 repetitions
    response_value_numerator = "xx", # Define which outcome is a success
    as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
    direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

- Conditions necessary to use simulation-based methods for inference for two categorical variables:

  - There must be independence of observational units within groups and between groups

Study design

- **Explanatory variable**: the variable researchers think *may be* affecting the other variable. (Some other textbooks call this the "independent" variable.)

- **Response variable**: the variable researchers think *may be* influenced by the other variable. (Some other textbooks call this the "dependent" variable.)

- **Observational study**: a study design in which observational units are merely "observed"; no manipulation is done. Examples include surveys and opinion polls.

- **Randomized experiment**: a study design where researchers **randomly assign** observational units to treatment groups (the explanatory variable). Examples include clinical trials where subjects are randomly assigned to either a placebo or a drug.

- **Confounding variable**: a third variable that is both (1) associated with the explanatory variable, and (2) associated with the response variable.

8.1.2.1 Scope of inference

- The **scope of inference** for a study answers two questions:

  1. To what population can my results be *generalized*?

  2. Does the study design allow us to assess whether changes in the explanatory variable *cause* changes in the response variable?

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type	
	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, **and** can generalize results to population.	Cannot conclude causal relationship, **but** can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, **but** cannot generalize results to a population.	Cannot conclude causal relationship, **and** cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

8.2 Video Notes: Inference for Two Categorical Variables using Simulation-based Methods

Read Sections 2.2 - 2.4, 15.1, 15.2, Chapter 4 and Chapter 16 in the course textbook. Use the following videos to complete the video notes for Module 8.

8.2.1 Course Videos

- 2.2to2.4
- 4.1_TwoProp
- 4.2_TwoProp
- 4.4
- 15.1
- 15.2
- RelativeRisk

Observational studies, experiments, and scope of inference: Video 2.2to2.4

- Review
  - Explanatory variable: the variable researchers think *may be* affecting the other variable.
  - Response variable: the variable researchers think *may be* influenced by the other variable.
- Confounding variable:
  - associated with both the explanatory and the response variable
  - explains the association shown by the data

Example:

Study design

- Observational study:

- Experiment:

Principles of experimental design

- Control: hold other differences constant across groups
- Randomization: randomized experiment

- Replication: large sample size or repeat of study

- Blocking: group based on certain characteristics

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the Journal of Experimental Psychology (Levin 2000): Human Perception and Performance investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the study design?

Example: 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.

- What is the study design?

- Assuming these data provide evidence that the low-dose aspirin group had a lower rate of heart attacks than the placebo group, is it valid for the researchers to conclude the lower rate of heart attacks was caused by the daily low-dose aspirin regimen?

Scope of Inference

1. How was the sample selected?

- Random sample with no sampling bias:

- Non-random sample with sampling bias:

2. What is the study design?

- Randomized experiment:

- Observational study:

Scope of Inference Table:

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type	
	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, **and** can generalize results to population.	Cannot conclude causal relationship, **but** can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, **but** cannot generalize results to a population.	Cannot conclude causal relationship, **and** cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the Journal of Experimental Psychology (Levin 2000): Human Perception and Performance investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the scope of inference for this study?

Summarizing two categorical variables - Video 4.1_TwoProp

- The summary measure for two categorical variables is the _____ in _____.

Notation used for the population difference in proportion:

- Two categorical variables:

    - Subscripts represent the _____ variable groups

Notation used for the sample difference in proportion:

- Two categorical variables

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.

For today's videos we will again use the `moving_to_mt` data set.

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")
```

To look at the relationship between the variable, `Age_Group` and the variable, `From` create the following two-way table using the `R` output below. Note, we are using `From` as the explanatory variable to predict whether a home sale has a buyer that is over or under the age of 30.

```
moving %>%
    group_by(Age_Group) %>% count(From) %>% print(n=8)
```

```
#> # A tibble: 8 x 3
#> # Groups:   Age_Group [2]
#>   Age_Group From      n
#>   <chr>     <chr> <int>
#> 1 Over30    CA        6
#> 2 Over30    CO        2
#> 3 Over30    MT       47
#> 4 Over30    WA       10
#> 5 Under30   CA        6
#> 6 Under30   CO        6
#> 7 Under30   MT       14
#> 8 Under30   WA        9
```

	State				
Age Group	CA	CO	MT	WA	Total
Over30	6	2	47	10	65
Under30	6	6	14	9	35
Total	12	8	61	19	100

- Using the table above, how many of the sampled home sales have buyers who were under 30 years old and from Montana?

If we want to know what proportion of each age group is from each state, we would calculate the proportion of home sales with buyers from each state within each age group. In other words, divide the number of home sales from each state with buyers that are over 30 by the total for row 1, the total number of home sales with buyers over 30.

- What proportion of sampled home sales with buyers under 30-years-old were from California?

- What notation should be used for this value?

Additionally, we could find the proportion of home sales with buyers in each state for each age group. Here we would calculate the proportion of home sales with buyers in each age group within each state. Divide the number of home sales with buyers in each age group from CA by the total for column 1, the total number of home sales with buyers from CA.

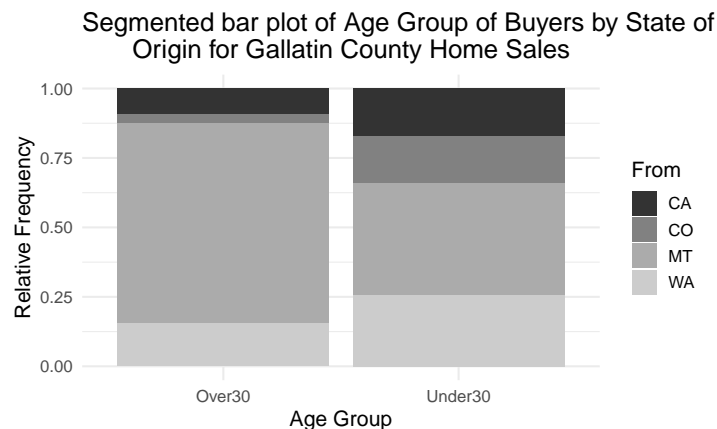|	State				
Age Group	CA	CO	MT	WA	Total
Over30	6	2	47	10	65
Under30	6	6	14	9	35
Total	12	8	61	19	100

- Using the table, calculate the proportion of home sales in Gallatin County with in-state buyers who are over 30 years old? Use appropriate notation with informative subscripts.

- Using the table, calculate the proportion of home sales in Gallatin County with California buyers who are over 30 years old? Use appropriate notation with informative subscripts.

- Calculate the difference in proportion of home sales in Gallatin County over 30 years old from other parts of Montana and from California. Use MT - CA as the order of subtraction. Give appropriate notation.

- Interpret the difference in proportion in context of the study.

Plots for two categorical variables - Video 4.2_TwoProp

In a segmented bar plot, the bar for each category will sum to 1. In this first plot, we are plotting the row proportions calculated conditional on the age group.

```
moving %>%
  ggplot(aes(x = Age_Group, fill = From))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Age Group of Buyers by State of
      Origin for Gallatin County Home Sales",
      #Title your plot
      y = "Relative Frequency", #y-axis label
      x = "Age Group") + #x-axis label
  scale_fill_grey()
```


Segmented bar plot of Age Group of Buyers by State of Origin for Gallatin County Home Sales

In this second plot, we are plotting the column proportions calculated conditional on the state of origin for the buyer.

```
moving %>%
  ggplot(aes(x = From , fill = Age_Group))+ #Enter variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of State of Origin of Buyers by Age
      Group for Gallatin County Home Sales",
      #Title your plot
      y = "Relative Frequency", #y-axis label
      x = "State of Origin") + #x-axis label
  scale_fill_grey()
```

Segmented bar plot of State of Origin of Buyers by Age
Group for Gallatin County Home Sales
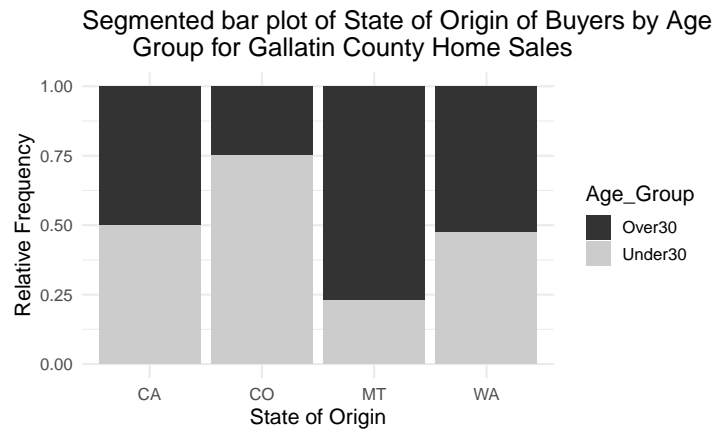
Mosaic plot:

```r
moving$Age_Group <- factor(moving$Age_Group, levels = c("Under30", "Over30"))
moving %>% # Data set piped into...
  ggplot() +    # This specifies the variables
  geom_mosaic(aes(x=product(From), fill = Age_Group)) +
    # Tell it to make a mosaic plot
  labs(title = "Mosaic plot of State of Origin Segmented by
Age Group for Gallatin County Home Sales",
      # Title your plot
      x = "State of Origin",   # Label the x axis
      y = "") +  # Remove y axis label
    scale_fill_grey(guide = guide_legend(reverse = TRUE)) # Make figure color
```

Mosaic plot of State of Origin Segmented by
Age Group for Gallatin County Home Sales
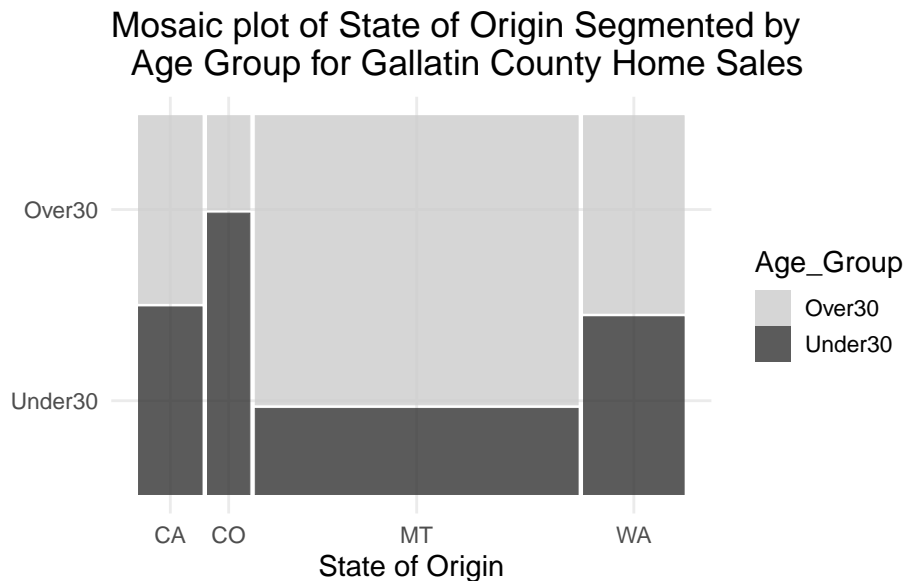
- Why is the bar for MT the widest on the mosaic plot?

Simpson's paradox - Video 4.4

- When an apparent _____ between explanatory and response variables reverses when accounting for _____ variable.

Example: The "Berkeley Dataset" contains all 12,763 applicants to UC-Berkeley's graduate programs in Fall 1973. This dataset was published by UC Berkeley researchers in an analysis to understand the possible gender bias in admissions and has now become a classic example of Simpson's Paradox.

```r
discrim <- read.csv ("data/berkeley.csv")

discrim %>%
  ggplot(aes(x =Gender, fill = Admission))+
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Sex of Berkley Applicants by
       Admission Status",
       y = "Relative Frequency",
       x = "Sex") +
  scale_fill_grey()
```
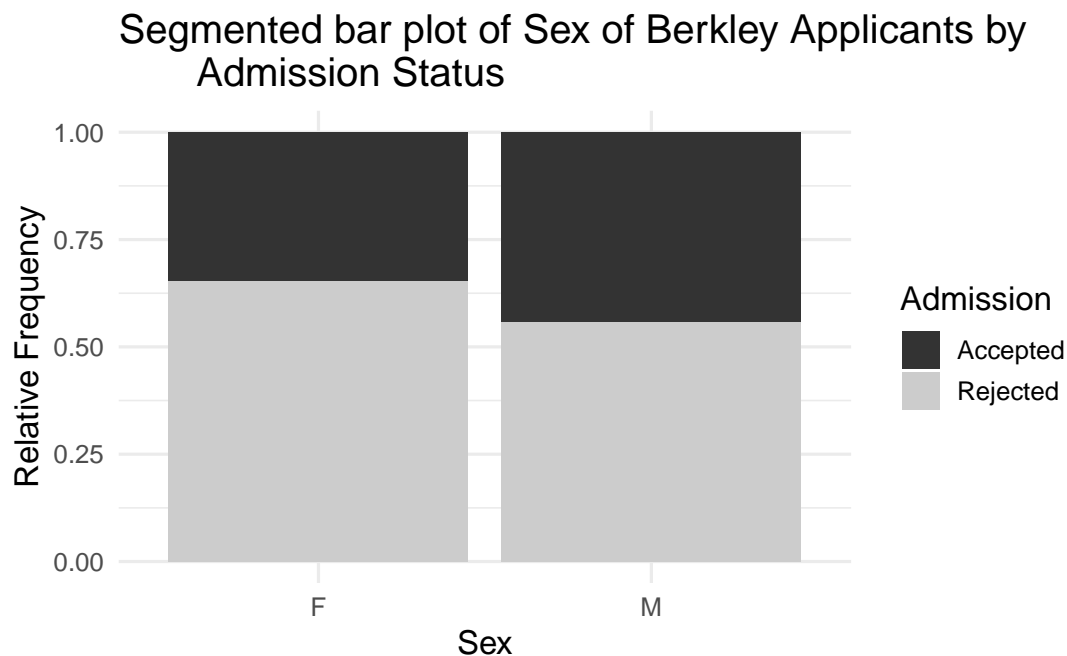

The data showed that 44% of male applicants were accepted and 35% of female applicants were accepted. Does it appear that the female students are discriminated against?

We can break down the data by major. A major code (either A, B, C, D, E, F, or Other) was used.

Here we look at the relationship between admission status and sex for Program A and for Program B.
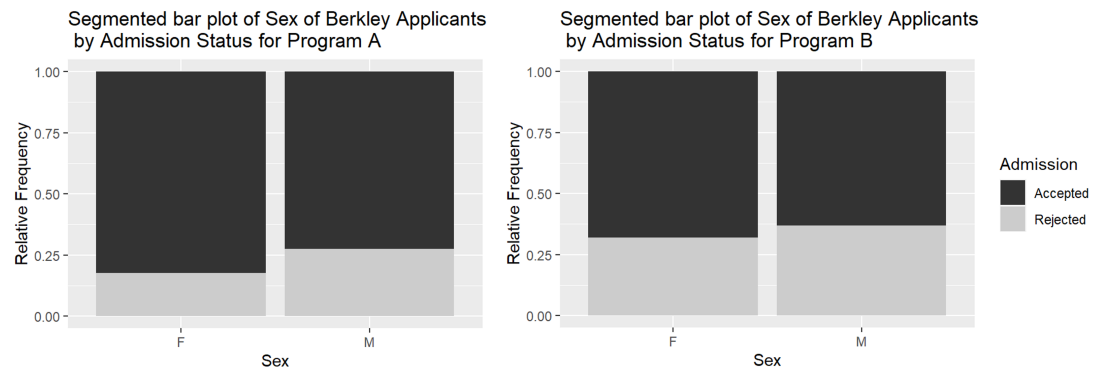
Showing Program C and Program D.
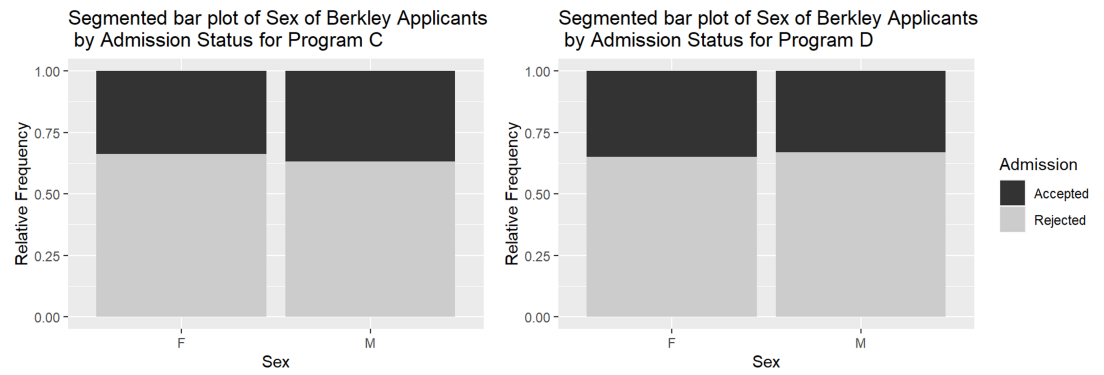
And finally, Program E and F.
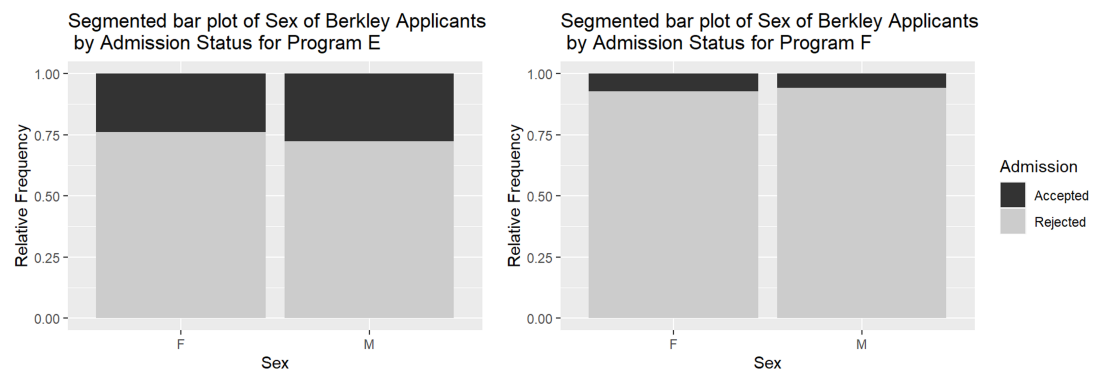
We can see in several programs the acceptance rate is higher for females than for males.

Two categorical variables - Video 15.1

- In this module, we will study inference for a _____ explanatory variable

  and a _____ response.

Example: In a double-blind experiment (Weiss 1988) on 48 cocaine addicts hoping to overcome their addiction, half were randomly assigned to a drug called desipramine and the other half a placebo. The addicts were followed for 6 weeks to see whether they were still clean. Is desipramine more effective at helping cocaine addicts overcome their addiction than the placebo?

Observational units:

Explanatory variable:

Response variable:

Notation:

- Population proportion for group 1:

- Population proportion for group 2:

- Sample proportion for group 1:

- Sample proportion for group 2:

- Sample difference in proportions:

- Sample size for group 1:

- Sample size for group 2:

Hypothesis Testing

Conditions:

- Independence: the response for one observational unit will not influence another observational unit

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.

Write the null and alternative hypotheses for the cocaine study:

In notation:

H_0 :

H_A :

Summary statistics and plot

```
cocaine %>% group_by(drug) %>% count(outcome)
```

```
#> # A tibble: 4 x 3
#> # Groups:   drug [2]
#>   drug         outcome       n
#>   <chr>        <chr>     <int>
#> 1 desipramine clean        14
#> 2 desipramine relapsed     10
#> 3 placebo      clean         4
#> 4 placebo      relapsed     20
```

Summary statistic:

Interpretation:


```
cocaine%>%
  ggplot(aes(x = drug, fill = outcome))+
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Bar plot of Type of Drug, Segmented by
       Outcome for Cocaine Addicts",
       y = "Relative Frequency",
       x = "Drug or Placebo") +
    scale_fill_grey()
```
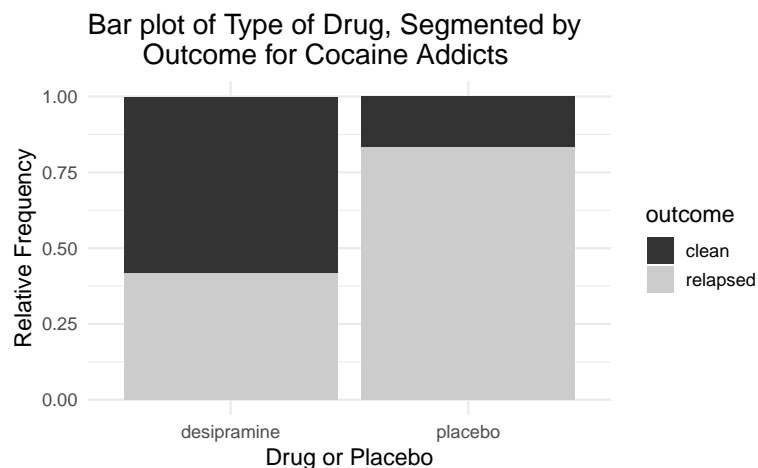
Is the independence condition met for simulation inference?

Simulation-based method

- Simulate many samples assuming $H_0 : \pi_1 = \pi_2$
  - Write the response variable values on cards
  - Mix the explanatory variable groups together
  - Shuffle cards into two explanatory variable groups to represent the sample size in each group (n_1 and n_2)
  - Calculate and plot the simulated difference in sample proportions from each simulation
  - Repeat 10000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\hat{p}_1 - \hat{p}_2$

```
set.seed(216)
two_proportion_test(formula = outcome~drug, # response ~ explanatory
    data = cocaine, # Name of data set
    first_in_subtraction = "desipramine", # Order of subtraction: enter the name of Group 1
    number_repetitions = 10000, # Always use a minimum of 10000 repetitions
    response_value_numerator = "clean", # Define which outcome is a success
    as_extreme_as = 0.417, # Calculated observed statistic (difference in sample proportions)
    direction="greater") # Alternative hypothesis direction ("greater","less","two-sided")
```
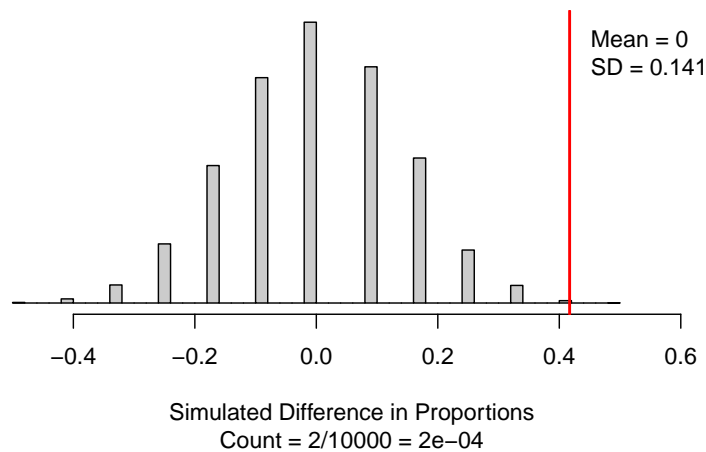

Mean = 0
SD = 0.141

Simulated Difference in Proportions
Count = 2/10000 = 2e−04

Explain why the null distribution is centered at the value of zero:

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 with scope of inference:

- Amount of evidence

- Parameter of interest

- Direction of the alternative hypothesis

- Generalization

- Causation

Confidence interval - Video 15.2

To estimate the difference in true proportion we will create a confidence interval.

Simulation-based method

- Write the response variable values on cards

- Keep explanatory variable groups separate

- Sample with replacement n_1 times in explanatory variable group 1 and n_2 times in explanatory variable group 2

- Calculate and plot the simulated difference in sample proportions from each simulation

- Repeat 10000 times (simulations) to create the bootstrap distribution

- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

Returning to the cocaine example, we will estimate the difference in true proportion of cocaine addicts that stay clean for those on the desipramine and those on the placebo.

```
set.seed(216)
two_proportion_bootstrap_CI(formula = outcome ~ drug,
        data=cocaine, # Name of data set
        first_in_subtraction = "desipramine", # Order of subtraction: enter the name of Group 1
        response_value_numerator = "clean", # Define which outcome is a success
        number_repetitions = 10000, # Always use a minimum of 10000 repetitions
        confidence_level = 0.99) # Enter the level of confidence as a decimal
```
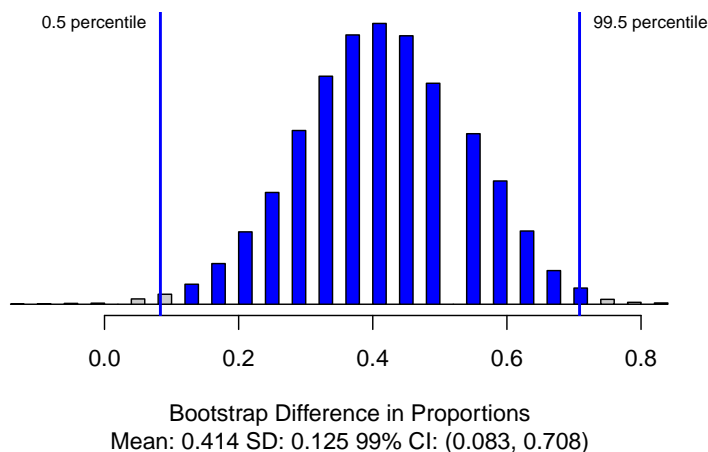

Bootstrap Difference in Proportions
Mean: 0.414 SD: 0.125 99% CI: (0.083, 0.708)

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

Relative Risk - Video RelativeRisk

- Relative risk is the ratio of the risks in two different categories of an explanatory variable.

Relative Risk:

Example: In a study reported in the New England Journal of Medicine (Du Toit 2015), one-hundred fifty (150) children who had shown sensitivity to peanuts were randomized to receive a flour containing a peanut protein or a placebo flour for 2.5 years. At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no). 71% of those in the peanut flour group no longer demonstrated a peanut allergy compared to 2% of those in the placebo group.

- Calculate the relative risk of desensitization comparing the peanut flour group to the placebo group.

- Interpretation:

  - The proportion of successes in group 1 is the RR _____ the proportion of successes in group 2.

Increase in risk:

- Interpretation:

  - The proportion of successes in group 1 is the $(RR-1)$ _____ higher/lower than the proportion of successes in group 2.

Percent increase in risk:

- Interpretation:

  - The proportion of successes in group 1 is the $(RR-1) \times 100$ _____ higher/lower than the proportion of successes in group 2.

- Interpret the value of relative risk from the peanut study in context of the problem.

- Find the increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

- Find the percent increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

Within the peanut flour group, the percent desensitized within each age group (at start of study) is as follows:

1-year-olds: 71%; 2-year-olds: 35%; 3-year-olds: 19%

- Calculate the relative risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.

- Interpret the percent increase (or decrease) in risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.

Relative risk in the news

People 50 and older who have had a mild case of covid-19 are 15% more likely to develop shingles (herpes zoster) within six months than are those who have not been infected by the coronavirus, according to research published in the journal Open Forum Infectious Diseases (Bhavsar 2022).

- What was the calculated relative risk of developing shingles when comparing those who has mild COVID-19 to those who had not had COVID-19, among the 50 and older population?

Testing Relative Risk

In Unit 2, we tested for a difference in proportion. We could also test for relative risk.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

8.2.2 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. Explain why the null distribution is centered at the value of zero.

2. Does the confidence interval agree with the p-value?

3. What is the difference between a mosaic plot and a segmented bar plot?

4. What does relative risk measure?

8.3 Activity 16: Study Design

8.3.1 Learning outcomes

- Explain the purpose of random assignment and its effect on scope of inference.

- Identify whether a study design is observational or an experiment.

- Identify confounding variables in observational studies and explain why they are confounding.

8.3.2 Terminology review

In this activity, we will examine different study designs, confounding variables, and how to determine the scope of inference for a study. Some terms covered in this activity are:

- Scope of inference

- Explanatory variable

- Response variable

- Confounding variable

- Experiment

- Observational study

To review these concepts, see Sections 2.2 through 2.5 in the textbook.

8.3.3 Atrial fibrillation

Atrial fibrillation is an irregular and often elevated heart rate. In some people, atrial fibrillation will come and go on its own, but others will experience this condition on a permanent basis. When atrial fibrillation is constant, medications are required to stabilize the patient's heart rate and to help prevent blood clots from forming. Pharmaceutical scientists at a large pharmaceutical company believe they have developed a new medication that effectively stabilizes heart rates in people with permanent atrial fibrillation. They set out to conduct a trial study to investigate the new drug. The scientists will need to compare the proportion of patients whose heart rate is stabilized between two groups of subjects, one of whom is given a placebo and the other given the new medication.

1. Identify the explanatory and response variable in this trial study.

   Explanatory variable:

   Response variable:

Suppose 24 subjects with permanent atrial fibrillation have volunteered to participate in this study. There are 16 subjects that self-identified as male and 8 subjects that self-identified as female.

2. One way to separate into two groups would be to give all the males the placebo and all the females the new drug. Explain why this is not a reasonable strategy.

3. Could the scientists fix the problem with the strategy presented in question 2 by creating equal sized groups by putting 4 males and 8 females into the drug group and the remaining 12 males in the placebo group? Explain your answer.

4. A third strategy would be to **block** on sex. In this type of study, the scientists would assign 4 females and 8 males to each group. Using this strategy, out of the 12 individuals in each group what **proportion** are males?

5. Assume the scientists used the strategy in question 4, but they put the four tallest females and eight tallest males into the drug group and the remaining subjects into the placebo group. They found that the proportion of patients whose heart rate stabilized is higher in the drug group than the placebo group.

   Could that difference be due to the sex of the subjects? Explain your answer.

   Could it be due to other variables? Explain your answer.

While the strategy presented in question 5 controlled for the sex of the subject, there are more potential **confounding variables** in the study. A confounding variable is a variable that is *both*

1. associated with the explanatory variable, *and*
2. associated with the response variable.

When both these conditions are met, if we observe an association between the explanatory variable and the response variable in the data, we cannot be sure if this association is due to the explanatory variable or the confounding variable—the explanatory and confounding variables are "confounded."

Random assignment means that subjects in a study have an equally likely chance of receiving any of the available treatments.

6. You will now investigate how randomly assigning subjects impacts a study's scope of inference.

- Navigate to the "Randomizing Subjects" applet under the "Other Applets" heading at: http://www.rossmanchance.com/ISIapplets.html. This applet lists the sex and height of each of the 24 subjects. Click "Show Graphs" to see a bar chart showing the sex of each subject. Currently, the applet is showing the strategy outlined in question 3.

- Click "Randomize".

In this random assignment, what proportion of males are in group 1 (the placebo group)?

What proportion of males are in group 2 (the drug group)?

What is the difference in proportion of males between the two groups (placebo - drug)?

7. Notice the difference in the two proportions is shown as a dot in the plot at the bottom of the web page. Un-check the box for Animate above "Randomize" and click "Randomize" again. Did you get the same difference in proportion of males between the placebo and drug groups?

8. Change "Replications" to 998 (for 1000 total). Click "Randomize" again. Sketch the plot of the distribution of difference in proportions from each of the 1000 random assignments here. Be sure to include a descriptive x-axis label.

9. Does random assignment *always* balance the placebo and drug groups based on the sex of the participants? Does random assignment *tend* to make the placebo and drug groups *roughly* the same with respect to the distribution of sex? Use your plot from question 8 to justify your answers.

10. Change the drop-down menu below Group 2 from "sex" to "height". The applet now calculates the average height in the placebo and drug groups for each of the 1000 random assignments. The dot plot displays the distribution of the difference in mean heights (placebo - drug) for each random assignment. Based on this dot plot, is height distributed equally, on average, between the two groups? Explain how you know.

The diagram below summarizes these ideas about confounding variables and random assignment. When a confounding variable is present (such as sex or height), and an association is found in a study, it is impossible to discern what caused the change in the response variable. Is the change the result of the explanatory variable or the confounding variable? However, if all confounding variables are *balanced* across the treatment groups, then only the explanatory variable differs between the groups and thus *must have caused* the change seen in the response variable.
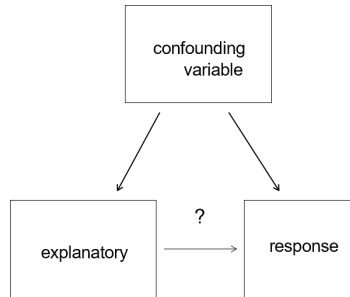
11. What is the purpose of random assignment of the subjects in a study to the explanatory variable groups? Cross out the arrow in the figure above that is eliminated by random assignment.

12. Suppose in this study on atrial fibrillation, the scientists did randomly assign groups and found that the drug group has a higher proportion of subjects whose heart rates stabilized than the placebo group. Can the scientists conclude the new drug *caused* the increased chance of stabilization? Explain your answer.

13. Is the sample of subjects a simple random sample or a convenience sample?

14. Both the sampling method and the study design will help to determine the *scope of inference* for a study: To *whom* can we generalize, and can we conclude *causation or only association*? Use your answers to question 12 and 13 and the table on the next page to determine the scope of inference of this trial study described in question 12.

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type	
	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, **and** can generalize results to population.	Cannot conclude causal relationship, **but** can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, **but** cannot generalize results to a population.	Cannot conclude causal relationship, **and** cannot generalize results to a population.

Inferences to population can be made

Can only generalize to those similar to the sample due to potential sampling bias

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

8.3.4 Scope of Inference

The two main study designs we will cover are **observational studies** and **experiments**. In observational studies, researchers have no influence over which subjects are in each group being compared (though they can control other variables in the study). An experiment is defined by assignment of the treatment groups of the *explanatory variable*, typically via random assignment.

For the next exercises identify the study design (observational study or experiment), the sampling method, and the scope of inference.

15. The pharmaceutical company Moderna Therapeutics, working in conjunction with the National Institutes of Health, conducted Phase 3 clinical trials of a vaccine for COVID-19 in the Fall of 2021. US clinical research sites enrolled 30,000 volunteers without COVID-19 to participate. Participants were randomly assigned to receive either the candidate vaccine or a saline placebo. They were then followed to assess whether or not they developed COVID-19. The trial was double-blind, so neither the investigators nor the participants knew who was assigned to which group.

    Study design:

    Sampling method:

    Scope of inference:

16. In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for one year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.

    Study design:

    Sampling method:

    Scope of inference:

8.3.5 Take-home messages

1. The study design (observational study vs, experiment) determines if we can draw causal inferences or not. If an association is detected, a randomized experiment allows us to conclude that there is a causal (cause-and-effect) relationship between the explanatory and response variable. Observational studies have potential confounding variables within the study that prevent us from inferring a causal relationship between the variables studied.

2. Confounding variables are variables not included in the study that are related to both the explanatory and the response variables. When there are potential confounding variables in the study we cannot draw causal inferences.

3. Random assignment balances confounding variables across treatment groups. This eliminates any possible confounding variables by breaking the connections between the explanatory variable and the potential confounding variables.

4. Observational studies will always carry the possibility of confounding variables. Randomized experiments, which use random assignment, will have no confounding variables.

8.3.6 Additional notes

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

8.4 Activity 17: Summarizing Two Categorical Variables

8.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question involving categorical variables.

- Plots for association between two categorical variables: segmented bar plot, mosaic plot.

- Calculate and interpret relative risk

8.4.2 Terminology review

In today's activity, we will review summary measures and plots for categorical variables. Some terms covered in this activity are:

- Conditional proportions

- Segmented bar plots

- Mosaic plots

- Relative risk

To review these concepts, see Chapter 4 in the textbook.

8.4.3 Graphing categorical variables

Follow these steps to upload the necessary R script file for today's activity:

- Download the RScript file for this Activity from D2L

- Upload the file to the RStudio server

- Open the RScript file

Highlight and run lines 1–2 to load the packages needed for today's activity and load the data set. Notice the use of the # symbol in the R script file. The # sign 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 # sign when executing the code. Refer to the instructions following the # sign to understand what you need to enter in the code.

Nightlight use and myopia

In a study reported in Nature (Quinn et al. 1999), a survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of two had a higher incidence of nearsightedness (myopia) later in childhood.

In this study, there are two variables studied: `Light`: level of light in room at night (no light, nightlight, full light) and `Sight`: level of myopia developed later in childhood (high myopia, myopia, no myopia).

1. Which variable is the explanatory variable? Which is the response variable?

An important part of understanding data is to create visual pictures of what the data represent. In this activity, we will create graphical representations of categorical data.

R code

The line of code shown below (line 5 in the R script file) reads in the data set and names the data set `myopia`. Highlight and run line 5 in the R script file to load the data from the Stat 216 webpage.

```
# This will read in the data set
myopia <- read.csv("https://math.montana.edu/courses/s216/data/ChildrenLightSight.csv")
```

2. Click on the data set name (`myopia`) in the Environment tab (upper right window). This will open the data set in a 2nd tab in the Editor window (upper left window). R is case sensitive, which means that you must always type the name of a variable EXACTLY as it is written in the data set including upper and lower case letters and without misspellings! Write down the name of each variable (column names) as it is written in the data set.

Summarizing two categorical variables

Is there an association between the level of light in a room and the development of myopia? Fill in the name of the explanatory variable, `Light` for explanatory and name of the response variable, `Sight` in line 9 in the R script file, highlight and run line 9 to get the counts for each combination of levels of variables.

```
myopia %>% group_by(explanatory) %>% count(response)
```

3. Fill in the following table with the values from the R output.

	Light Level			
Myopia Level	Full Light	Nightlight	No Light	Total
High Myopia				
Myopia				
No Myopia				
Total				

In the following questions, use the table to calculate the described proportions. Notation is important for each calculation. Since this is sample data, it is appropriate to use statistic notation for the proportion, $\hat{p}$. When calculating a proportion dependent on a single level of a variable, subscripts are needed when reporting the notation.

4. Calculate the proportion of children with no myopia. Use appropriate notation.

5. Calculate the proportion of children with no myopia among those that slept with full light. Use appropriate notation.

6. Calculate the proportion of children with no myopia among those that slept with no light. Use appropriate notation.

7. Calculate the difference in proportion of children with no myopia for those that slept with full light minus those who slept with no light. Give the appropriate notation. Use full light minus no light as the order of subtraction.

8. Interpret the calculated difference in proportion in context of the study.

Displaying two categorical variables

Two types of plots can be created to display two categorical variables. To examine the differences in level of myopia for the level of light, we will first create a segmented bar plot of `Light` segmented by `Sight`. To create the segmented bar plot enter the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in the R script file in line 14. Highlight and run lines 13–20.

```
myopia %>% # Data set piped into...
ggplot(aes(x = explanatory, fill = response)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Night Light Use by Level of Myopia",
       # Make sure to title your plot
       x = "Level of Light",   # Label the x axis
       y = "")  + # Remove y axis label
  scale_fill_grey()  # Make figure black and weight
```

9. Sketch the segmented bar plot created here. Be sure to label the axes.

10. From the segmented bar plot, which level of light has the highest proportion of `No Myopia`?

11. Based on the plot, is there an association between level of light and level of myopia?

We could also plot the data using a mosaic plot which is shown below.

```
myopia$Sight <- factor(myopia$Sight, levels = c("No Myopia", "Myopia", "High Myopia"))
myopia %>% # Data set piped into...
  ggplot() +    # This specifies the variables
  geom_mosaic(aes(x=product(Light), fill = Sight)) +  # Tell it to make a mosaic plot
  labs(title = "Mosaic Plot of Night Light Use by Level of Myopia",  # Make sure to title your plot
       x = "Level of Light",   # Label the x axis
       y = "") +  # Remove y axis label
  scale_fill_grey(guide = guide_legend(reverse = TRUE))  # Make figure black and white
```

Mosaic Plot of Night Light Use by Level of Myopia
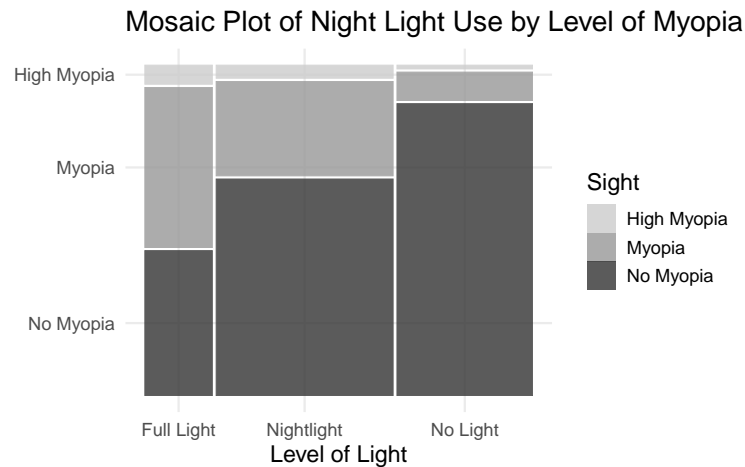
12. What is similar and what is different between the segmented bar chart and the mosaic bar chart?

13. Explain why the bar for `Nightlight` is the widest in the mosaic plot.

Relative Risk

14. Calculate the relative risk of myopia for children that slept with full light compared to those that slept with no light.

15. Interpret the value of relative risk in context of the problem.

16. Calculate the percent increase/decrease in risk of myopia for children that slept with full light compared to those that slept with no light.

17. Interpret as a percent increase/decrease in risk in context of the problem.

8.4.4 Take-home messages

1. Bar charts can be used to graphically display a single categorical variable either as counts or proportions. Segmented bar charts and mosaic plots are used to display two categorical variables.

2. Segmented bar charts always have a scale from 0 - 100%. The bars represent the outcomes of the explanatory variable. Each bar is segmented by the response variable. If the heights of each segment are the same for each bar there is no association between variables.

3. Mosaic plots are similar to segmented bar charts but the widths of the bars also show the number of observations within each outcome.

8.4.5 Additional notes

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

8.5 Activity 18: The Good Samaritan

8.5.1 Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Investigate the process of creating a null distribution for two categorical variables

- Find and evaluate a p-value from the null distribution

8.5.2 Terminology review

In today's activity, we will use simulation-based methods to analyze two categorical variables. Some terms covered in this activity are:

- Conditional proportion

- Null hypothesis

- Alternative hypothesis

To review these concepts, see Chapter 15 in your textbook.

8.5.3 The Good Samaritan

Researchers at the Princeton University wanted to investigate influences on behavior (Darley and Batson 1973). The researchers randomly selected 67 students from the Princeton Theological Seminary to participate in a study. Only 47 students chose to participate in the study, and the data below includes 40 of those students (7 students were removed from the study for various reasons). As all participants were theology majors planning a career as a preacher, the expectation was that all would have a similar disposition when it comes to helping behavior. Each student was then shown a 5-minute presentation on the Good Samaritan, a parable in the Bible which emphasizes the importance of helping others. After the presentation, the students were told they needed to give a talk on the Good Samaritan parable at a building across campus. Half the students were told they were late for the presentation; the other half told they could take their time getting across campus (the condition was randomly assigned). On the way between buildings, an actor pretending to be a homeless person in distress asked the student for help. The researchers recorded whether the student helped the actor or not. The results of the study are shown in the table below. Do these data provide evidence that those in a hurry will be less likely to help people in need in this situation? Use the order of subtraction hurry – no hurry.

	Hurry Condition	No Hurry Condition	Total
Helped Actor	2	11	13
Did Not Help Actor	18	9	27
Total	20	20	40

These counts can be found in R by using the `count()` function:

- Download the R script file from D2L and upload to the RStudio server.

- Highlight and run lines 1–7 to get the counts for each group.

```
# Read data set in
good <- read.csv("https://math.montana.edu/courses/s216/data/goodsam.csv")
good %>% group_by(Condition) %>% count(Behavior)

#> # A tibble: 4 x 3
#> # Groups:   Condition [2]
#>   Condition Behavior      n
#>   <chr>     <chr>     <int>
```

```
#> 1 Hurry     Help        2
#> 2 Hurry     No help    18
#> 3 No hurry  Help       11
#> 4 No hurry  No help     9
```

Ask a research question

The research question as stated above is: Do these data provide evidence that those in a hurry will be less likely to help people in need in this situation? In order to set up our hypotheses, we need to express this research question in terms of parameters.

Remember, we define the parameter for a single categorical variable as the true proportion of observational units that are labeled as a "success" in the response variable.

For this study we are identifying two parameters and looking at the difference between these two parameters.

- π_{hurry} = long-run proportion of Princeton Theological Seminary students assigned to hurry that helped the actor

- $\pi_{\text{no hurry}}$ = long-run proportion of Princeton Theological Seminary students assigned not to hurry that helped the actor

- $\pi_{\text{hurry}} - \pi_{\text{no hurry}}$ = the difference in long-run proporiton of Princeton Theological Seminary Students that helped the actor between those who were assigned to hurry and those who were not assigned to hurry

When comparing two groups, we assume the two parameters are equal in the null hypothesis—there is no association between the variables.

1. Write the null hypothesis out in words.

2. Based on the research question, fill in the appropriate sign for the alternative hypothesis ($<$, $>$, or $\neq$):

   $H_A : \pi_{\text{hurry}} - \pi_{\text{no hurry}}$ _____ 0

Summarize and visualize the data

To create the segmented bar plot:

- Enter the name of the explanatory variable for explanatory

- Enter the name of the response variable for response

- Highlight and run lines 13–20

```
good %>%
  ggplot(aes(x = explanatory, fill = response))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented Bar Plot of Princeton Seminary Students that Help the actor \nbetween those tha
       y = "Relative Frequency", #y-axis label
       x = "Condition") + #x-axis label
  scale_fill_grey()
```

3. Based on the segmented bar plot, is there an association between whether a Seminary student helps the actor and condition assigned?

195

4. Using the two-way table given in the introduction, calculate the conditional proportion of students in the hurry condition who helped the actor. Use appropriate notation.

5. Using the two-way table given in the introduction, calculate the conditional proportion of students in the no hurry condition who helped the actor. Use appropriate notation.

6. Calculate the summary statistic (difference in sample proportion) for this study. Use Hurry - No hurry as the order of subtraction. Use appropriate notation.

Interpretation of the summary statistic:

The proportion of Princeton Theological Seminary students that helped the actor is 0.45 less for those assigned to hurry compared to those assigned not to hurry.

Hypothesis Test

We will now simulate a **null distribution** of sample differences in proportions. The null distribution is created under the assumption the null hypothesis is true.

7. Using the cards provided by your instructor, simulate one sample under the assumption the null hypothesis is true.

- Start with 40 cards (13 labeled helped, 27 labeled did not help)

- Mix the cards together

- Shuffle the cards into two piles (20 in hurry, 20 in no hurry)

- Calculate the proportion of simulated students that helped in each group.

- Report the difference in proportion of simulated students that helped (hurry - no hurry)

The segmented bar plot below shows the relationship between the variables for **one simulation assuming the null hypothesis is true**.

Segmented bar plot of Simulated Data
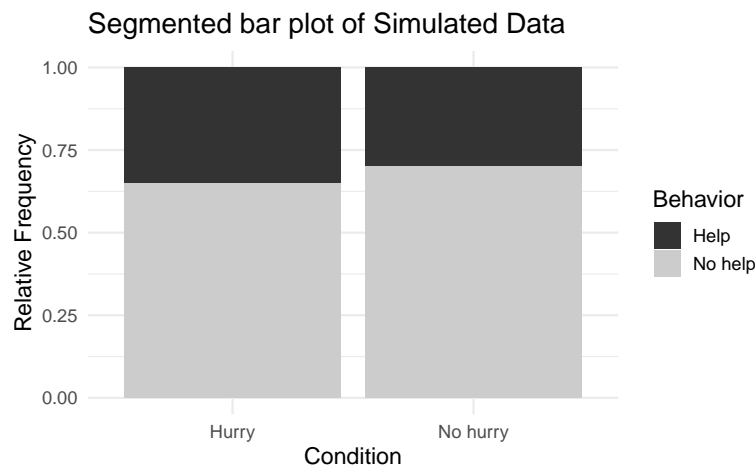
To create the null distribution of differences in sample proportions, we will use the `two_proportion_test()` function in R (in the `catstats` package). We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `good`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the direction of the alternative hypothesis.

The response variable name is `Behavior` and the explanatory variable name is `Condition`.

8. What inputs should be entered for each of the following to create the simulation?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"Hurry"` or `"No hurry"`):

- Number of repetitions:

- Response value numerator (What is the outcome for the response variable that is considered a success? `"Help"` or `"No help"`):

- As extreme as (enter the value for the sample difference in proportions):

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

Using the R script file for this activity, enter your answers for question 8 in place of the **xx**'s to produce the null distribution with 10000 simulations; highlight and run lines 24–30.

```
two_proportion_test(formula = Behavior~Condition, # response ~ explanatory
    data = good, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
    number_repetitions = 10000, # Always use a minimum of 10000 repetitions
    response_value_numerator = "xx", # Define which outcome is a success
    as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
    direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

9. Sketch the null distribution created here.

10. Explain why the null distribution is centered around the value of zero?

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

12. Write a conclusion in context of the study.

8.5.4 Take-home messages

1. When comparing two groups, we are looking at the difference between two parameters. In the null hypothesis, we assume the two parameters are equal, or that there is no difference between the two proportions.

2. To create one simulated sample on the null distribution for a difference in sample proportions, label $n_1 + n_2$ cards with the response variable outcomes from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2, representing the explanatory variable groups. Calculate and plot the difference in proportion of successes.

8.5.5 Additional notes

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

Theory-based Hypothesis Testing and Simulation-based and Theory-based Confidence Intervals for Two Categorical Variables:

9.1 Vocabulary Review and Key Topics

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

9.1.1 Key topics

Module 9 introduces theory-based hypothesis testing methods and both simulation-based and theory-based confidence intervals for two categorical variables.

9.1.2 Vocabulary

Theory-based inference

- **Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution**: The following conditions must be met in order to use theory-based methods for two categorical variables.

    - **Independence**: the sample's observations are independent both within and between the two groups. (*Remember*: This also must be true to use simulation-based methods!)

    - **Success-failure condition**: we *expect* to see at least 10 successes and 10 failures in the *each* sample. We consider this condition to be met if we observe at least 10 successes and 10 failures in our data set in both groups: $n_1\hat{p}_1 \geq 10$, $n_1(1 - \hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1 - \hat{p}_2) \geq 10$. Equivalently, we check that all four cells in the table have at least 10 observations.

- **Standard error of a difference in sample proportions assuming the null is true**:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pooled} \times (1 - \hat{p}_{pooled}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\hat{p}_{pooled}$ is the **pooled sample proportion**: the total number of successes divided by the total sample size $(n_1 + n_2)$.

- **Standardized difference in sample proportion**:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

    - Measures the number of standard errors the sample difference in proportions is above or below the null value of zero

    - If the conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution are met, and if the true difference in proportions is equal to zero, the standardized difference in sample proportions, Z, will have an approximate *standard* normal distribution.

- Use the `pnorm` function in R to find a theory-based p-value for a hypothesis test involving a difference in proportions by finding the area under a standard normal distribution where Z is as or more extreme as the value observed (in the direction of H_A).

- **Standard error of a difference in sample proportions for a confidence interval** (not assuming the null is true):

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

  - Calculation of the confidence interval for a difference in sample proportions:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

  - Use the `qnorm` function in R to find the z^* multiplier.

Simulation-based confidence interval

- Review the conditions necessary for simulation-based inference in Module 8.

- R code to find the simulation-based confidence interval using the `two_proportion_bootstrap_CI` function from the `catstats` package.

```
two_proportion_bootstrap_CI(formula = response~explanatory,
    data=object, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
    response_value_numerator = "xx", # Define which outcome is a success
    number_repetitions = 10000, # Always use a minimum of 10000 repetitions
    confidence_level = xx) # Enter the level of confidence as a decimal
```

9.2 Video Notes: Theoretical Inference for Two Categorical Variables

Read Sections 15.3 and 15.4 in the course textbook. Use the following videos to complete the video notes for Module 9.

9.2.1 Course Videos

- 15.3TheoryTests
- 15.3TheoryIntervals

Hypothesis testing using theory-based methods - Video 15.4TheoryTests

Example: In Modules 3 and 4, we investigated data on higher education institutions in the United States, collected by the Integrated Postsecondary Education Data System (IPEDS) for the National Center for Education Statistics (NCES) (Education Statistics 2018). A random sample of 2900+ higher education institutions in the United States was collected in 2018. Two variables measured on this data set is whether the institution is a land grant university and whether the institution offers tenure. Does the proportion of universities that offer tenure differ between land grant and non-land-grant institutions?

What is the explanatory variable?

What is the response variable?

Write the parameter of interest:

Hypotheses:

In notation:

H_0 :

H_A :

```
IPED <-read.csv("https://math.montana.edu/courses/s216/data/IPEDS_2018.csv")

IPEDS <- IPED %>%
    drop_na(Tenure)

IPEDS %>% # Data set piped into...
    ggplot(aes(x = LandGrant, fill = Tenure)) +    # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Tenure Availability
       by Type of Institution for Higher Ed Institutions",
       # Make sure to title your plot
       x = "Land Grant",   # Label the x axis
       y = "") + # Remove y axis label
```

```
    scale_fill_grey()

IPEDS %>% group_by(LandGrant) %>% count(Tenure)
```

```
#> # A tibble: 4 x 3
#> # Groups:   LandGrant [2]
#>   LandGrant Tenure     n
#>   <chr>     <chr>  <int>
#> 1 No        No       976
#> 2 No        Yes     1829
#> 3 Yes       No        31
#> 4 Yes       Yes       72
```

Segmented Bar Plot of Tenure Availability
by Type of Institution for Higher Ed Institutions
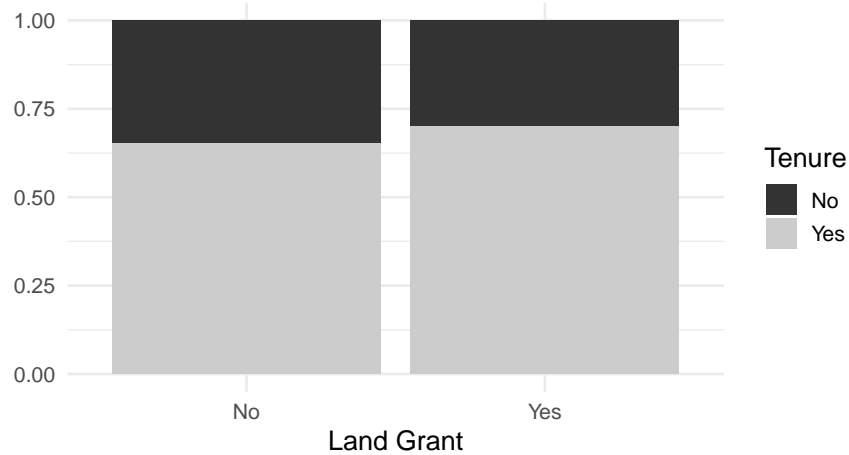
Report the summary statistic:

Conditions for inference using theory-based methods for two categorical variables:

- Independence: the response for one observational unit will not influence another observational unit

- Large enough sample size:

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

Steps to use theory-based methods:

- Calculate the standardized statistic
- Find the area under the standard normal distribution at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample proportions assuming the null hypothesis is true:

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

Equation for the standardized difference in sample proportions:

- This value measures how many standard errors the sample difference in proportions is above/below the null value.

Calculate the standardized difference in sample proportion of higher education institutions that offer tenure between land grant universities and non-land grant universities.

- First calculate the standard error of the difference in proportion assuming the null hypothesis is true
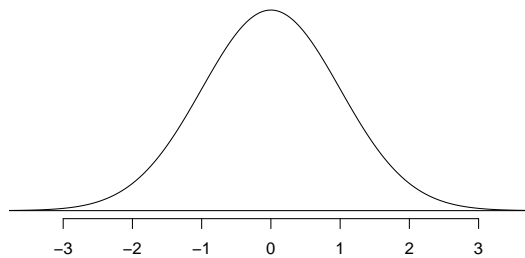
- Then calculate the Z score

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(0.985, lower.tail = FALSE)*2
```

```
#> [1] 0.3246241
```

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 with scope of inference:

- Amount of evidence

- Parameter of interest

- Direction of the alternative hypothesis

- Generalization

- Causation

Confidence interval - Video 15.3TheoryIntervals

- Estimate the _____ in true _____
- $CI = \text{statistic} \pm \text{margin of error}$

Theory-based method for a two categorical variables

- $CI = \hat{p}_1 - \hat{p}_2 \pm (z^* \times SE(\hat{p}_1 - \hat{p}_2))$

- When creating a confidence interval, we no longer assume the _____ hypothesis is true. Use the sample _____ to calculate the sample to sample variability, rather than $\hat{p}_{pooled}$.

Equation for the standard error of the difference in sample proportions *NOT* assuming the null is true:

Example: Estimate the difference in true proportions of higher education institutions that offer tenure between land grant universities and non-land grant universities.

Find a 90% confidence interval:

- 1st find the z^* multiplier

```r
qnorm(0.95, lower.tail=TRUE)
```

```
#> [1] 1.644854
```

- Next, calculate the standard error for the difference in proportions **NOT** assuming the null hypothesis is true

- Calculate the margin of error

- Calculate the endpoints of the 90% 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

9.2.2 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 for the difference in sample proportions?

2. Explain why a theory-based confidence interval for the Good Samaritan study from last module would NOT be similar to the bootstrap interval created.

9.3 Activity 19: Winter Sports Helmet Use and Head Injuries — Theory-based Methods

9.3.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.

- Create and interpret a theory-based confidence interval for a difference in proportions.

- Calculate and interpret the standardized difference in sample proportion

- Use the standard normal distribution to find the p-value for the test

9.3.2 Terminology review

In today's activity, we will use theory-based methods to estimate the difference in two proportions. Some terms covered in this activity are:

- Standard normal distribution

- Independence and success-failure conditions

To review these concepts, see Chapter 15 in your textbook.

9.3.3 Winter sports helmet use and head injury

In this activity we will focus on theory-based methods to calculate a confidence interval. The sampling distribution of a difference in proportions can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution:

- **Independence**: The data are independent within and between the two groups. (*Remember*: This also must be true to use simulation methods!)

- **Success-failure condition**: This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.

A study was reported in "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), on the use of helmets and head injuries for skiers and snowboarders involved in accidents. The summary results from a random sample of 3562 skiers and snowboarders involved in accidents is shown in the two-way table below.

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

- Download the R script file from D2L and upload to the RStudio server

- Highlight and run 1–13 to import the data set and create the segmented bar plot

```
skiers <- read.csv("https://www.math.montana.edu/courses/s216/data/HeadInjuries.csv") # Read data set in
skiers %>% # Data set piped into...
  ggplot(aes(x = Helmet, fill = Outcome)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Head Injuries for Skiers/Snowboarders
      Involved in Injuries between Helmet Use",  # Make sure to title your plot
```

```
    x = "Helmet Use",    # Label the x axis
    y = "") +    # Remove y axis label
scale_fill_grey()  # Make figure black and white
```

Segmented Bar Plot of Head Injuries for Skiers/Snowbo Involved in Injuries between Helmet Use
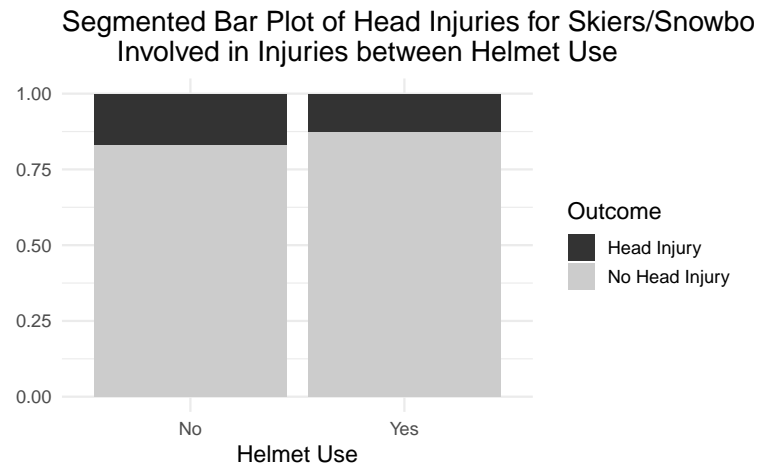
1. Verify the independence condition is met.

2. Verify the success failure condition is met to use theory-based methods.

3. Calculate the difference in sample proportion of skiers and snowboarders involved in accidents with a head injury for those who wear helmets and those who do not. Use appropriate notation with informative subscripts.

Hypothesis test

4. Write the null and alternative hypotheses in notation.

H_0:

H_A:

Use statistical analysis methods to draw inferences from the data

To test the null hypothesis, we could use simulation-based methods as we did in the activities in Module 8. In this activity, we will focus on theory-based methods. Like with a single proportion, the sampling distribution of a difference in sample proportions can be mathematically modeled using the normal distribution if certain conditions are met.

To calculate the standardized statistic we use:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \text{null value}}{SE_0(\hat{p}_1 - \hat{p}_2)},$$

where the null standard error is calculated using the pooled proportion of successes:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

For this study we would first calculate the pooled proportion of successes.

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}}$$

5. Calculate the pooled proportion of head injuries.

6. Use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$ assuming the null hypothesis is true.

7. Use the value of the null standard error to calculate the standardized statistic (standardized difference in proportion).

8. Mark the value of the standardized difference in proportion on the standard normal distribution shown below. Interpret the standardized statistic in context of the problem.
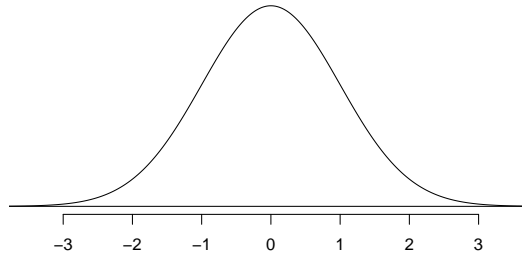
We will use the `pnorm()` function in R to find the p-value.

- Enter the value of z from question 7 for xx

- Highlight and run lines 18–20

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

9. Write a conclusion to the test.

How would an increase in sample size impact the p-value of the test?

	Helmet Use	No Helmet Use	Total
Head Injury	135	674	809
No Head Injury	921	3270	4191
Total	1056	3944	5000

Note that the sample proportions for each group are the same as the smaller sample size.

$$\hat{p}_h = \frac{135}{1056} = 0.128, \ \ \hat{p}_n = \frac{674}{3944} = 0.171$$

First calculate the pooled proportion of successes.

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{809}{5000} = 0.162$$

209

We use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$.

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{1056} + \frac{1}{3944} \right)} = 0.013$$

Standardized Statistic Calculation:

$$Z = \frac{0.128 - 0.171 - 0}{0.013} = -3.308$$

Use Rstudio to find the p-value for this new sample.

```
pnorm(-3.308, # Enter value of standardized statistic
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
      lower.tail=TRUE) # Gives a p-value greater than the standardized statistic
```

```
#> [1] 0.000469824
```

10. How does the increase in sample size affect the p-value?

9.3.4 Take-home messages

1. Simulation-based methods and theory-based methods should give similar results for a study *if the validity conditions are met*. For both methods, observational units need to be independent. To use theory-based methods, additionally, the success-failure condition must be met. Check the validity conditions for each type of test to determine if theory-based methods can be used.

2. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*.

3. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and a larger standardized statistic.

9.3.5 Additional notes

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

9.4 Activity 20: Diabetes

9.4.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Describe and perform a simulation-based confidence interval for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.

9.4.2 Glycemic control in diabetic adolescents

Researchers compared the efficacy of two treatment regimens to achieve durable glycemic control in children and adolescents with recent-onset type 2 diabetes (Group 2012). A convenience sample of patients 10 to 17 years of age with recent-onset type 2 diabetes were randomly assigned to either a medication (rosiglitazone) or a lifestyle-intervention program focusing on weight loss through eating and activity. Researchers measured whether the patient still needs insulin (failure) or had glycemic control (success). Of the 233 children who received the Rosiglitazone treatment, 143 had glycemic control, while of the 234 who went through the lifestyle-intervention program, 125 had glycemic control. Is there evidence that there is difference in proportion of patients that achieve durable glycemic control between the two treatments? Use Rosiglitazone – Lifestyle as the order of subtraction.

- Upload and open the R script file. Upload the csv file, `diabetes`.
- Enter the name of the data set for `datasetname.csv` in the R script file in line 7.
- Highlight and run lines 1–8 to get the counts for each combination of categories.

```
glycemic <- read.csv("datasetname.csv")
glycemic %>% group_by(treatment) %>% count(outcome)
```

1. Is this an experiment or an observational study?

2. Complete the following two-way table using the R output.

	Treatment		
Outcome	rosiglitazone	lifestyle	Total
glycemic control (success)			
insulin required (failure)			
Total			

3. Is the independence condition met for this study? Explain your answer.

4. Is the success failure condition met for this study? Explain your answer.

5. Write the parameter of interest for the research question.

6. **Calculate the summary statistic (difference in proportions). Use appropriate notation.**

Simulation methods

First we will use simulation methods to find the confidence interval. This will give an interval estimate for the parameter of inference.

We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a 90% confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `glycemic`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

7. What inputs should be entered for each of the following to create the bootstrap simulation?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"rosi"` or `"lifestyle"`):

- Number of repetitions:

- Response value numerator (What is the outcome for the response variable that is considered a success? `"success"` or `"failure"`):

- confidence_level:

- Fill in the missing values/names in the R script file in the two_proportion_bootstrap_CI function to create a simulation 90% confidence interval.

- Highlight and run lines 12–17

```
two_proportion_bootstrap_CI(formula = response~explanatory,
        data=glycemic, # Name of data set
        first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
        response_value_numerator = "xx", # Define which outcome is a success
        number_repetitions = 10000, # Always use a minimum of 10000 repetitions
        confidence_level = xx) # Enter the level of confidence as a decimal
```

8. Report the 90% confidence interval.

9. Interpret the confidence interval in context of the problem.

Theory-based Methods

Next we will use theory-based methods to find the 90% confidence interval. Review the conditions for using theory-based methods from Activity 19.

10. Is the sample size large enough to use theory-based methods to find the confidence interval? Explain in context of the study,

To find a confidence interval for the difference in proportions we will add and subtract the margin of error from the point estimate to find the two endpoints.

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \quad \text{where}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

In this formula, we use the sample proportions for each group to calculate the standard error for the difference in proportions since we are not assuming that the true difference is zero.

11. Calculate the standard error of the sample proportion not assuming the null hypothesis is true.

Recall that the z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 90%, we find the Z values that encompass the middle 90% of the standard normal distribution. If 90% of the standard normal distribution should be in the middle, that leaves 10% in the tails, or 5% in each tail. The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, $90\% + 5\% = 95\%$ percentile).

```
qnorm(0.95, lower.tail = TRUE) # Multiplier for 90% confidence interval
```

```
#> [1] 1.644854
```

12. Mark the value of the z^* multiplier and the percentages used to find this multiplier on the standard normal distribution shown below.
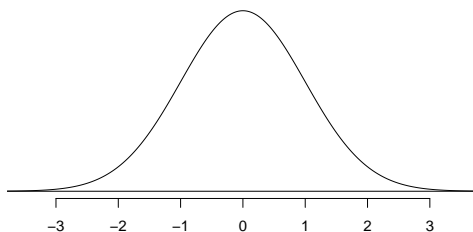
Remember that the margin of error is the value added and subtracted to the sample difference in proportions to find the endpoints for the confidence interval.

$$ME = z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

13. Using the multiplier of $z^* = 1.645$ and the calculated standard error, calculate the margin of error for a 90% confidence interval.

14. Calculate the 90% confidence interval for the parameter of interest.

9.4.3 Take-home messages

1. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true.* For a confidence interval, we are not assuming a null hypothesis, so we use the values of the two conditional proportions to calculate the standard error. Make note of the difference in these two formulas.

2. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and a larger standardized statistic.

3. Since we add and subtract the margin of error to the point estimate, the margin of error is half the width of the confidence interval.

9.4.4 Additional notes

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

9.5 Module 8 and 9 Lab: Poisonous Mushrooms

9.5.1 Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a difference in proportions.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in proportions.

- Interpret and evaluate a confidence interval for a simulation-based confidence interval for a difference in proportions.

9.5.2 Poisonous Mushrooms

Wild mushrooms, such as chanterelles or morels, are delicious, but eating wild mushrooms carries the risk of accidental poisoning. Even a single bite of the wrong mushroom can be enough to cause fatal poisoning. An amateur mushroom hunter is interested in finding an easy rule to differentiate poisonous and edible mushrooms. They think that the mushroom's gills (the part which holds and releases spores) might be related to a mushroom's edibility. They used a data set of 8124 mushrooms and their descriptions. For each mushroom, the data set includes whether it is edible (e) or poisonous (p) and the size of the gills (broad (b) or narrow (n)). Is there evidence gill size is associated with whether a mushroom is poisonous? PLEASE NOTE: According to The Audubon Society Field Guide to North American Mushrooms, there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

- Upload and open the R script file for the Module 9 lab. Upload and import the csv file, `mushrooms_edibility`.

- Enter the name of the data set (see the environment tab) for datasetname.csv in the R script file in line 8.

- Highlight and run lines 1–9 to get the counts for each combination of categories.

```
mushrooms <- read.csv("datasetname.csv") # Read data set in
mushrooms %>% group_by(gill_size) %>% count(edibility) #finds the counts in each group
```

1. What is the explanatory variable? How are the two levels of the explanatory variable written in the data set?

2. What is the response variable? How are the two levels of the response variable written in the data set?

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

4. Write the null hypothesis for this study in notation.

5. **Using the research question, write the alternative hypothesis in words.**

6. Fill in the following two-way table using the R output.

	Gill Size			
Edibility	Broad (b)	Narrow (n)	Total
Poisonous (p)			
Edible (e)			
Total			

7. **Calculate the difference in proportion of mushrooms that are poisonous for broad gill mushrooms and narrow gill mushrooms. Use broad - narrow for the order of subtraction. Use appropriate notation.**

- Fill in the missing values/names in the R script file for the `two-proportion_test` function to create the null distribution and find the p-value for the test.

```
two_proportion_test(formula = response~explanatory, # response ~ explanatory
    data= mushrooms, # Name of data set
    first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
    number_repetitions = 10000, # Always use a minimum of 10000 repetitions
    response_value_numerator = "xx", # Define which outcome is a success
    as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
    direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

8. Report the p-value for the study.

9. **Do you expect that a 90% confidence interval would contain the null value of zero? Explain your answer.**

- Fill in the missing values/names in the R script file in the two_proportion_bootstrap_CI function to create a simulation 90% confidence interval.

- **Upload a copy of the bootstrap distribution to Gradescope.**

```
two_proportion_bootstrap_CI(formula = response~explanatory,
        data=mushrooms, # Name of data set
        first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
        response_value_numerator = "xx", # Define which outcome is a success
        number_repetitions = 10000, # Always use a minimum of 10000 repetitions
        confidence_level = xx) # Enter the level of confidence as a decimal
```

10. Report the 90% confidence interval.

11. Write a paragraph summarizing the results of the study as if writing a press release. 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 your group's confidence interval interpretation and conclusion to Gradescope.

Paragraph:

MODULE 10

Unit 2 Review

The following section contains both a list of key topics covered in Unit 2 as well as Module Review Worksheets.

10.0.1 Key Topics

Review the key topics for Unit 2 to review prior to the exams. All of these topics will be covered in Modules 6–9.

10.0.2 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 2 Review folder after the study sessions.

10.1   Key Topics Exam 2

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 the study design (observational study or randomized experiment).

6. Identify the sampling method and potential types of sampling bias (non-response, response, selection).

7. Calculate and interpret the difference in proportions, relative risk, and percent increase/decrease in risk for a study involving two categorical variables.

Hypothesis testing

8. Identify which of the two scenarios applies to the study: one quantitative variable or two categorical variables.

9. Write the parameter of interest in words and correct notation.

10. Find the value of the observed statistic (point estimate, summary statistic). Use correct notation.

11. State the null and alternative hypotheses in words and in correct notation.

12. Verify the validity condition is met to use simulation-based methods to find a p-value.

13. Verify the validity conditions are met to use theory-based methods to find a p-value from the theoretical distribution.

14. In a simulation-based hypothesis test, describe how to create one dot on a dotplot of the null distribution using coins, cards, or spinners.

15. Explain where the null distribution is centered and why.

16. Describe and illustrate how R calculates the p-value for a simulation-based test.

17. Describe and illustrate how R calculates the p-value for a theory-based test.

18. 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.

19. Calculate and interpret the standard error of the statistic under the null using the correct formula on the Golden ticket.

20. Calculate and interpret the appropriate standardized statistic using the correct formula on the Golden ticket.

21. Interpret the p-value in context of the study: it is the probability of _____, assuming _____.

22. Evaluate the p-value for strength of evidence against the null: how much evidence does the p-value provide against the null?

23. Write a conclusion about the research question based on the p-value.

24. Given a significance level, what decision can be made about the research question based on the p-value.

25. Describe which features of the study could be changed to increase power and how.

26. Describe which features of the study impact the p-value and how.

27. Write a Type I error in context of the problem.

28. Write a Type II error in context of the problem.

29. Interpret power in context of the problem.

30. Based on your p-value, identify what type of error could have occurred.

Confidence interval

31. Describe how to simulate one bootstrapped sample using cards.

32. Explain where the bootstrap distribution is centered and why.

33. Find an appropriate percentile confidence interval using a bootstrap distribution from R output.

34. Verify the validity condition is met to use simulation-based methods to find the confidence interval.

35. Verify the validity conditions are met to use theory-based methods to calculate a confidence interval.

36. Describe and illustrate how the bootstrap distribution is used to find the confidence interval for a given confidence level.

37. Describe and illustrate how the standard normal distribution or t-distribution is used to find the multiplier for a given confidence level.

38. Calculate and interpret the standard error of the statistic (not assuming the null hypothesis) using the correct formula on the Golden ticket

39. Calculate the appropriate margin of error and confidence interval using theory-based methods.

40. Interpret the confidence interval in context of the study.

41. 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.

42. Interpret the confidence level in context of the study. What does "confidence" mean?

43. Describe which features of the study have an effect on the width of the confidence interval and how.

10.2 Module 6 Review - One Mean Testing

There are about 4 million tourists to Yellowstone National Park per year. One of the most visited sites within the park is the Old Faithful Geyser. The reason this geyser is called old faithful is because of the regularity of eruptions. Tourists report a typical wait time of 30 minutes, on average. A sample of 299 tourists reported their wait time to see Old Faithful erupt. Is there evidence that the average wait time differs from 30 minutes?

```
#>   min Q1 median Q3 max     mean       sd   n missing
#> 1  43 59     76 83 108 72.31438 13.89032 299       0
```

The following code created the boxplot of waiting time.

Boxplot of Tourist Reported Wait Time for Old Faithrul to Erupt
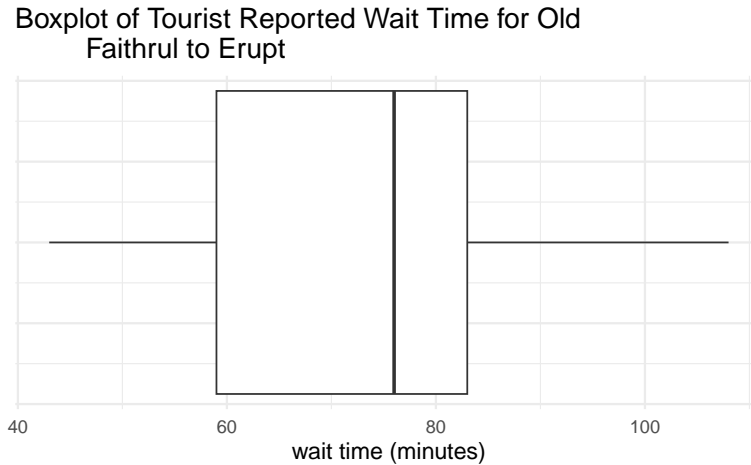
1. Report and interpret the value of Q_1 in context of the study.

2. Report and interpret the standard deviation of wait time in context of the study.

3. Descripe the plot using the four characteristics for boxplots.

4. Write the parameter of interest for this study in context of the study.

5. Write the null hypothesis in notation.

6. Write the alternative hypothesis in words.

We will start with simulation methods.

7. Calculate the difference $\mu_0 - \bar{x}$. Will we need to shift the data up or down?

```
set.seed(216)
one_mean_test(data = geyser$waiting,    #Object and variable
              null_value = 30, #null value for the study
              shift = -42.31438,    #Shift needed for bootstrap hypothesis test
              summary_measure = "mean",
              as_extreme_as = 72.314,   #Observed statistic
              direction = "two-sided",   #Direction of alternative
              number_repetitions = 10000)   #Number of simulated samples for null distribution
```


Mean = 30.002
SD = 0.804

Simulated Mean
Count = 0/10000 = 0
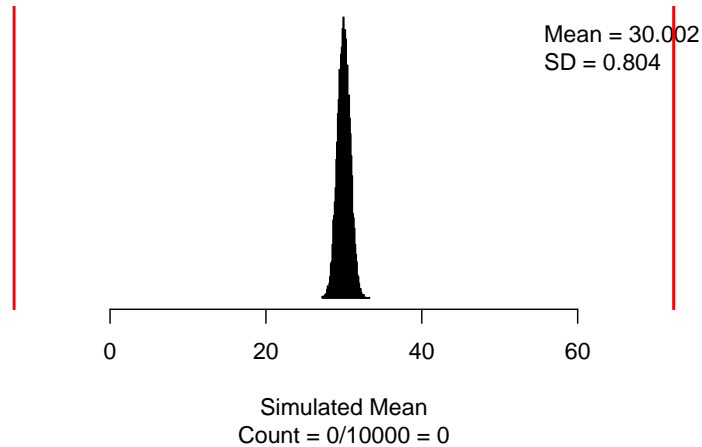
8. Interpret the p-value of the test.

Now let's focus on theory-based methods.

Conditions for the sampling distribution of $\bar{x}$ 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!)

- **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

  - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $30 \leq n < 100$: If the sample size n is betwe 30 and 100 and there are no particularly extreme outliers in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

9. Is the independence condition met?

10. Is the normality condition met to use theory-based methods?

To find the standardized statistic for the mean we will use the following formula:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

where the standard error of the sample mean difference is:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

11. Calculate the standard error of the sample mean.

12. Calculate the standardized mean for the study.

13. Mark on the t-distribution shown below on how to find the p-value of the test.

t Distribution with 298 df
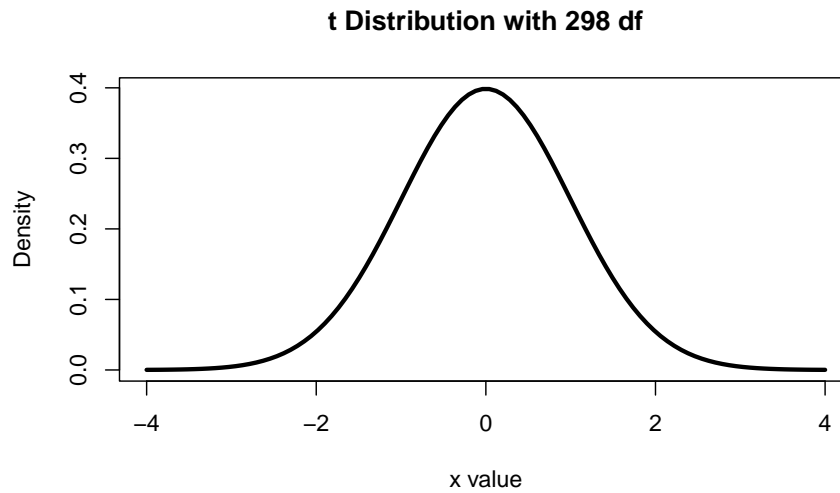
14. Interpret the standardized mean in context of the study.

The following code calculates the p-value for the study.

```
2*pt(-52.676, df=298, lower.tail=TRUE)
#> [1] 5.045442e-153
```

15. Write a conclusion to the test.

10.3 Module 7 Review - One Mean Confidence Interval

There are about 4 million tourists to Yellowstone National Park per year. One of the most visited sites within the park is the Old Faithful Geyser. The reason this geyser is called old faithful is because of the regularity of eruptions. Tourists report a typical wait time of 30 minutes, on average. A sample of 299 tourists reported their wait time to see Old Faithful erupt. How long, on average, do tourists wait for Old Faithful to erupt?

```
#>   min Q1 median Q3 max     mean       sd   n missing
#> 1  43 59     76 83 108 72.31438 13.89032 299       0
```

The following code created the boxplot of waiting time.

Boxplot of Tourist Reported Wait Time for Old Faithrul to Erup
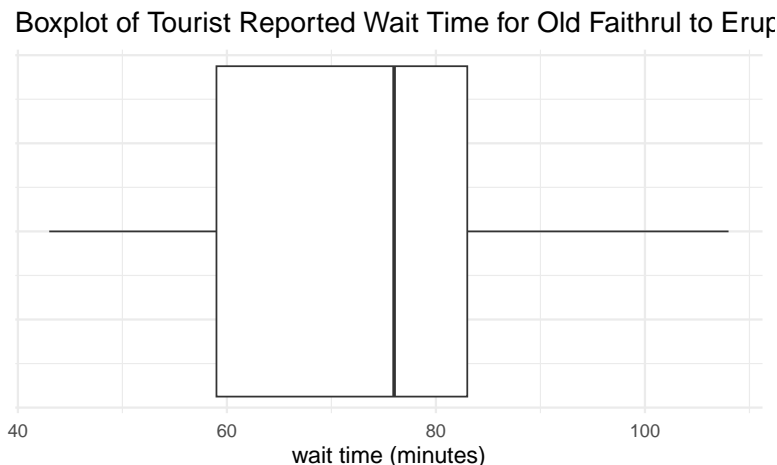
1. Write the parameter of interest in context of the study.

2. In the last module review, we saw very strong evidence that the true mean wait time reported by tourists for Old Faithful to erupt differs from 30 minutes. Do you expect the 99% confidence inteval to contain the null value of zero? Explain your answer.

We will start with simulation methods to create the 99% confidence interval.

```
set.seed(216)
one_mean_CI(data = geyser$waiting,   #Object and variable
            summary_measure = "mean",
            confidence_level = 0.99, #Level of context as a decimal
            number_repetitions = 10000)  #Number of simulated samples for null distribution
```

0.5 percentile   99.5 percentile

69   70   71   72   73   74   75

Bootstrap Mean
99% CI: (70.241, 74.385)
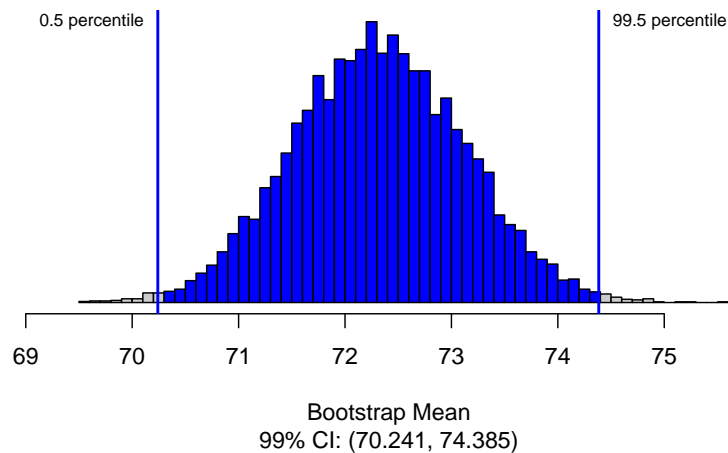
3. How many simulations are at and below the value of 70.241?

4. Report the 99% confidence interval.

Now let's focus on theory-based methods. **In the last module review, we verified the normality conditions were met.**

Conditions for the sampling distribution of $\bar{x}$ 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!)

- **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

  - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $30 \leq n < 100$: If the sample size n is betwe 30 and 100 and there are no particularly extreme outliers in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

To calculate a theory-based confidence interval for the a single mean, use the following formula:

$$\bar{x} \pm t^* \times SE(\bar{x}).$$

We will need to find the t^* multiplier using the function `qt()`.

- Enter the appropriate percentile (0.995) in the R code to find the multiplier for a 99% confidence interval.
- Enter the df $n - 1 = 299 - 1 = 298$

```
qt(0.995, df = 298, lower.tail=TRUE)
```

```
#> [1] 2.592428
```

5. Mark on the t-distribution found below the values of $\pm t^*$. Draw a line at each multiplier and write the percentiles used to find each.
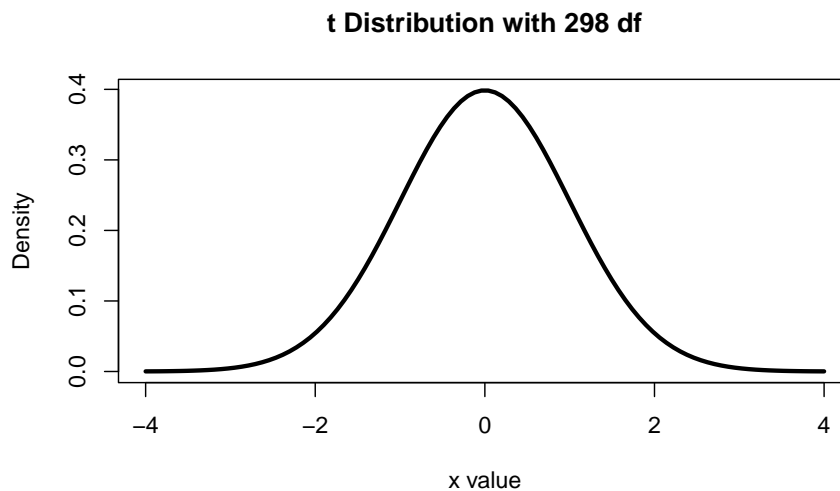

Figure 10.1: t-distribution with 602 degrees of freedom

6. Calculate the 99% confidence interval using theory-based methods.

7. Interpret the confidence interval in context of the study.

Types of Errors:

8. What type of error may have occurred for this study?

9. Interpret this error in context of the study.

10.4 Module 8 - 9 Review - Inference for Two Categorical Variables

```
allergy <- read.csv("https://math.montana.edu/courses/s216/data/PeanutAllergy.csv")
allergy %>% group_by(Treatment) %>% count(Allergy)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Treatment [2]
#>   Treatment Allergy     n
#>   <chr>     <chr>   <int>
#> 1 Avoiders  No        220
#> 2 Avoiders  Yes        35
#> 3 Peanuts   No        240
#> 4 Peanuts   Yes         5
```

In the last 10 years, the proportion of children who are allergic to peanuts has doubled in Western countries. However, the allergy is not very common in some other countries where peanut protein is an important part of peoples' diets. The LEAP randomized trial, reported by Du Toit, et.al in the New England Journal of Medicine in February 2015 identified over 500 children ages 4 to 10 months who showed some sensitivity to peanut protein. They randomly assigned them to two groups:

• Peanut avoiders: parents were told to not give their kids any food which contained peanuts

• Peanut eaters: parents were given a snack containing peanut protein and told to feed it to their child several times per week (target dose was at least 6g of peanut protein per week).

At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no). Is there evidence that exposure to peanuts reduces the likelihood of developing peanut allergies?

	Peanut Avoiders	Peanut Eaters	Total
Allergy	35	5	40
No Allergy	220	240	460
Total	255	245	500

For this study we will use the order of subtraction avoiders – eaters.

1. Fill in the blanks with one answer from each set of parentheses:

The variable whether or not a child is given peanut protein is the _____ (explanatory/response) variable and it is _____ (categorical/quantitative).

The variable whether or not a child developed a peanut allergy is the _____ (explanatory/response) variable and it is _____ (categorical/quantitative).

2. Write the parameter of interest for this study.

3. Write the null hypothesis in notation.

4. Write the alternative hypothesis in words.

5. Calculate the conditional proportion of children that developed a peanut allergy among those that avoided peanuts. Use proper notation.

6. Calculate the conditional proportion of children that developed a peanut allergy among those that ate peanuts. Use proper notation.

7. Calculate the difference in proportion of children that developed a peanut allergy for those that avoided peanuts and those who ate peanuts. Use proper notation.

8. First, let's think about how one simulation would be created on the null distribution using cards.

   How many cards would you need?

   What would be written on each card?

9. Next, we would mix the cards together and shuffle into two piles. How many cards would be in each pile? What would each pile represent?

10. Once we have one simulated sample, what would we calculate and plot on the null distribution? *Hint*: What statistic are we calculating from the data?

```
two_proportion_test(formula = Allergy ~ Treatment, #response~explanatory
                    data=allergy, #name of dataset
                    first_in_subtraction = "Avoiders", #order of subtraction: avoiders - peanuts
                    number_repetitions = 10000, #always use a minimum of 10000 repetitions
                    response_value_numerator = "Yes", #define a success as having an allergy
                    as_extreme_as = 0.117, #type your calculated observed statistic (difference in sample
                    direction="greater") #type your selected direction to match the alternative hypothesi
```

Mean = 0
SD = 0.024

Simulated Difference in Proportions
Count = 0/10000 = 0
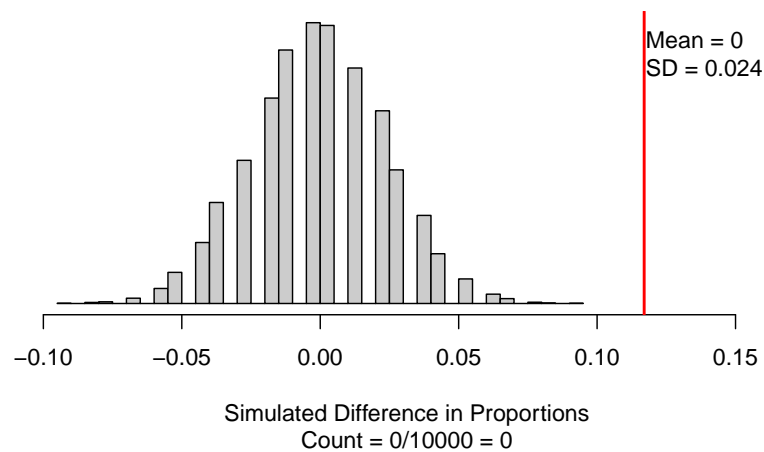
11. Interpret the p-value in context of the problem:

12. Write a conclusion to the test in context of the study.

We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `allergy`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

```r
two_proportion_bootstrap_CI(formula = Allergy~Treatment,
        data=allergy, # Name of data set
        first_in_subtraction = "Avoiders", # Order of subtraction: enter the name of Group 1
        response_value_numerator = "Yes", # Define which outcome is a success
        number_repetitions = 10000, # Always use a minimum of 10000 repetitions
        confidence_level = 0.90) # Enter the level of confidence as a decimal
```


Bootstrap Difference in Proportions
Mean: 0.117 SD: 0.023 90% CI: (0.078, 0.156)
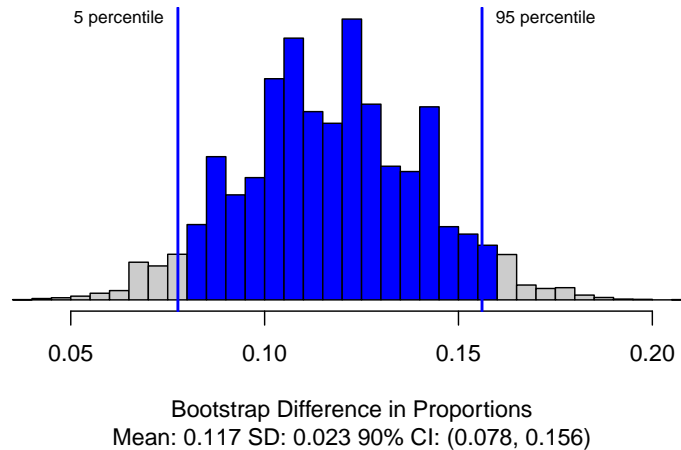
13. Interpret the 90% confidence interval in context of the problem.

Theory-based Methods

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution:

- **Independence**: The data are independent within and between the two groups. (*Remember*: This also must be true to use simulation methods!)

- **Success-failure condition**: This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.

14. Are the conditions met to use theory-based methods?

To calculate the standardized statistic we use:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \text{null value}}{SE_0(\hat{p}_1 - \hat{p}_2)},$$

where the null standard error is calculated using the pooled proportion of successes:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

15. Calculate the null standard error of the difference in proportion.

16. Calculate the standardized statistic.

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

```
pnorm(4.814, lower.tail = FALSE)
#> [1] 7.39694e-07
```

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \quad \text{where}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

18. Calculate the standard error of the difference in proportions to calculate the confidence interval.

```
qnorm(0.90, lower.tail = TRUE)
#> [1] 1.281552
```

19. Calculate the 90% confidence interval.

20. What is the scope of inference for this study?

Exploratory Data Analysis and Inference for a Quantitative Response with Paired Samples

11.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of paired data with a quantitative response.

11.1.1 Key topics

Module 11 will cover exploratory data analysis and both simulation-based and theory-based methods of inference for a quantitative response variable with paired samples. The **summary measure** for paired data is a **mean difference**.

- Notation for a sample mean difference: $\bar{x}_d$

- Notation for a population mean difference: μ_d

- Paired differences are treated as a single mean. Review the summary of Module 6 for interpretations of other summary measures from quantitative data and for the type of plots used. Additionally, we can create a plot of paired data in R using the `paired_observed_plot` function in the `catstats` function:

```
paired_observed_plot(object)
#Note you can use this plot if you ONLY have two columns of paired data in the data set
```

- R code to find the summary statistics for a paired differences:

```
object %>% # Data set piped into...
    summarise(favstats(differences))
```

11.1.2 Vocabulary

- **Hypotheses in notation for a paired mean difference**: In the hypotheses below, the **null value** is equal to zero.

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} 0$$

Simulation-based inference for a paired mean difference

- **Conditions necessary to use simulation-based methods for inference for paired data with a quantitative response**:

  - **Independence**: there must be independence of the sample differences; the pairs must be independent of each other. (Note that since this is paired data, measurements within a single pair will be dependent.)

- **Simulation-based methods to create the null distribution**: R code to use simulation-based methods for paired data with a quantitative response to find the p-value, `paired_test` (from the `catstats` package), is shown below.

```
paired_test(data = object$differences,    # Vector of differences
                                # or data set with column for each group
    shift = xx,    # Shift needed for bootstrap hypothesis test
    as_extreme_as = xx,   # Observed statistic
    direction = "xx",   # Direction of alternative
    number_repetitions = 10000,   # Number of simulated samples for null distribution
    which_first = 1)   # Not needed when using calculated differences
```

- **Simulation-based methods to create the bootstrap distribution**: R code to find the simulation-based confidence interval using the `paired_bootstrap_CI` function from the `catstats` package is shown below.

```
paired_bootstrap_CI(data = object$differences, # Enter vector of differences
                number_repetitions = 10000, # Number of bootstrap samples for CI
                confidence_level = xx,   # Confidence level in decimal form
                which_first = 1)   # Not needed when entering vector of differences
```

  – The interpretation of the confidence interval is very similar for that of a single mean. Just make sure to include the order of subtraction for the differences.

Theory-based inference for a paired mean difference

- **Conditions for the sampling distribution of $\bar{x}_d$ 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!)

  – **Normality Condition**: either the sample differences come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

    * $n < 30$: The distribution of the sample differences must be approximately normal with no outliers.

    * $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample differences must have no extreme outliers or skewness.

    * $n \geq 100$: Can assume the sampling distribution of $\bar{x}_d$ is nearly normal, even if the underlying distribution of individual observations is not.

- **Standard error of the sample mean difference**:

$$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$$

- **Standardized sample mean difference**:

$$T = \frac{\bar{x}_d - 0}{SE(\bar{x}_d)}$$

  – Use the `pt` function in R to find a theory-based p-value for a hypothesis test involving a mean difference by finding the area under a t-distribution with $n-1$ degrees of freedom where T is as or more extreme as the value observed (in the direction of H_A).

- **Margin of error**: half the width of the confidence interval. For a mean difference, the margin of error is:

$$ME = t^* \times SE(\bar{x}_d)$$

where t^* is the **multiplier**, corresponding to the desired confidence level found from a t-distribution with $n - 1$ degrees of freedom.

– Use the `qt` function in R to find the t^* multiplier with $n - 1$ degrees of freedom.

– To find the endpoints of a confidence interval, add and subtract the margin of error to the sample statistic. The confidence interval for a population mean difference is:

$$\bar{x}_d \pm ME$$

11.2 Video Notes: Inference for Paired Data

Read Chapters 17 and 18 in the course textbook. Use the following videos to complete the video notes for Module 9.

11.2.1 Course Videos

- PairedData

- 18.1and18.2

- 18.3

Single categorical, single quantitative variables Video Paired_Data

- In this module, we will study inference for a _____ explanatory variable and a _____ response variable where the two groups are _____.

Paired vs. Independent Samples

Two groups are paired if an observational unit in one group is connected to an observational unit in another group

Data are paired if the samples are _____

Examples:

- Change in test score from pre and post test

- Weight of college students before and after 1st year

- Change in blood pressure

Independent Samples

Sample 1	Sample 2
x_{1a}	x_{2a}
x_{1b}	x_{2b}
x_{1c}	x_{2c}
.	.
.	.
.	.
x_{1g}	x_{2g}
$\bar{x}_1$	$\bar{x}_2$
Difference in Means	$\bar{x}_1$-$\bar{x}_2$

Paired Data

Sample 1	Sample 2	Difference
x_{1a}	x_{2a}	x_{1a}-x_{2a}
x_{1b}	x_{2b}	x_{1b}-x_{2b}
x_{1c}	x_{2c}	x_{1c}-x_{2c}
.	.	.
.	.	.
.	.	.
x_{1g}	x_{2g}	x_{1g}-x_{2g}
Mean of the Differences		$\bar{x}_d$

Figure 11.1: Illustration of Independent vs. Paired Samples

Example 1: Three hundred registered voters were selected at random to participate in a study on attitudes about how well the president is performing. They were each asked to answer a short multiple-choice questionnaire and then they watched a 20-minute video that presented information about the job description of the president. After watching the video, the same 300 selected voters were asked to answer a follow-up multiple-choice questionnaire.

- Is this an example of a paired samples or independent samples study?

Example 2: Thirty dogs were selected at random from those residing at the humane society last month. The 30 dogs were split at random into two groups. The first group of 15 dogs was trained to perform a certain task using a reward method. The second group of 15 dogs was trained to perform the same task using a reward-punishment method.

- Is this an example of a paired samples or independent samples study?

Example 3: Fifty skiers volunteered to study how different waxes impacted their downhill race times. The participants were split into groups of two based on similar race times from the previous race. One of the two then had their skis treated with Wax A while the other was treated with Wax B. The downhill ski race times were then measured for each of the 25 volunteers who used Wax A as well as for each of the 25 volunteers who used Wax B.

- Is this an example of a paired samples or independent samples study?

Example: Is there a difference in heights between husbands and wives? The heights were measured on the husband and wife in a random sample of 199 married couples from Great Britain ("Great Britain Married Couples: Great Britain Office of Population Census and Surveys," n.d.).

For a paired experiment, we look at the difference between responses for each unit (pair), rather than just the average difference between treatment groups

```
hw <-read.csv("data/husbands_wives_ht.csv")
paired_observed_plot(hw)
```
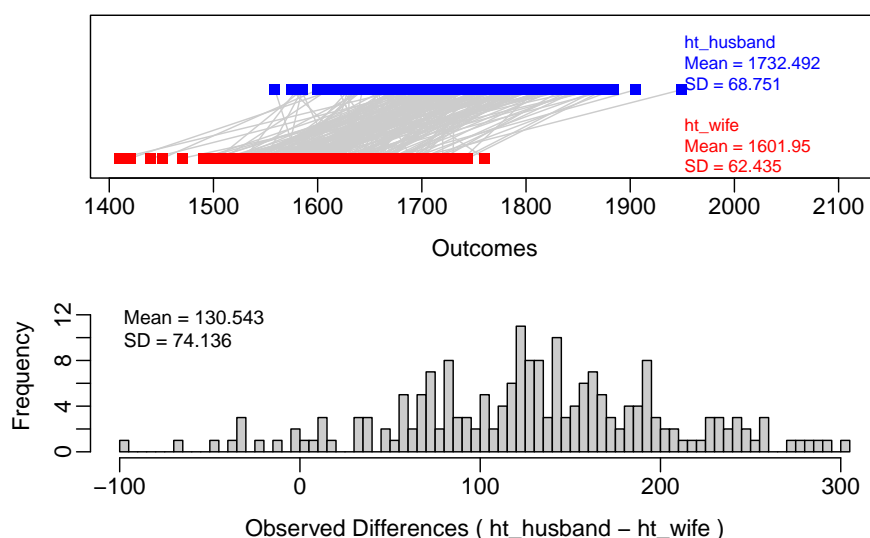


```
hw_diff %>%
    summarise(fav_stats(ht_diff))
```

```
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

- The summary measure for paired data is the _____.

240

- Mean difference: the average _____ in the _____ variable outcomes for observational units between _____ variable groups

Notation for the Paired differences

- Population mean of the differences:
- Population standard deviation of the differences:
- Sample mean of the differences:
- Sample standard deviation of the differences:

Conditions for inference for paired data:

- Independence:

Is the independence condition met for the height study?

Hypothesis testing

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

- Treat the differences like a single mean
- Always of form: "parameter" = null value

H_0 :

H_A :

- Research question determines the direction of the alternative hypothesis.

Write the null and alternative for the height study:

In notation:

H_0 :

H_A :

Simulation-based method

- Simulate many samples assuming $H_0 : \mu_d = 0$
    - Shift the data by the difference between μ_0 and $\bar{x}_d$
    - Sample with replacement n times from the shifted data
    - Plot the simulated shifted sample mean from each simulation
    - Repeat 10000 times (simulations) to create the null distribution

241

– Find the proportion of simulations at least as extreme as $\bar{x}_d$

Reminder of summary statistics:

```
hw_diff %>%
    summarise(fav_stats(ht_diff))
```

```
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

Find the difference:

$\mu_0 - \bar{x}_d =$

Simulated null distribution:

```
set.seed(216)
paired_test(data = hw_diff$ht_diff,   # Vector of differences
                                      # or data set with column for each group
            shift = -130.543,   # Shift needed for bootstrap hypothesis test
            as_extreme_as = 130.543,  # Observed statistic
            direction = "two-sided",  # Direction of alternative
            number_repetitions = 10000,  # Number of simulated samples for null distribution
            which_first = 1)   # Not needed when using calculated differences
```
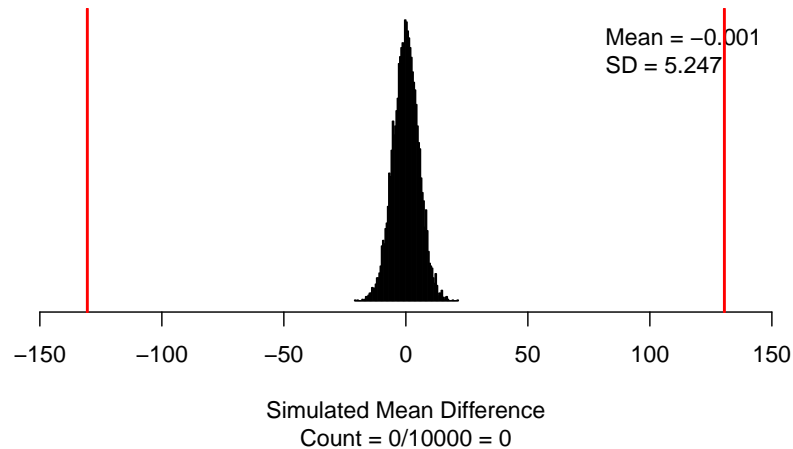

Interpret the p-value:

- 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

Confidence interval

Simulation-based method

- Label cards with the values (differences) from the data set
- Sample with replacement (bootstrap) from the original sample n times
- Plot the simulated sample mean on the bootstrap distribution
- Repeat at least 10000 times (simulations)
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.
    - i.e., 95% CI = (2.5th percentile, 97.5th percentile)

Simulated bootstrap distribution:

```
set.seed(216)
paired_bootstrap_CI(data = hw_diff$ht_diff, # Enter vector of differences
            number_repetitions = 10000, # Number of bootstrap samples for CI
            confidence_level = 0.95,  # Confidence level in decimal form
            which_first = 1)  # Not needed when entering vector of differences
```


Bootstrap Mean Difference
95% CI: (120.246, 140.86)
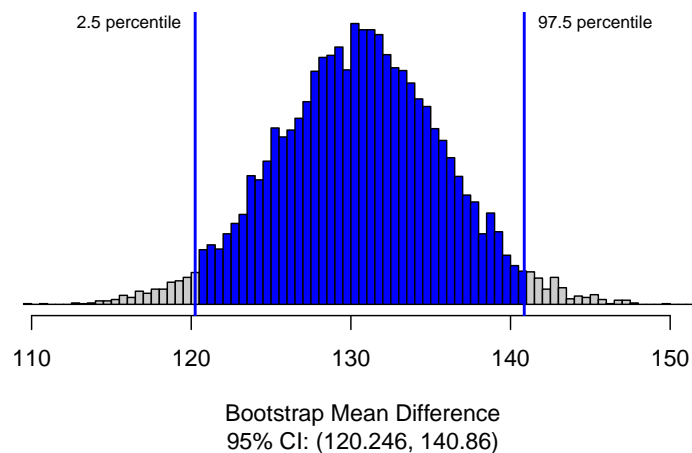
Interpret the 99% confidence interval:

- How confident you are (e.g., 90%, 95%, 98%, 99%)

- Parameter of interest

- Calculated interval

- Order of subtraction when comparing two groups

Theory-based method - Video 18.3

t-distribution

In the theoretical approach, we use the CLT to tell us that the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0 and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N(\mu_0, \frac{\sigma_d}{\sqrt{n}})$$

- Estimate the population standard deviation, σ_d, with the _____ standard deviation, _____.

- For a single quantitative variable we use the _____ - distribution with _____ degrees of freedom to approximate the sampling 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!)

- **Normality Condition**: either the sample differences come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

    $n < 30$:

    $30 \leq n < 100$:

    $n \geq 100$:

Theory-based Hypothesis Test:

- Calculate the standardized statistic

- Find the area under the t-distribution with $n - 1$ df at least as extreme as the standardized statistic

Equation for the standard error for the sample mean difference:

Equation for the standardized sample mean difference:

Reminder of summary statistics for height data:

```
hw_diff %>%
    summarise(fav_stats(ht_diff))
```

```
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

Calculate the standardized sample mean difference in height:

- 1st calculate the standard error of the sample mean difference

- Then calculate the T score

What theoretical distribution should we use to find the p-value using the value of the standardized statistic?

t–distribution with 198 df
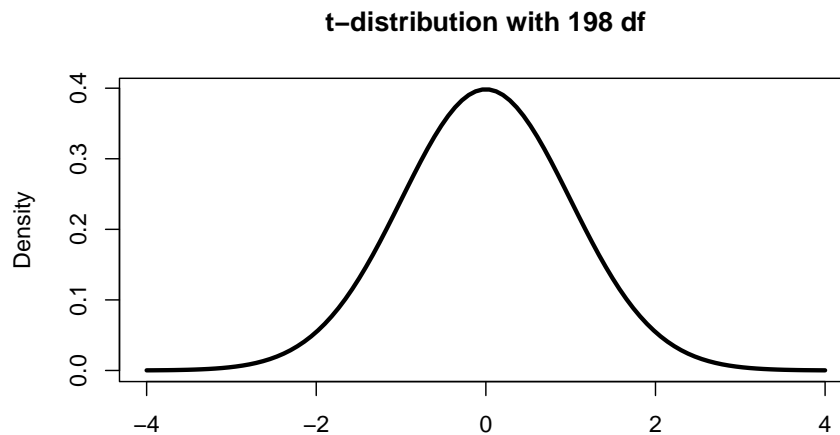
To find the p-value:

```
pt(24.84, df = 198, lower.tail=FALSE)*2
```

```
#> [1] 9.477617e-63
```

Theory-based Confidence Interval:

$$\text{statistic} \pm \text{margin of error}$$

The t^* multiplier is the value at the given percentile of the t-distribution with $n - 1$ degrees of freedom.

For the height data, we will use a t-distribution with _____ df.

t–distribution with 198 df
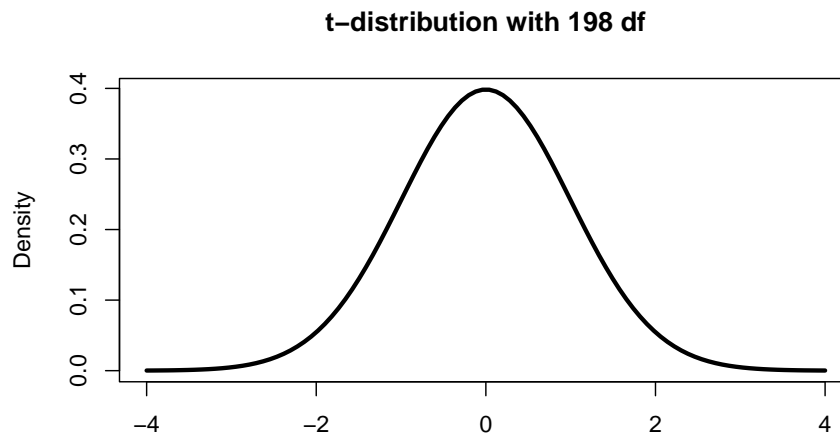
To find the t^* multiplier for a 99% confidence interval:

```r
qt(0.975, df=198, lower.tail = TRUE)
```

```
#> [1] 1.972017
```

Calculate the margin of error:

Calculate the theory-based confidence interval.

11.2.2 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 theoretical distribution is used to approximate paired quantitative data?

2. What is the difference between a paired and independent study design?

11.3 Activity 21: Paired vs. Independent Samples

11.3.1 Learning outcomes

- Determine if a data set is paired or two independent samples

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.

- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

11.3.2 Terminology review

In today's activity, we will review summary measures and plots for paired data. Some terms covered in this activity are:

- Mean difference

- Median difference

- Standard deviation

- Quartiles

To review these concepts, see Chapter 5 and 18 in the textbook.

11.3.3 Paired vs. Independent Samples

For each of the following scenarios, determine whether the samples are paired or independent.

1. Researchers interested in studying the effect of a medical treatment on insulin rate measured insulin rates of 30 patients before and after the medical treatment.

2. A university is planning to bring emotional support animals to campus during finals week and wants to determine which type of animals are more effective at calming students. Anxiety levels will be measured before and after each student interacts with either a dog or a cat. The university will then compare change in anxiety levels between the 'dog' people and the 'cat' people.

3. An industry leader is investigating a possible wage gap between male and non-male employees. Twenty companies within the industry are randomly selected and the average salary for all males and non-males in mid-management positions is recorded for each company.

4. Researchers conducted a study to evaluate the effectiveness of a brief yoga intervention on working memory improvement. A sample of 43 undergraduate students at Texas State University was recruited for the study. Participants completed six yoga sessions. Working memory was assessed before and after the intervention using the Digit Span Forward task, which involves recalling a sequence of numbers in the same order they were presented. The maximum number of digits correctly recalled in the proper order served as the measure of working memory. Is there evidence that working memory is higher, on average, after completing six yoga sessions than before six yoga sessions?

11.3.4 Tattoo Effect on Sweat Rate

The popularity of tattoos has increased tremendously in the last 10 years particularly among athletes and military personnel. A study reported in *Medicine & Science in Sports & Exercise* (LUETKEMEIER 2017) looked at whether skin tattoos altered a person's sweat rate. The study participants were 10 healthy men with a tattoo. According to the article, sweat was stimulated by iontophoresis using agar gel disks impregnated with 0.5% pilocarpine nitrate. The sweat rate was determined by weighing the disk before and after sweat collection. Sweat rate was measured on both the tattooed skin and untattooed skin from the same participant. We will use these data to assess the difference in sweat rate between the tattooed and untattooed skin (tat - notat).

11.3.5 Exploring Paired Data

5. Explain why this is a paired study design and not independent groups.

- Download the R script file and the data file for this activity from D2L
- Upload the R script file and data set to the RStudio server
- Open the R script file and enter the name of the dataset for datasetname.csv
- Run lines 1–9 to create the paired plot

```
tats <- read.csv("data/tattoos.csv")
paired_observed_plot(tats)
```

6. What is the value of the mean difference? Give the appropriate notation.

7. What is the value of the standard deviation of the differences. Give the appropriate notation.

8. You may notice that the difference in the means $(0.178 - 0.345 = -0.167)$ is equal to the mean difference you reported in Q6. That is not true for the standard deviations. Explain why?

9. Interpret the value of the standard deviation of the differences

10. What is the shape of the distribution of the differences?

To find the differences to continue to assess this data and create a boxplot of the differences...

- Enter Tattoo for measurement_1 and No_Tattoo for measurement_2 in line 15
- Highlight and run lines 13 - 21

```
tat_diff <- tats %>%
    mutate(differences = Tattoo - No_Tattoo)
tat_diff %>%
    summarise(favstats(differences))
tat_diff %>%
    ggplot(aes(x = differences)) +
    geom_boxplot()+
    labs(title="Boxplot of the difference in Sweat Rate (mg·cm^2) per min for Adult Men
         with Tattoos comparing Tattooed and Untattooed Skin (tat-notat)")
```

11. What four characteristics do we use to describe the plot of a quantitative variable?

12. Identify the value of Q_3. Interpret this value in context of the study.

13. Interpret the mean from the five number summary in context of the study. Identify where this value is found in the paired plot.

14. What other measure of center could we use to describe the distribution of differences in sweat rates between tattooed and untattooed skin?

15. Based on the plots, which measure of center would be more appropriate to describe the distribution of the differences in sweat rates between tattooed and untattooed skin? Explain why.

16. If we wanted to test that there is evidence that the sweat rate is lower for the tattooed skin than for the untattooed skin, on average, what null value would be used for this study? What direction of the alternative would be used?

17. Using the Golden Ticket, write the null hypothesis for this study in notation.

18. The authors reported that the confidence interval for the mean difference was -0.17 ± 0.11 $\frac{mg*cm^2}{min}$. Does this interval provide evidence in support of the alternative hypothesis? Explain why.

1. The differences in a paired data set are treated like a single quantitative variable when performing a statistical analysis. Paired data (or paired samples) occur when pairs of measurements are collected. We are only interested in the population (and sample) of differences, and not in the original data.

2. Plots and interpretations of summary statistics are very similar to that for a single mean.

11.3.6 Additional notes

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

11.4 Activity 22: Snakes

11.4.1 Learning outcomes

- Given a research question involving paired differences, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a paired mean difference.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a paired mean difference.

- Use bootstrapping to find a confidence interval for a paired mean difference.

- Interpret a confidence interval for a paired mean difference.

- Use a confidence interval to determine the conclusion of a hypothesis test.

11.4.2 Terminology review

In today's activity, we will analyze paired quantitative data using simulation-based methods. Some terms covered in this activity are:

- Mean difference

- Paired data

- Independent groups

- Shifted bootstrap (null) distribution

To review these concepts, see Section 18 in the textbook.

11.4.3 Snake mazes

"Sidewinding" refers to a method of locomotion used by snakes to travel over slippery or loose terrain, such as sand. This manner of movement is a form of lateral undulation, in which an animal uses wave-like movement patterns to propel themselves forward. The desert horned viper is one of a handful of snake species that primarily uses sidewinding. As part of a recent study (Subach 2022), researchers exploring foraging behaviors of these vipers in the Sahara and western Negev deserts captured 27 unique vipers with the goal of examining how environment complexity affected movement. Each snake was placed in a circular "maze" of 1045 poles stuck into a sand dune (meant to simulate a dense-vegetation area). The snake was timed to see how long it took to get out of the maze. Researchers also measured how far the snake had "traveled" within the maze by examining the tracks left in the sand. Once the snake had completed the maze, it was allowed to travel in an open area for the same length of time as it spent in the maze; distance was again measured by the tracks in the sand. Is there evidence the distance traveled (in meters) by snakes is less in areas of dense vegetation than in open spaces, on average? Use circular maze - open area as the order of subtraction.

- Download the R script file and csv file from D2L and upload both to the RStudio server

- Open the R script file and enter the name of the data set for datasetname.csv

- Highlight and run lines 1 - 9

```
snakes <- read_csv("datasetname.csv")
paired_observed_plot(snakes)
```

To find the difference in distance traveled in the circular maze vs in open spaces for each snake we will create the variable differences.

- Enter `DistanceMaze` for measurement_1 and `DistanceOpen` for measurement_2 in line 16

- Highlight and run 14–22

```
snakes_diff <- snakes %>%
  mutate(differences = measurement_1 - measurement_2)
snakes_diff %>%
    summarise(favstats(differences))

snakes_diff %>%
    ggplot(aes(x = differences)) +
    geom_boxplot()+
    labs(title="Boxplot of the Difference in Distance for Snakes to
        Complete the Open Area vs the Maze
        (Maze - Open)")
```

1. Explain why simulation methods should be used to analyze these data.

2. Explain why this is a paired study design.

Ask a research question

3. Write the null hypothesis in words.

4. Write the alternative hypothesis in notation.

Use statistical inferential methods to draw inferences from the data

5. Report the sample mean difference with appropriate notation.

Hypothesis test To simulate the null distribution of paired sample mean differences we will use a bootstrapping method. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping, we will need to *shift* each data point by the difference $\mu_0 - \bar{x}_d$. This will ensure that the mean of the shifted data is μ_0 (rather than the mean of the original data, $\bar{x}_d$), and that the simulated null distribution will be centered at the null value. This is the same process we used to create the simulation of the null distribution for a single quantitative variable.

6. Calculate the difference $\mu_0 - \bar{x}_d$. Will we need to shift the data up or down?

We will use the `paired_test()` function in R (in the `catstats` package) to simulate the shifted bootstrap (null) distribution of sample mean differences and compute a p-value.

- Use the provided R script file and enter the calculated value from question 6 for **xx** to simulate the null distribution and enter the summary statistic from question 5 for **yy** to find the p-value.

- Highlight and run lines 27–33.

```
paired_test(data = snakes_diff$differences,    # Vector of differences
                                               # or data set with column for each group
            shift = xx,    # Shift needed for bootstrap hypothesis test
            as_extreme_as = yy,  # Observed statistic
            direction = "less",  # Direction of alternative
            number_repetitions = 10000,  # Number of simulated samples for null distribution
            which_first = 1)  # Not needed when using calculated differences
```

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

Confidence interval   We will use the `paired_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample mean differences and calculate a 99% confidence interval.

- Enter missing values for xx

- Highlight and run lines 38 - 41

```
paired_bootstrap_CI(data = snakes_diff$differences, # Enter vector of differences
                    number_repetitions = 10000, # Number of bootstrap samples for CI
                    confidence_level =xx,  # Confidence level in decimal form
                    which_first = 1)  # Not needed when entering vector of differences
```

8. Report the 99% confidence interval.

9. Interpret the 99% confidence interval in the context of the problem.

10. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

11. Write a conclusion to the test.

11.4.4 Take-home messages

1. The differences in a paired data set are treated like a single quantitative variable when performing a statistical analysis. Paired data (or paired samples) occur when pairs of measurements are collected. We are only interested in the population (and sample) of differences, and not in the original data.

2. When using bootstrapping to create a null distribution centered at the null value for both paired data and a single quantitative variable, we first need to shift the data by the difference $\mu_0 - \bar{x}_d$, and then sample with replacement from the shifted data.

3. When analyzing paired data, the summary statistic is the 'mean difference' NOT the 'difference in means'[1]. This terminology will be *very* important in interpretations.

4. To create one simulated sample on the null distribution for a sample mean or mean difference, shift the original data by adding $(\mu_0 - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Calculate and plot the sample mean or the sample mean difference.

5. To create one simulated sample on the bootstrap distribution for a sample mean or mean difference, label n cards with the original response values. Randomly draw with replacement n times. Calculate and plot the resampled mean or the resampled mean difference.

11.4.5 Additional notes

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

---

[1]Technically, if we calculate the differences and then take the mean (mean difference), and we calculate the two means and then take the difference (difference in means), the value will be the same. However, the *sampling variability* of the two statistics will differ, as we will see in Module 12.

11.5 Activity 23: Color Interference

11.5.1 Learning outcomes

- Given a research question involving paired differences, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a theory-based hypothesis test for a paired mean difference.

- Interpret and evaluate a p-value for a theory-based hypothesis test for a paired mean difference.

- Use theory-based methods to find a confidence interval for a paired mean difference.

- Interpret a confidence interval for a paired mean difference.

- Use a confidence interval to determine the conclusion of a hypothesis test.

11.5.2 Terminology review

In today's activity, we will analyze paired quantitative data using theory-based methods. Some terms covered in this activity are:

- Paired data

- Mean difference

- Independent observational units

- Normality

- t-distribution

- Degrees of freedom

- T-score

To review these concepts, see Chapter 18 in the textbook.

11.5.3 Color Interference

The abstract of the article "Studies of interference in serial verbal reactions" in the *Journal of Experimental Psychology* (Stroop 1935) reads:

> In this study pairs of conflicting stimuli, both being inherent aspects of the same symbols, were presented simultaneously (a name of one color printed in the ink of another color—a word stimulus and a color stimulus). The difference in time for reading the words printed in colors and the same words printed in black is the measure of interference of color stimuli upon reading words. … The interference of conflicting color stimuli upon the time for reading 100 words (each word naming a color unlike the ink-color of its print) caused an increase of 2.3 seconds or 5.6% over the normal time for reading the same words printed in black.

The article reports on the results of a study in which seventy college undergraduates were given forms with 100 names of colors written in black ink, and the same 100 names of colors written in another color (i.e., the word purple written in green ink). The total time (in seconds) for reading the 100 words printed in black, and the total time (in seconds) for reading the 100 words printed in different colors were recorded for each subject. The order in which the forms (black or color) were given was randomized to the subjects. Does printing the name of colors in a different color increase the time it takes to read the words? Use color − black as the order of subtraction.

Identify the scenario

1. Should these observations be considered paired or independent? Explain your answer.

2. Based on your answer to question 1, is the appropriate summary measure to be used to analyze these data the difference in mean times or the mean difference in times?

Ask a research question

3. Write out the null hypothesis in words, in the context of this study.

4. Write out the alternative hypothesis in proper notation for this study.

In general, the sampling distribution for a sample mean, $\bar{x}$, based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a Normal distribution when certain conditions are met. Note, that since we are treating paired data as a single mean, the conditions are the same as for a single mean.

Conditions for the sampling distribution of $\bar{x}$ 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!)

- **Normality Condition**: either the sample differences come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

  - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Summarize and visualize the data

Since the original data from the study are not available, we simulated data to match the means and standard deviations reported in the article. We will use these simulated data in the analysis below.

The following code plots each subject's time to read the colored words (above) and time to read the black words (below) connected by a grey line, a histogram of the differences in time to read words between the two conditions, and a boxplot displaying the pairwise differences in time (color − black).

- Download the R script file for this activity and upload to the R studio server.

- Follow the instructions given in the R file to create the paired plot and boxplot of the differences.

```
color <- read.csv("https://math.montana.edu/courses/s216/data/interference.csv")
paired_observed_plot(color)

color_diff <- color %>%
  mutate(differences = DiffCol-Black)
color_diff %>%
  ggplot(aes(x = differences))+
  geom_boxplot()+
  labs(title="Boxplot of the Difference in Time to Read Words
       Between Color and Black for College Undergraduates",
       x = "Differences in time to read words (Color - Black)")
```
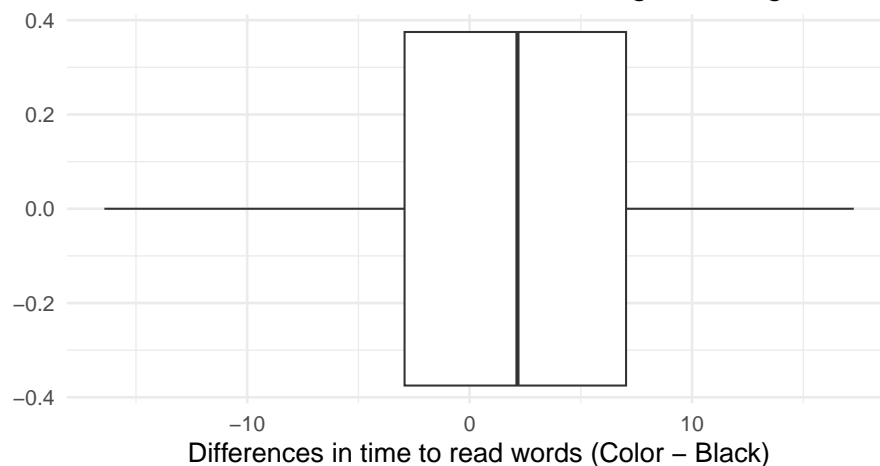

The following code gives the summary statistics for the pairwise differences.

- Enter the variable `differences` for variable

- Highlight and run lines 22 - 23

```
color_diff %>%
  summarise(favstats(differences))
```

Check theoretical conditions

5. How do you know the independence condition is met for these data?

6. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the paired differences we will use the following formula:

$$T = \frac{\bar{x}_d - \mu_0}{SE(\bar{x}_d)},$$

where the standard error of the sample mean difference is:

$$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}.$$

7. Calculate the standard error of the sample mean difference.

8. How many standard errors is the observed mean difference from the null mean difference?

To find the p-value

- Enter the value for the standardized statistic for xx in the pt function.

- For a single sample or paired data, degrees of freedom are found by subtracting 1 from the sample size. You should therefore use $df = n - 1 = 70 - 1 = 69$ and `lower.tail = FALSE` to find the p-value.

- Enter the df for yy in the pt function.

- Highlight and run line 28

```
pt(xx, df=yy, lower.tail=FALSE)
```

9. What does this p-value mean, in the context of the study? Hint: it is the probability of what...assuming what?

Next we will calculate a theory-based confidence interval. To calculate a theory-based confidence interval for the paired mean difference, use the following formula:

$$\bar{x}_d \pm t^* \times SE(\bar{x}_d).$$

We will need to find the t^* multiplier using the function `qt()`.

- Enter the appropriate percentile in the R code to find the multiplier for a 90% confidence interval.

- Enter the df for yy.

- Highlight and run line 34

```
qt(percentile, df = yy, lower.tail=TRUE)
```

10. Mark on the t-distribution found below the values of $\pm t^*$. Draw a line at each multiplier and write the percentiles used to find each.

t Distribution with 69 df
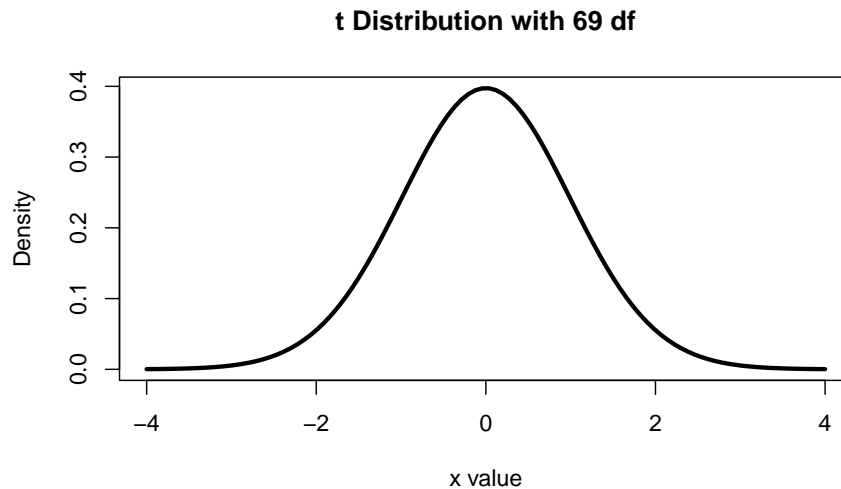
Figure 11.2: t-distribution with 69 degrees of freedom

11. Calculate the margin of error for the true paired mean difference using theory-based methods.

12. Calculate the confidence interval for the true paired mean difference using theory-based methods.

13. Interpret the confidence interval in context of the study.

14. Write a conclusion to the test in context of the study.

15. The abstract states, that the conflicting color stimuli "caused an increase of 2.3 seconds or 5.6% over the normal time for reading the same words printed in black." Is this statement valid? Explain.

11.5.4 Take-home messages

1. In order to use theory-based methods for dependent groups (paired data), the independent observational units and normality conditions must be met.

2. A T-score is compared to a t-distribution with $n-1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.

3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

11.5.5 Additional notes

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

11.6   Module 11 Lab: Swearing

11.6.1   Learning outcomes

- Identify whether a study is a paired design or independent groups

- Given a research question involving paired data, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a mean difference.

- Interpret and evaluate a p-value for a hypothesis test for a mean difference.

- Use bootstrapping methods to find a confidence interval for a mean difference.

- Interpret a confidence interval for a mean difference.

11.6.2   Swearing

Profanity (language considered obscene or taboo) and society's attitude about its acceptableness is a highly debated topic, but does swearing serve a physiological purpose or function? Previous research has shown that swearing produces increased heart rates and higher levels of skin conductivity. It is theorized that since swearing provokes intense emotional responses, it acts as a distracter, allowing a person to withstand higher levels of pain. To explore the relationship between swearing and increased pain tolerance, researchers from Keele University (Staffordshire, UK) recruited 83 native English-speaking participants (Stephens and Robertson 2020). Each volunteer performed two trials holding a hand in an ice-water bath, once while repeating the "f-word" every three seconds, and once while repeating a neutral word ("table"). The order of the word to repeat was randomly assigned. Researchers recorded the length of time, in seconds, from the moment the participant indicated they were in pain until they removed their hand from the ice water for each trial. They hope to find evidence that pain tolerance is greater (longer times) when a person swears compared to when they say a neutral word, on average. Use Swear – Neutral as the order of subtraction.

1. What is the explanatory variable for this study? What is the response?

2. What does μ_d represent in the context of this study?

3. Write out the null hypothesis in proper notation for this study.

4. What sign ($<$, $>$, or $\neq$) would you use in the alternative hypothesis for this study? Explain your choice.

- Upload and open the R script file for Module 11 lab.

- Upload and import the csv file, `pain_tolerance`.

- Enter the name of the data set for datasetname.csv in the R script file in line 8.

- Highlight and run lines 1–9 to load the data and create a paired plot of the data.

```
swearing <- read.csv("datasetname.csv")
paired_observed_plot(swearing)
```

5. Based on the plots, does there appear to be some evidence in favor of the alternative hypothesis? How do you know?

- Enter the outcome for group 1 (`Swear`) for `measurement_1` and the outcome for group 2 (`Neutral`) for `measurement_2` in line 16.

- Highlight and run lines 14–25 to get the summary statistics and boxplot of the differences.

```
swearing_diff <- swearing %>%
  mutate(differences = measurement_1 - measurement_2)
swearing_diff %>%
    summarise(favstats(differences))

swearing_diff %>%
    ggplot(aes(x = differences)) +
    geom_boxplot()+
    labs(title="Boxplot of the Difference in Time Participants Held Their Hand
         in Ice Water while Swearing or while Saying a Neutral Word (Swearing - Neutral)")
```

6. What is the value of $\bar{x}_d$? What is the sample size?

7. **How far, on average, is each difference in time the participant holds their hand in ice water from the mean of the differences in time? What is the appropriate notation for this value?**

Use statistical inferential methods to draw inferences from the data

8. Using the provided graphs and summary statistics, determine if both theory-based methods and simulation methods could be used to analyze the data. Explain your reasoning.

Hypothesis test

Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that swearing does not affect pain tolerance, or that the length of time a subject kept their hand in the water would be the same whether the patient was swearing or not.

We will use the `paired_test()` function in R (in the `catstats` package) to simulate the null distribution of sample mean differences and compute a p-value.

9. When using the `paired_test()` function, we need to enter the name of the data set, either the order of subtraction (if the data set has both measurements) or the name of the differences (if the data set contains them). We will also need to provide R with the observed mean difference, the direction of the alternative hypothesis, and the shift required in order to force the null hypothesis to be true. The name of the data set as shown above is `swearing_diff` and the column of differences is called `differences`. What values should be entered for each of the following to create 10000 simulated samples?

- shift:

- As extreme as:

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

- Number of repetitions:

10. Simulate a null distribution and compute the p-value. Using the R script file for this lab, enter your answers for question 9 in place of the `xx`'s to produce the null distribution with 10000 simulations. Highlight and run lines 30–36.

```
paired_test(data = swearing$differences,    # Vector of differences
                                # or data set with column for each group
       shift = xx,    # Shift needed for bootstrap hypothesis test
       as_extreme_as = xx,  # Observed statistic
       direction = "xx",  # Direction of alternative
       number_repetitions = xx,  # Number of simulated samples for null distribution
       which_first = 1)  # Not needed when using calculated differences
```

Sketch the null distribution created using the `paired_test` code.

Communicate the results and answer the research question

11. **Report the p-value. Based off of this p-value and a 1% significance level, what decision would you make about the null hypothesis? What potential error might you be making based on that decision?**

12. Do you expect the 98% confidence interval to contain the null value of zero? Explain.

Confidence interval

We will use the `paired_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample mean differences and calculate a confidence interval.

13. Using bootstrapping and the provided R script file, find a 98% confidence interval. Fill in the missing values/numbers in the `paired_bootstrap_CI()` function to create the 98% confidence interval. Highlight and run lines 41–44. **Upload a copy of the bootstrap distribution created to Gradescope for your group.**

```
paired_bootstrap_CI(data = swearing_diff$differences, # Enter vector of differences
                    number_repetitions = 10000, # Number of bootstrap samples for CI
                    confidence_level = xx,  # Confidence level in decimal form
                    which_first = 1)  # Not needed when entering vector of differences
```

Report the 98% confidence interval in interval notation.

14. Interpret the *confidence level* of the interval found in question 12.

15. Write a paragraph summarizing the results of the study. **Upload a copy of your group's paragraph to Gradescope.** 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)?

Exploratory Data Analysis and Inference for a Quantitative Response with Independent Samples

12.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a categorical explanatory variable and a quantitative response variable for independent samples.

12.1.1 Key topics

Module 12 will cover exploratory data analysis and both simulation-based and theory-based methods of inference for a quantitative response variable with independent samples. The **summary measure** for a quantitative response with independent samples is a **difference in means**.

- Notation for a difference in sample means: $\bar{x}_1 - \bar{x}_2$, where 1 represents the 1st group of the explanatory variable and 2 represents the 2nd group

- Notation for a difference in population means: $\mu_1 - \mu_2$

Types of plots for a quantitative response variable with independent samples:

- Side-by-side boxplots

- Stacked histograms

- Stacked dotplots

R code to find the summary statistics for a quantitative response variable with independent samples:

```
object %>%
  reframe(favstats(response ~ explanatory))
```

12.1.2 Vocabulary

Plotting two categorical variables

- **Side-by-side boxplots**: plots a boxplot of the five number summary for each categorical level. R code to create side-by-side boxplots:

```
object %>%  # Data set piped into...
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Don't forget to include a title",  # Title: should include the type of plot,
   # observational units, variables
   x = "x-axis label",    # x-axis label
   y = "y-axis label")  # y-axis label
```

Hypotheses

- **Hypotheses in notation for a difference in means**: In the hypotheses below, the **null value** is equal to zero.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 - \mu_2 \left\{ \begin{matrix} < \\ \neq \\ < \end{matrix} \right\} 0 \quad \text{or} \quad H_A : \mu_1 \left\{ \begin{matrix} < \\ \neq \\ < \end{matrix} \right\} \mu_2$$

Simulation-based inference for a difference in means

- **Conditions necessary to use simulation-based methods for inference for a quantitative response with independent groups**:

  - **Independence**: there must be independence of observational units within groups and between groups.

- **Simulation-based methods to create the null distribution**: R code for simulation-based methods to find the p-value using the `two_mean_test` function in the `catstats` package.

```
two_mean_test(response~explanatory, #Enter the names of the variables
          data = object,  # Enter the name of the dataset
          first_in_subtraction = "xx", # First outcome in order of subtraction
          number_repetitions = 10000,  # Number of simulations
          as_extreme_as = xx,  # Observed statistic
          direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

- **Simulation-based methods to create the bootstrap distribution**: R code to find the simulation-based confidence interval using the `two_mean_bootstrap_CI` function from the `catstats` package.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                  data = object,  # Enter the name of the data set
                  first_in_subtraction = "xx", # First value in order of subtraction
                  number_repetitions = 10000,  # Number of simulations
                  confidence_level = xx)
```

  - Review how to interpret a confidence interval for two groups from Module 8.

Theory-based inference for a difference in means

- **Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution**:

  - **Independence**: the sample's observations are independent, e.g., are from a simple random sample and there is independence between groups. (*Remember*: This also must be true to use simulation methods!)

  - **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. *When we have two samples, we need to check this condition for each group!* To check this condition, use the following rules of thumb (for both n_1 and n_2):

    * $n < 30$: The distribution of the sample must be approximately normal with no outliers.

    * $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.

    * $n \geq 100$: Can assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

- **Standard error of the sample difference in means**:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}$$

- **Standardized sample difference in means**:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$$

  – Use the `pt` function in R to find a theory-based p-value for a hypothesis test involving a difference in means by finding the area under a t-distribution with $\min(n_1 - 1, n_2 - 1)$ (the minimum sample size minus 1) degrees of freedom where T is as or more extreme as the value observed (in the direction of H_A).

- **Margin of error**: half the width of the confidence interval. For a difference in means, the margin of error is:

$$ME = t^* \times SE(\bar{x}_1 - \bar{x}_2)$$

where t^* is the **multiplier**, corresponding to the desired confidence level found from a t-distribution with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom.

  – Use the `qt` function in R to find the t^* multiplier with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom.

  – To find the endpoints of a confidence interval, add and subtract the margin of error to the sample statistic. The confidence interval for a population difference in means is:

$$\bar{x}_1 - \bar{x}_2 \pm ME$$

12.2   Video Notes: Inference for Independent Samples

Read Chapters 19 and 20 in the course textbook. Use the following videos to complete the video notes for Module 10.

12.2.1   Course Videos

- 19.1
- 19.2
- 19.3TheoryTests
- 19.4TheoryInterval

Single categorical, single quantitative variable with independent samples

- In this module, we will study inference for a _____ explanatory variable and a _____ response variable where the two groups are _____.

- Independent groups: When the measurements in one sample are not related to the measurements in the other sample.

- Two random samples taken separately from two populations and the same response variable is recorded. Compare the average number of sick days off from work for people who had a flu shot and people who didn't.

- Participants are randomly assigned to one of two treatment conditions, and the same response variable is recorded.

Rather than analyzing the differences as a single mean we will calculate summary statistics on each sample.

Example: Fifty-one (51) college students volunteered to look at impacts on memorization, specifically if putting letters into recognizable patterns (like FBI, CIA, EDA, CDC, etc.) would increase the number letters memorized. (Miller 1956) The college students were randomly assigned to either a recognizable or non-recognizable letter group. After a period of study time, the number of letters memorized was collected on each study. Is there evidence that putting letters into recognizable letter groups improve memory?

- The summary measure for two independent groups is the _____ in _____.

Notation for Independent Groups

- Population mean for group 1:

- Population mean for group 2:

- Sample mean for group 1:

- Sample mean for group 2:

- Sample difference in means:

- Population standard deviation for group 1:

- Population standard deviation for group 2:

- Sample standard deviation for group 1:

- Sample standard deviation for group 2:

- Sample size for group 1:

- Sample size for group 2:

Why should we treat this as two independent groups rather than paired data?

Hypothesis Testing

Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit

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 alternative hypothesis.

Write the null and alternative hypotheses for the letters study:

In notation:

H_0 :

H_A :

```
letters<-read.csv("data/letters.csv")
letters %>%
    reframe(favstats(Memorized~Grouped))

#>           Grouped min Q1 median    Q3 max     mean       sd  n missing
#> 1 NotRecognizable   1  6     12 14.75  24 11.15385 6.576883 26       0
#> 2    Recognizable   1  6     15 21.00  30 14.32000 8.518216 25       0
```

270

Summary statistic:

Interpret the summary statistic in context of the problem:
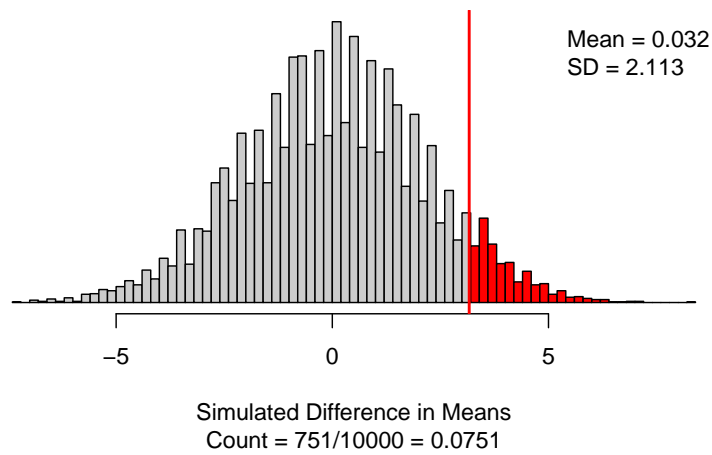
```
letters%>%
  ggplot(aes(y = Memorized, x = Grouped))  + #Enter the name of the explanatory and response variable
  geom_boxplot()+
  labs(title = "Boxplot of Number of Letters memorized by Type
       of Grouping for College Students", #Title your plot
       y = "Number of letters memorized", #y-axis label
       x = "Letter Grouping") #x-axis label
```
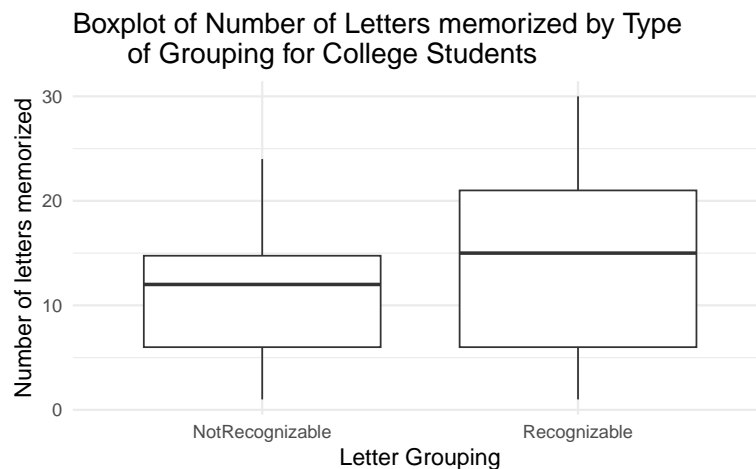

Boxplot of Number of Letters memorized by Type of Grouping for College Students

Simulation-based method

- Simulate many samples assuming $H_0 : \mu_1 = \mu_2$
  - Write the response variable values on cards
  - Mix the explanatory variable groups together
  - Shuffle cards into two explanatory variable groups to represent the sample size in each group (n_1 and n_2)
  - Calculate and plot the simulated difference in sample means from each simulation
  - Repeat 10000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}_1 - \bar{x}_2$

```
set.seed(216)
two_mean_test(Memorized~Grouped, #Enter the names of the variables
              data = letters,  # Enter the name of the dataset
              first_in_subtraction = "Recognizable", # First outcome in order of subtraction
              number_repetitions = 10000,  # Number of simulations
              as_extreme_as = 3.166,  # Observed statistic
              direction = "greater")  # Direction of alternative: "greater", "less", or "two-sided"
```

271

Mean = 0.032
SD = 2.113

Simulated Difference in Means
Count = 751/10000 = 0.0751

Explain why the null distribution is centered at the value of zero:

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

Confidence interval

To estimate the difference in true mean we will create a confidence interval.

Simulation-based method - Video 19.2

- Write the response variable values on cards

- Keep explanatory variable groups separate

- Sample with replacement n_1 times in explanatory variable group 1 and n_2 times in explanatory variable group 2

- Calculate and plot the simulated difference in sample means from each simulation

- Repeat 10000 times (simulations) to create the bootstrap distribution

- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

For the letters example, we will estimate the difference in true mean number of letters recognized for students given recognizable letter groupings and students given non-recognizable letter groupings.

```
set.seed(216)
two_mean_bootstrap_CI(Memorized ~ Grouped, #Enter the name of the variables
                      data = letters,  # Enter the name of the data set
                      first_in_subtraction = "Recognizable", # First value in order of subtraction
                      number_repetitions = 10000,  # Number of simulations
                      confidence_level = 0.95)
```


Bootstrap Difference in Means
95% CI: (−0.928, 7.211)
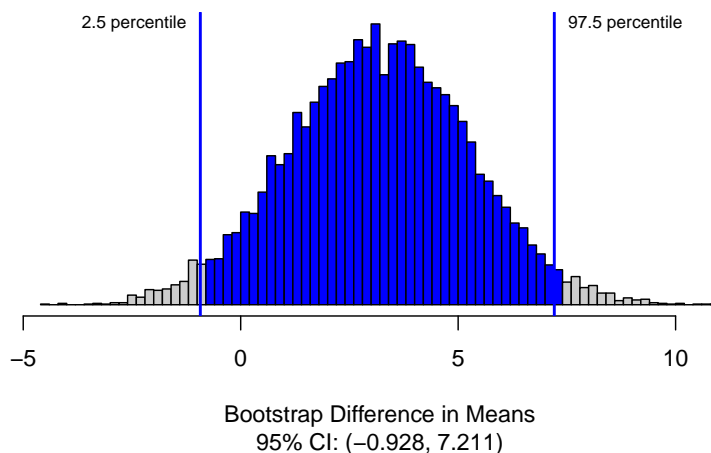
Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)

- Parameter of interest

- Calculated interval

- Order of subtraction when comparing two groups

Theory-based method - Video 19.3TheoryTests

Example: Every year, orange and black monarch butterflies migrate from their summer breeding grounds in the US and Canada to mountain forests in central Mexico, where they hibernate for the winter. Due to abnormal weather patterns and drought affecting monarch habitats and feeding grounds, the population of monarch butterflies is estimated to have decreased by 53% from the 2018-2019 wintering season to the 2019-2020 wintering season (WWF, 2020). While conservationists often resort to captive-rearing with the goal of raising biologically indistinct individuals for release into the wild, tagging studies have shown that captive-reared monarchs have lower migratory success compared to wild monarchs. For this study, the researchers raised 67 monarchs (descended from wild monarchs) from eggs to maturity and then compared them to a group of 40 wild-caught monarchs. The researchers want to explore whether the maximum grip strength (how many Newtons a butterfly exerts at the moment of release when gently tugged from a mesh-covered perch) differs between captive-reared and wild-caught monarchs. Use Captive – Wild for order of subtraction.

Write the null and alternative hypotheses in notation.

H_0 :

H_A :


```r
butterfly <-read.csv("data/butterfly1.csv")

butterflies <- butterfly %>% na.omit() %>%
    rename(Monarch_Group = "Monarch.Group",
           MaxGrip = "Max.Grip.Strength..N.") %>%
    mutate(Monarch_Group = factor(Monarch_Group),
           Sex = factor(Sex)) %>%
    mutate(Monarch_Group = fct_collapse(Monarch_Group,"Captive" = c("Incubator – Fall conditions", "Reari

butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group   min    Q1 median     Q3   max      mean          sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```


Boxplot of Max Grip of Monarch Butterflies by Environment Reared
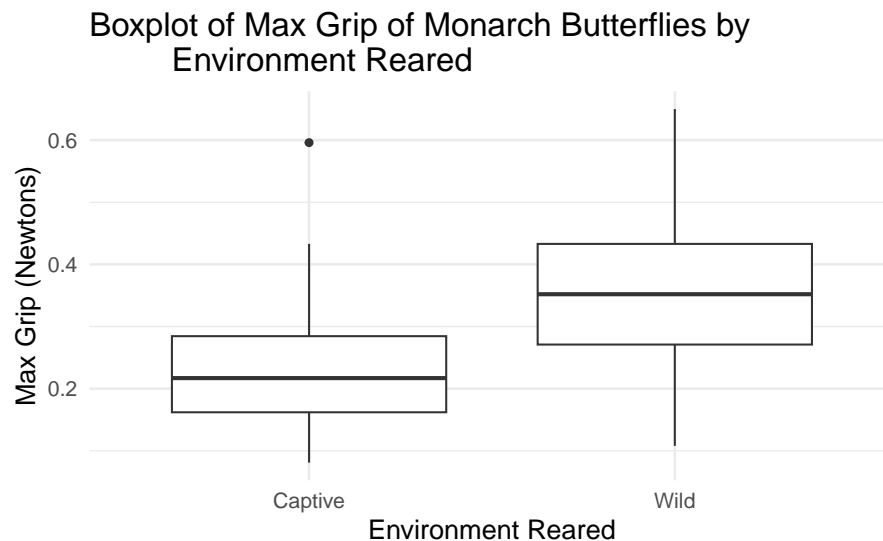
Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit

- Large enough sample size

Like with paired data the t-distribution can be used to model the difference in means.

- For independent samples we use the _____- distribution with _____ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic

- Find the area under the t-distribution with the smallest $n - 1$ df $[\min(n_1 - 1, n_2 - 1)]$ at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample mean:

Equation for the standardized difference in sample mean:

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

Calculate the standardized difference in mean max grip strength.

- First calculate the $SE(\bar{x}_1 - \bar{x}_2)$

- Then calculate the T-score

What theoretical distribution should we use to find the p-value?

To find the theory-based p-value:

```r
pt(-5, df=39, lower.tail=FALSE)*2
```

```
#> [1] 1.999987
```

Conclusion:

- Amount of evidence

- Parameter of interest

- Direction of the alternative hypothesis

Confidence Interval - Video 19.3TheoryIntervals

- Calculate the interval centered at the sample statistic

  statistic $\pm$ margin of error

Using the butterfly data, calculate the 99% confidence interval.

```r
butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group   min    Q1 median     Q3   max      mean         sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```

- Need the t^* multiplier for a 99% confidence interval from a t-distribution with _____ df.

```r
qt(0.995, df=39, lower.tail = TRUE)
```

```
#> [1] 2.707913
```

- We will use the same value for the $SE(\bar{x}_1 - \bar{x}_2)$ as calculated for the standardized statistic.

Calculate the margin of error for a 99% confidence interval for the parameter of interest.

Calculate a 99% confidence interval for the parameter of interest.

12.2.2 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. Why is the recognizable letter study analyzed as two independent groups rather than paired data?

2. Write out the equation for the standard error for a difference in sample means.

12.3   Activity 24: Does behavior impact performance?

12.3.1   Learning outcomes

- Create a side-by-side boxplot of one categorical explanatory variable and one quantitative response variable

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a difference in means.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in means.

- Use bootstrapping to find a confidence interval for a difference in means.

- Interpret a confidence interval for a difference in means.

- Use a confidence interval to determine the conclusion of a hypothesis test.

12.3.2   Terminology review

In today's activity, we will use simulation-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Independent groups

- Difference in means

To review these concepts, see Chapter 19 in the textbook.

12.3.3   Behavior and Performance

A study in the Academy of Management Journal (Porath 2017) investigated how rude behaviors influence a victim's task performance. Randomly selected college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in five minutes; this value (total number of uses) was used as a performance measure for each student, where higher values indicate better performance. During this time another individual showed up late for class. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to the late-arriving person). No comments were made about the late-arriving person for students in the control group. Is there evidence that the average performance score for students in the rudeness condition is lower than for students in the control group? Use the order of subtraction of rudeness – control.

- Download the R script file from D2L and upload to the RStudio server

- Highlight and run lines 1–7 to load the data

```
# Read in data set
rude <- read.csv("https://math.montana.edu/courses/s216/data/rude.csv")
```

To create a plot of the data and a table of summary statistics:

- Highlight and run lines 11–19

```
# Side-by-side box plots
rude %>%
ggplot(aes(x = condition, y = number_of_uses)) +
    geom_boxplot() +
    labs(title = "Number of Uses for a Brick based on Behavior Condition
        for College Students in a Management Course",
        x = "Behavior")
# Summary statistics
rude %>%
    reframe(favstats(number_of_uses ~ condition))
#>   condition min Q1 median Q3 max       mean       sd  n missing
#> 1   control   0  6     12 17  30 11.811321 7.382559 53       0
#> 2  rudeness   0  6      9 11  18  8.511111 3.992164 45       0
```


Number of Uses for a Brick based on Behavior Condition
for College Students in a Management Course
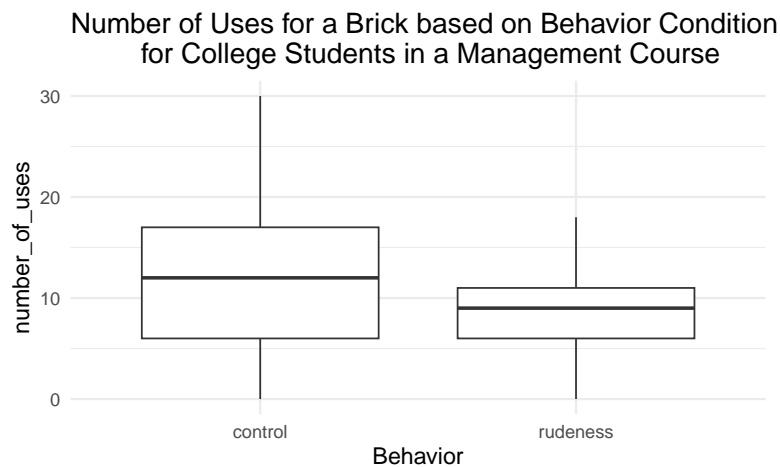
Quantitative variables review

1. Compare the distributions of the number of bricks between the two treatment conditions.

   - What is the shape of each group?

   - Which group has the higher center?

   - What group has the larger spread?

   - Does either distribution have outliers?

2. Is this an experiment or an observational study? Justify your answer.

3. Explain why this is two independent samples and not paired data.

Ask a research question

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

4. What assumption are we making about the difference in true mean?

5. Write the alternative hypothesis in notation.

Numerically Summarize the data

6. Calculate the summary statistic of interest (difference in means). What is the appropriate notation for this statistic?

Interpret this calculated value.

7. Write out the parameter of interest for this study in context of the study.
   - To write in context:
     - Population word (true, long-run, population)
     - Summary measure (depends on the type of data)
     - Context
       * Observational units
       * Variable(s)

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

Use statistical inferential methods to draw inferences from the data

Hypothesis test   Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that there is no association between the two variables. This means that the values observed in the data set would have been the same regardless of the behavior condition.

To demonstrate this simulation, we could create cards to simulate a sample.

- Write the number of uses for a brick given by each student on one card.
- Mix together and shuffle into two piles, one with 45 cards to represent the rudeness condition and one with 53 cards to represent the control group.
- Calculate the difference in mean number of uses for a brick (rudeness - control)

We will use the `two_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of differences in sample means and compute a p-value.

- Fill in the response and explanatory variable names
- Fill in the missing values/names for the xx's in the R script file
- Highlight and run lines 25–30

```
set.seed(216)
two_mean_test(response~explanatory, #Enter the names of the variables
              data = rude,  # Enter the name of the dataset
              first_in_subtraction = "xx", # First outcome in order of subtraction
              number_repetitions = 10000,  # Number of simulations
              as_extreme_as = xx,  # Observed statistic
              direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

8. Report the p-value. Based off of this p-value, write a conclusion to the hypothesis test.

Confidence interval   We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `rude`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (the count for the numerator when calculating a sample proportion), and the confidence level as a decimal.

The response variable name is `number_of_uses` and the explanatory variable name is `condition`.

9. What values should be entered for each of the following into the simulation to create a 99% confidence interval?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"rudeness"` or `"control"`):

- Number of repetitions:

- Confidence level (entered as a decimal):

Using the R script file for this activity, enter your answers for question 9 in place of the **xx**'s to produce the bootstrap distribution with 10000 simulations; highlight and run lines 35–39.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = rude,  # Enter the name of the data set
                      first_in_subtraction = "xx", # First value in order of subtraction
                      number_repetitions = 10000,  # Number of simulations
                      confidence_level = xx)
```

10. Where is the bootstrap distribution centered? Explain why.

11. Report the bootstrap 99% confidence interval.

12. What percentile of the bootstrap distribution does the upper value of the confidence interval represent?

13. Interpret the 99% confidence interval.

12.3.4   Take-home messages

1. This activity differs from the activities in Module 11 because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.

2. To create one simulated sample on the null distribution for a difference in sample means, label cards with the response variable values from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2. Calculate and plot the difference in means.

3. To create one simulated sample on the bootstrap distribution for a difference in sample means, label $n_1 + n_2$ cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.

12.3.5   Additional notes

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

12.4 Activity 25: Moon Phases and Virtual Reality

12.4.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a theory-based hypothesis test for a difference in means.

- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.

- Use theory-based methods to find a confidence interval for a difference in means.

- Interpret a confidence interval for a difference in means.

- Use a confidence interval to determine the conclusion of a hypothesis test.

12.4.2 Terminology review

In today's activity, we will use theory-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Difference in means

- Independence within and between groups

- Normality

To review these concepts, see Chapter 19 in the textbook.

12.4.3 Moon Phases and Virtual Reality

In a study comparing immersive virtual reality (VR) to traditional hands-on methods, researchers recruited 115 undergraduate students to assess the effectiveness of these approaches in teaching complex scientific concepts like Moon phases (Madden 2020). Participants were randomly assigned to experience either a VR simulation replicating the Sun-Earth-Moon system or a hands-on activity where they physically manipulated models to observe Moon phases. The students were given a 14 multiple choice question quiz about Moon phases and the Moon's motion relative to the Earth to evaluate their understanding of Moon phases and the Moon's motion. Each question had only one correct answer, and the participant's score was the sum of the number of correct answers, with all questions weighted equally (with a maximum score of 14). Is there evidence of a difference, on average, in student learning comparing those using VR methods to those using the traditional method? Use order of subtraction VR – Hands-on.

- Download the RScript file and dataset from D2L and upload to the RStudio server

- Open the RScript file

- Enter the name of the data set for datasetname in line 8

- Highlight and run lines 1–8

```
moon <- read.csv("datasetname.csv")
```

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

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence**: the sample's observations are independent, e.g., are from a simple random sample and there is independence between groups. (*Remember*: This also must be true to use simulation methods!)

- **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. *When we have two samples, we need to check this condition for each group!* To check this condition, use the following rules of thumb (for both n_1 and n_2):

  - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

To create the plots of the data:

- Enter a title in line 15 for the plot between the quotations

- Highlight and run lines 12 - 17

```
moon %>%  # Data set piped into...
  ggplot(aes(y = TestScore, x = Method))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Don't forget to title your plot!",  # Title
       x = "Methods",    # x-axis label
       y = "Test Score (points)")  # y-axis label
```

To find the summary statistic:

- Enter the response and explanatory variable names in line 22

- Highlight and run lines 21–22

```
moon %>%
  reframe(favstats(response~explanatory))
```

4. Can theory-based methods be used to analyze these data?

5. Calculate the summary statistic (difference in means) for this study. Use appropriate notation with clearly defined subscripts.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

6. Calculate the standard error for the difference in sample means.

7. Calculate the standardized statistic for the difference in sample means.

To find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. ($\texttt{df}$ = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the value of the standardized statistic for xx
- Enter the df for yy
- Highlight and run line 27

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

8. Report the p-value for the study. Why did we multiply by two to find the p-value?

To calculate a theory-based 95% confidence interval for a difference in means, in the questions on the next page, we will use the formula:

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$$

First, we will need to find the t^* multiplier using the function $\texttt{qt()}$.

To find the t^* multiplier

- Enter the percentile to find the multiplier for a 95% confidence level
- Enter the degrees of freedom for yy
- Highlight and run line 32

```
qt(percentile, df = yy, lower.tail=TRUE)
```

9. Calculate the margin of error for a 95% confidence interval.

10. Calculate the 95% confidence interval.

11. Write a conclusion to the test.

12.4.4 Take-home messages

1. In order to use theory-based methods for independent groups, the normality condition must be met for each sample.

2. A T-score is compared to a t-distribution with the minimum $n-1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.

3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with the minimum $n-1$ df.

12.4.5 Additional notes

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

12.5 Module 12 Lab: Trustworthiness

12.5.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a theory-based hypothesis test for a difference in means.

- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.

- Use theory-based methods to find a confidence interval for a difference in means.

- Interpret a confidence interval for a difference in means.

- Use a confidence interval to determine the conclusion of a hypothesis test.

12.5.2 Trustworthiness

Researchers in India wanted to find out how trustworthy famous YouTubers are (Kalra 2022). They went through a process in which they collected data on many videos from famous YouTubers to determine a trustworthiness score. Scientists randomly selected videos from famous YouTubers (>1000 subscribers) to include in the study. There were many different factors that went into calculating the trustworthiness score. Researchers also recorded if YouTubers were a subject matter expert (SME) or not a subject matter expert (non-SME). An example of an SME would be if one of your statistics professors made a YouTube video of how to do hypothesis testing. An example of someone who isn't an SME would be if one of your friends who has never taken a civil engineering class in their life decided to make a YouTube video about how to build a bridge. There were 621 Youtubers who are SMEs in the sample and 1026 who aren't SMEs. Is there evidence of a difference in mean trustworthiness score between subject matter experts (SME) YouTubers and non-SME YouTubers? Use SME – Non -SME as the order of subtraction

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

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

- **Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution**:

    - **Independence**: the sample's observations are independent, e.g., are from a simple random sample and there is independence between groups. (*Remember*: This also must be true to use simulation methods!)

– **Normality Condition**: either the sample observations come from a normally distributed population or we have a large enough sample size. *When we have two samples, we need to check this condition for each group!* To check this condition, use the following rules of thumb (for both n_1 and n_2):

– $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

– $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

– $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

- Upload and open the R script file for Module 10 lab. Upload the csv file, `Trustworthiness.csv`.

- Enter the name of the data set for datasetname in the R script file in line 10.

- Write a title for the boxplots in line 14.

- Highlight and run lines 1–16 to load the data and create plots of the data.

```
trust <- read.csv("datasetname")
trust %>%  # Data set piped into...
  ggplot(aes(y = Trustworthiness_Video, x = Creator_SME))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Don't forget to include a title",  # Title: should include the type of plot,
       # observational units, variables
       x = "Whether the Creator is SME",    # x-axis label
       y = "Trustworthiness Score")  # y-axis label
```

4. Is the independence condition met? Explain your answer.

5. Check that the normality condition is met to use theory-based methods to analyze these data.

- Enter the name of the explanatory variable for `explanatory` and the name of the response variable for `response` in line 22.

- Highlight and run lines 21–22 to get the summary statistics for the data.

```
trust %>%
  reframe(favstats(response~explantory))
```

6. **Calculate the summary measure (difference in means) for this study. Use appropriate notation with clearly defined subscripts.**

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

7. Calculate the standard error for the difference in sample means.

8. **Calculate the standardized statistic for the difference in sample means.**

9. When we are comparing two quantitative variables to find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (df = minimum of $n_1 - 1$ or $n_2 - 1$). Calculate the df for this study.

10. Using the provided R script file, enter the T-score (for `xx`) and the `df` calculated in question 9 for `yy` into the `pt()` function to find the p-value. Highlight and run line 27. Report the p-value calculated.

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

11. **Explain why we multiplied by 2 in the code above.**

12. Do you expect the 95% confidence interval to contain the null value of zero? Explain your answer.

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$(\bar{x}_1 - \bar{x}_2) \pm (t^* \times SE(\bar{x}_1 - \bar{x}_2))$$

We will need to find the t^* multiplier using the function `qt()`. For a 95% confidence level, we are finding the t^* value at the 97.5th percentile with (df = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the appropriate percentile value degrees of freedom into the `qt()` function at line 32 to find the appropriate t^* multiplier

```
qt(xx, df = yy, lower.tail=FALSE)
```

13. Report the t^* multiplier for the 95% confidence interval.

14. Calculate the 95% confidence interval using theory-based methods.

15. Do the results of the CI agree with the p-value? Explain your answer.

16. What type of error may be possible?

17. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:

- Summary statistic and interpretation
- 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

Paragraph continued:

Exploratory Data Analysis and Inference for Two Quantitative Variables

13.1   Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of two quantitative variables.

13.1.1   Key topics

Module 13 will cover exploratory data analysis and both simulation-based and theory-based methods of inference for two quantitative variables. The **summary measure** for two quantitative variables is either the **slope** of a regression line or the **correlation** between the two variables.

- Notation for a sample regression slope: b_1

- Notation for a population regression slope: β_1

- Notation for a sample correlation: r

- Notation for a population correlation: ρ

Types of plots for two quantitative variables:

- Scatterplot

13.1.2   Vocabulary

Plotting two quantitative variables

- **Scatterplot**: plots (x, y) pairs of observations with the explanatory variable on the x-axis and the response variable on the y-axis. R code to create a scatterplot:

```
object %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "x-axis label",  # Label x-axis
   y = "y-axis label",  # Label y-axis
   title = "Don't forget to add a title!") +
          # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

  - If there is a third categorical variable, you can use color or shape to include the third variable on the scatterplot.

- Four characteristics of scatterplots:

  - Form (linear or non-linear)

  - Direction (positive or negative)

  - Strength (weak, moderate, or strong)

– Outliers?

Sample statistics for two quantitative variables

- **Least-squares regression line**: a line fit to the data which minimizes the squared vertical distances from the observed y-value to the line

    – **Notation for the fitted least-squares regression line**:

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

    or

    $$\widehat{response} = b_0 + b_1 \times explanatory$$

    To write the equation of the regression line in context of the problem, include descriptive names of the response and explanatory variables for "y/response" and "x/explanatory" above.

    – b_0 is the y-**intercept** of the regression line: the *predicted* value of the response variable when the explanatory variable is equal to zero.

    – b_1 is the **slope** of the regression line: the *predicted* increase/decrease in the response variable associated with a one-unit increase in the explanatory variable.

    – The distance from an observation's y-value (observed response) to its fitted value, $\hat{y}$ (the value on the line) is called a **residual**:

    $$\text{residual} = \text{observed} - \text{fitted} = y - \hat{y}$$

    – A least-squares regression line is a special case of a **linear model**.

    – R code to find the least-squares regression line (fit the linear model):

```
linearmodel <- lm(response~explanatory, data=object)
round(summary(linearmodel)$coefficients,3) # Display coefficients
```

- **Correlation**: measures the magnitude and direction of the linear relationship between two quantitative variables.

    – Parameter notation: ρ

    – Sample notation: r

    – R code to find the **correlation** matrix between variables:

```
object %>%  # Data set pipes into
    select(c("variable1", "variable2", "variable3")) %>%
    #Selects the variables you want to compare
    cor(use="pairwise.complete.obs") %>%  #Calculates the correlation between each pair
    round(3) #Rounds to 3 decimal places
```

- **Coefficient of determination**: measures the proportion of total variability in the response variable that is explained by the linear relationship with the explanatory variable. The coefficient of determination can be calculated in three ways:

$$r^2 = (r)^2 = \frac{SST - SSE}{SST} = \frac{s_y^2 - s_{residual}^2}{s_y^2}$$

Hypotheses

Hypotheses involving two quantitative variables can be expressed either in terms of the slope or the correlation. When either the slope or correlation is equal to zero, there is no linear relationship between the two quantitative

variables (the null hypothesis).

- **Hypotheses in notation for slope**:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} 0$$

- **Hypotheses in notation for correlation**:

$$H_0 : \rho = 0$$

$$H_A : \rho \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} 0$$

Simulation-based inference for two quantitative variables

- **Conditions necessary to use simulation-based methods for inference for two quantitative variables**:

  - **Independence**: observational units (the (x, y) pairs) must be independent of one another.

  - **Linearity**: the form of the relationship (if any) between the two variables must be linear.

- **Simulation-based methods to create the null distribution**: R code for simulation methods to find the p-value using the `regression_test` function in the `catstats` package.

```
regression_test(response~explanatory, # response ~ explanatory
           data = object, # Name of data set
           direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
           summary_measure = "xx", # "slope" or "correlation"
           as_extreme_as = xx, # Observed slope or correlation
           number_repetitions = 10000) # Number of simulated samples for null distribution
```

- **Simulation-based methods to create the bootstrap distribution**: R code to find the simulation-based confidence interval using the `regression_bootstrap_CI` function from the `catstats` package.

```
regression_bootstrap_CI(response~explanatory, # response ~ explanatory
   data = object, # Name of data set
   confidence_level = xx, # Confidence level as decimal
   summary_measure = "xx", # Slope or correlation
   number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

Theory-based methods for two quantitative variables

- **Conditions necessary to use theory-based methods for inference for two quantitative variables**:

  - **Independence** (for both simulation-based and theory-based methods): observational units (the (x, y) pairs) must be independent of one another.

    * Check this assumption by investigating the sampling method and determining if the observational units are related in any way.

– **Linearity** (for both simulation-based and theory-based methods): the form of the relationship (if any) between the two variables must be linear.

  * Check this assumption by examining the scatterplot of the two variables, and a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The pattern in the residuals vs. fitted plot should display a horizontal line.

– **Constant variability** (for theory-based methods only): the variability of points around the least squares line remains roughly constant

  * Check this assumption by examining a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The variability in the residuals around zero should be approximately the same for all fitted values.

– **Nearly normal residuals** (for theory-based methods only): residuals must be nearly normal.

  * Check this assumption by examining a histogram of the residuals, which should appear approximately normal.

- **Standard error of the slope of the least-squares regression line** ($SE(b_1)$): obtain the value of the standard error of the slope from the linear model (`lm`) R output.

- **Standardized slope**:
$$T = \frac{\text{slope estimate} - nullvalue}{SE} = \frac{b_1 - 0}{SE(b_1)}.$$

  – The p-value can be found from the linear model (`lm`) R output or by using the `pt` function in R to find the area under a t-distribution with $n - 2$ degrees of freedom where T is as or more extreme as the value observed (in the direction of H_A).

- **Margin of error**: half the width of the confidence interval. For a regression slope, the margin of error is:
$$ME = t^* \times SE(b_1)$$

where t^* is the **multiplier**, corresponding to the desired confidence level found from a t-distribution with $n - 2$ degrees of freedom.

  – Use the `qt` function in R to find the t^* multiplier with $n - 2$ degrees of freedom.

  – To find the endpoints of a confidence interval, add and subtract the margin of error to the sample statistic. The confidence interval for a population slope is:
$$b_1 \pm ME$$

13.2 Video Notes: Regression and Correlation

Read Chapters 6, 7, 8, 21, and 22 in the course textbook. Use the following videos to complete the video notes for Module 13.

13.2.1 Course Videos

- 6.1
- 6.2
- 6.3
- Ch 7
- 21.1
- 21.3
- 21.4TheoryTests
- 21.4TheoryIntervals

Summary measures and plots for two quantitative variables - Videos 6.1 - 6.3

Example: Data were collected from 1236 births between 1960 and 1967 in the San Francisco East Bay area to better understand what variables contributed to child birthweight, as children with low birthweight often suffer from an array of complications later in life ("Child Health and Development Studies," n.d.). There were some missing values in the study and with those observations removed we have a total of 1223 births.

```
babies<-read.csv("data/babies.csv") %>%
    drop_na(bwt) %>%
    drop_na(gestation)
glimpse(babies)
#> Rows: 1,223
#> Columns: 8
#> $ case      <int> 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
#> $ bwt       <int> 120, 113, 128, 108, 136, 138, 132, 120, 143, 140, 144, 141, ~
#> $ gestation <int> 284, 282, 279, 282, 286, 244, 245, 289, 299, 351, 282, 279, ~
#> $ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
#> $ age       <int> 27, 33, 28, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, 38, ~
#> $ height    <int> 62, 64, 64, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, 63, ~
#> $ weight    <int> 100, 135, 115, 125, 93, 178, 140, 125, 136, 120, 124, 128, 9~
#> $ smoke     <int> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, ~
```

Here you see a glimpse of the data. The 1223 rows correspond to the sample size. The case variable is labeling each pregnancy 1 through 1223. Then 7 variables are recorded. birthweight (bwt), length of gestation in days, parity is called an indicator variable telling us if the pregnancy was a first pregnancy (labeled as 0) or not (labeled as 1) were recorded about the child and pregnancy. The age, height, and weight were recorded for the mother giving birth, as was smoke, another indicator variable where 0 means the mother did not smoke during pregnancy, and 1 indicates that she did smoke while pregnant.

Type of plot

A _____ is used to display the relationship between two _____ variables.

Four characteristics of the scatterplot:

- Form:

- Direction:

- Strength:

- Outliers:

  - Influential points: outliers that change the regression line; far from the line of regression

  - High leverage points: outliers that are extreme in the x- axis; far from the mean of the x-axis

The following shows a scatterplot of length of gestation as a predictor of birthweight.

```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight for Births
       between 1960 and 1967 in San Francisco") +
    # Be sure to title your plots with the type of plot, observational units, variable(s)
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
    theme_bw()
```


Scatterplot of Gestation vs. Birthweight for Births
between 1960 and 1967 in San Francisco
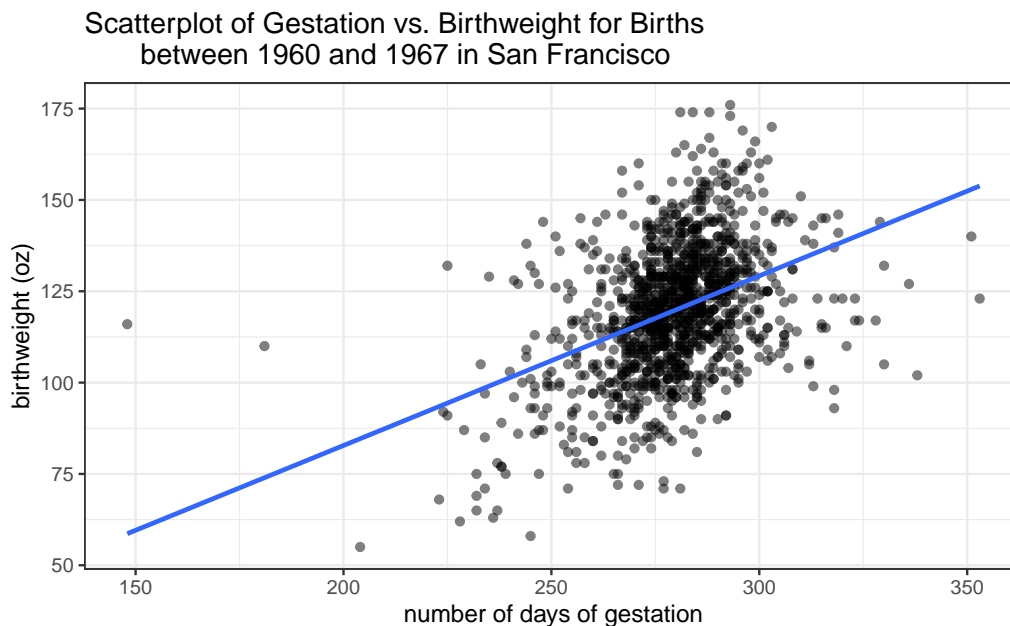
Describe the scatterplot using the four characteristics of a scatterplot.

The summary measures for two quantitative variables are:

- _____

- _____

- _____

Notation:

- Population slope:

- Population correlation:

- Sample slope:

- Sample correlation:

Correlation

Correlation is always between the values of _____ and _____.

- Measures the _____ and _____ of the linear relationship between two quantitative variables.

- The stronger the relationship between the variables the closer the value of _____ is to _____ or _____.

- The sign gives the _____.

The following code creates a correlation matrix between different quantitative variables in the data set.

```
babies %>%
    select(c("gestation", "age", "height", "weight", "bwt")) %>%
    cor(use="pairwise.complete.obs") %>%
    round(3)
```

```
#>          gestation    age height weight   bwt
#> gestation    1.000 -0.056  0.064  0.022 0.408
#> age         -0.056  1.000 -0.005  0.147 0.029
#> height       0.064 -0.005  1.000  0.436 0.201
#> weight       0.022  0.147  0.436  1.000 0.154
#> bwt          0.408  0.029  0.201  0.154 1.000
```

The value of correlation between gestation and birthweight is _____. This shows a _____, _____ relationship between gestation and birthweight.

Slope

- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem) or $\widehat{response} = b_0 + b_1 \times$ explanatory

- $\hat{y}$ or $\widehat{response}$ is

- b_0 is

- b_1 is

- x or explanatory is

298

- The estimates for the linear model output will give the value of the _____ and the _____.

- Interpretation of slope: an increase in the _____ variable of 1 unit is associated with an increase/decrease in the _____ variable by the value of slope, on average.

- Interpretation of the y-intercept: for a value of 0 for the _____ variable, the predicted value for the _____ variable would be the value of y-intercept.

- We can predict values of the _____ variable by plugging in a given _____ variable value using the least squares equation line.

- A prediction of a response variable value for an explanatory value outside the range of x values is called _____.

- To find how far the predicted value deviates from the actual value we find the _____.

- To find the least squares regression line the line with the _____ SSE is found.

  SSE = sum of squared errors

  – To find SSE, the residual for each data point is found, squared and all the squared residuals are summed together

The linear model output for this study is given below:

```
# Fit linear model: y ~ x
babiesLM <- lm(bwt ~ gestation, data=babies)
round(summary(babiesLM)$coefficients,3) # Display coefficient summary

#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  -10.064      8.322  -1.209    0.227
#> gestation      0.464      0.030  15.609    0.000
```

Write the least squares equation of the line.

Interpret the slope in context of the problem.

Interpret the y-intercept in context of the problem.

Predict the birthweight for a birth with a baby born at 310 days gestation.

Calculate the residual for a birth of a baby with a birthweight of 151 ounces and born at 310 days gestation.

Is this value (310, 151) above or below the line of regression? Did the line of regression overestimate or underestimate the birthweight?
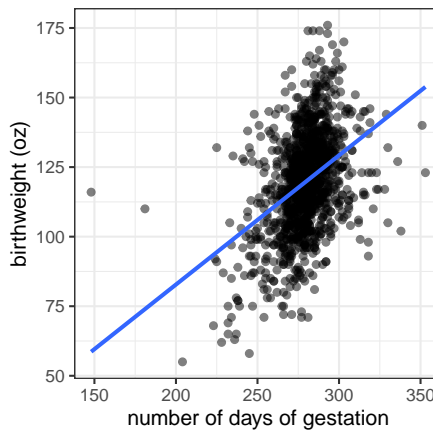
Coefficient of Determination

The coefficient of determination can be found by squaring the value of correlation, using the variances for each variable or using the SSE (sum of squares error) and SST (sum of squares total)

- $r^2 = (r)^2 = \frac{SST - SSE}{SST} = \frac{s_y^2 - s_{residual}^2}{s_y^2}$

- The coefficient of determination measures the _____ of total variation in the _____ variable that is explained by the changes in the _____ variable.

A   Scatterplot of Gestation vs. Birthweight for Births between 1960 and 1967 in SF with Horizontal Line
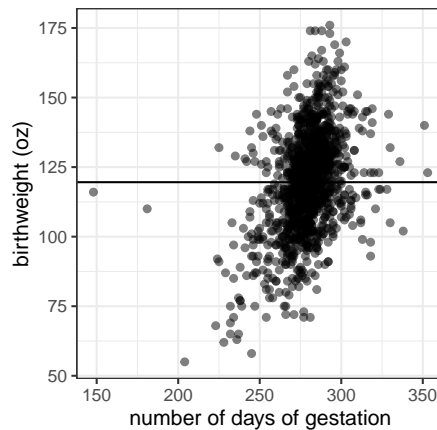
B   Scatterplot of Gestation vs. Birthweight for Births between 1960 and 1967 in SF with Regression Line

The value for SST was calculated as 406753.48. The value for SSE was calculated as 339092.13.

Calculate the coefficient of determination between gestation and birthweight.

Interpret the coefficient of determination between gestation and birthweight.

Multivariable plots - Video Chapter7

Aesthetics: visual property of the objects in your plot

- Position on the axes: groups for _____ variables, or a number line if the variable is

  _____

- Color or shape - to represent _____ variables

- Size - to represent _____ variables

Adding the quantitative variable maternal age to the scatterplot between gestation and birthweight.

```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+  # Specify variables
  geom_point(alpha=0.5, shape=1, aes(size=age)) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight by Age
       for Births between 1960 and 1967 in San Francisco") +
    # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```


Scatterplot of Gestation vs. Birthweight by Age for Births between 1960 and 1967 in San Francisco
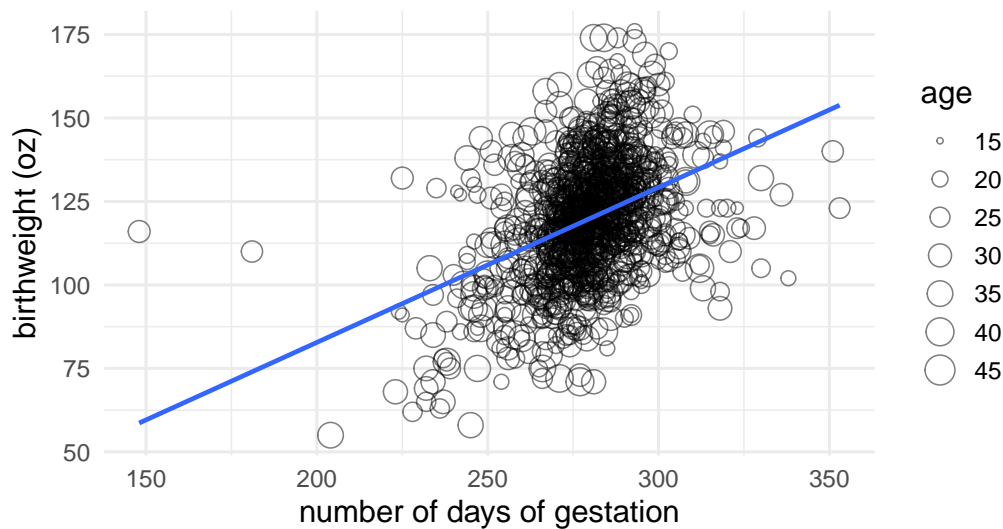
Let's add the categorical variable, whether a mother smoked, to the scatterplot between gestation and birthweight.

```
babies <- babies %>%
    mutate(smoke = factor(smoke)) %>%
    na.omit()

babies %>% # Data set pipes into...
    ggplot(aes(x = gestation, y = bwt, color = smoke))+  #Specify variables
    geom_point(aes(shape = smoke), size = 2) +  #Add scatterplot of points
    labs(x = "number of days of gestation",  #Label x-axis
        y = "birthweight (oz)",  #Label y-axis
        title = "Scatterplot of Gestation vs. Birthweight by
        Smoking Status for Births between 1960 and 1967
        in San Francisco") +
    #Be sure to title your plots
    geom_smooth(method = "lm", se = FALSE) + #Add regression line
    scale_color_grey()
```
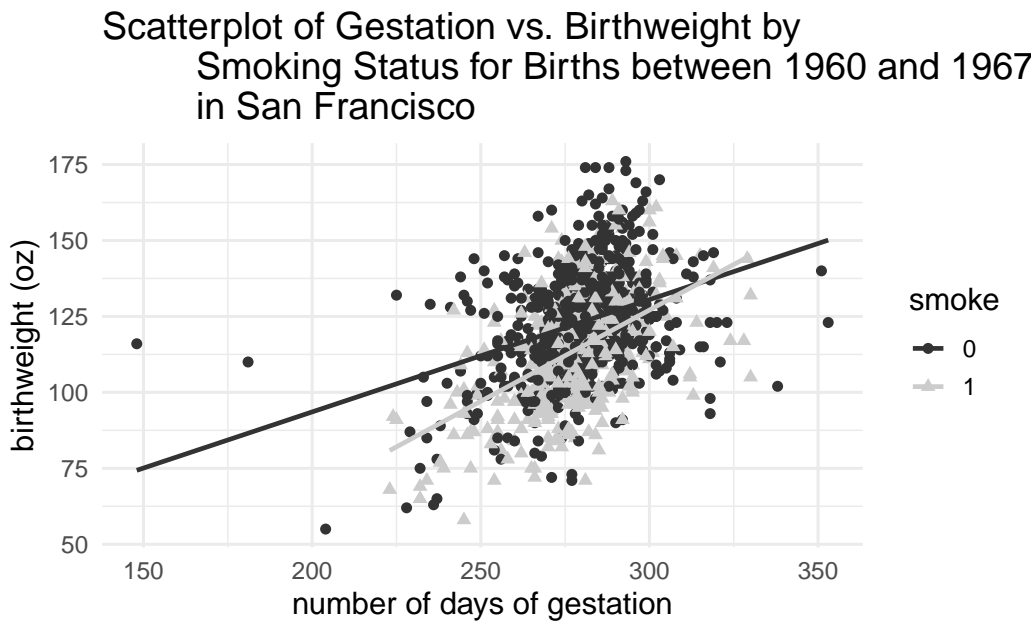

Does the relationship between length of gestation and birthweight appear to depend upon maternal smoking status?

Is the variable smoking status a potential confounding variable?

Adding a categorical predictor:

- Look at the regression line for each level of the _____

- If the slopes are _____, the two predictor variables do not _____ to help explain the response

- If the slopes _____, there is an interaction between the categorical predictor and the relationship between the two quantitative variables.

13.2.2 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 are the three summary measures for two quantitative variables?

2. What are the four characteristics used to describe a scatterplot?

3. When we add a categorical predictor variable to a scatterplot of two quantitative variables, what summary measure will we compare across the categories to assess the change in the relationship between the two quantitative variables.

13.2.3   Video Notes: Inference for Two Quantitative Variables

Example: Oceanic temperature is important for sea life. The California Cooperative Oceanic Fisheries Investigations has measured several variables on the Pacific Ocean for more than 70 years hoping to better understand weather patterns and impacts on ocean life. ("Ocean Temperature and Salinity Study," n.d.) For this example, we will look at the most recent 100 measurements of salt water salinity (measured in PSUs or practical salinity units) and the temperature of the ocean measured in degrees Celsius. Is there evidence that water temperature in the Pacific Ocean tends to decrease with higher levels of salinity?

Hypothesis Testing - Video 21.1

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 alternative hypothesis.

Write the null and alternative for the ocean study:

In notation:

H_0 :

H_A :

```
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "salinity (PSUs)",  # Label x-axis
       y = "temperature (C)",  # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
             # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

Scatterplot of Pacific Ocean Salinity vs Temperature
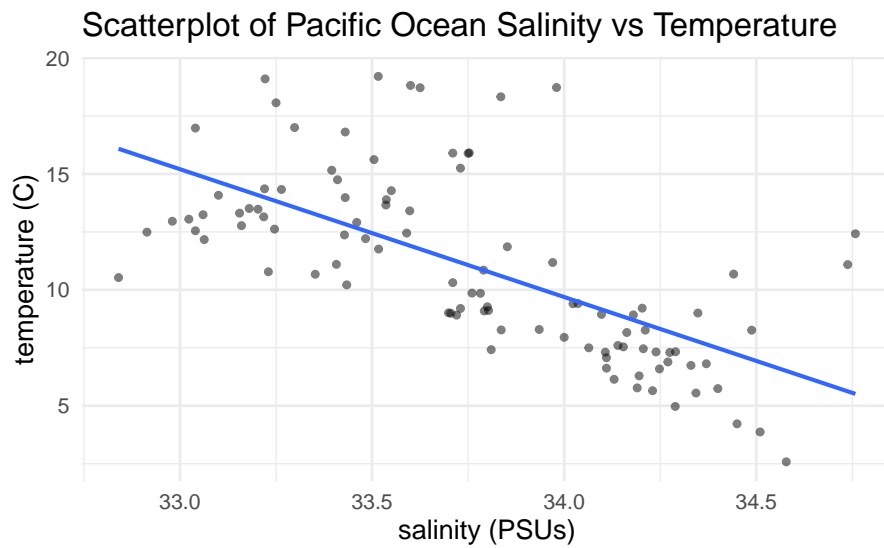
Describe the four characteristics of the scatterplot:

Linear model output:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients, 3)
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18        0
#> Salnty        -5.514      0.636   -8.67        0
```

Correlation:

```
cor(T_degC~Salnty, data=water)
```

```
#> [1] -0.6588365
```

Write the least squares equation of the line in context of the problem:

Interpret the value of slope in the context of the problem:

Report and describe the correlation value:

Calculate and interpret the coefficient of determination:

Simulation-based method

Conditions:

- Independence: the response for one observational unit will not influence another observational unit

- Linear relationship:

- Simulate many samples assuming $H_0 : \beta_1 = 0$ or $H_0 : \rho = 0$
  - Write the response variable values on cards
  - Hold the explanatory variable values constant
  - Shuffle a new response variable to an explanatory variable
  - Plot the shuffled data points to find the least squares line of regression
  - Calculate and plot the simulated slope or correlation from each simulation
  - Repeat 10000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as b_1 or r

To test slope:

```
set.seed(216)
regression_test(T_degC ~ Salnty, # response ~ explanatory
                data = water, # Name of data set
                direction = "less", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "slope", # "slope" or "correlation"
                as_extreme_as = -5.514, # Observed slope or correlation
                number_repetitions = 10000) # Number of simulated samples for null distribution
```


Mean = −0.004
SD = 0.839

Simulated Slope
Count = 0/10000 = 0
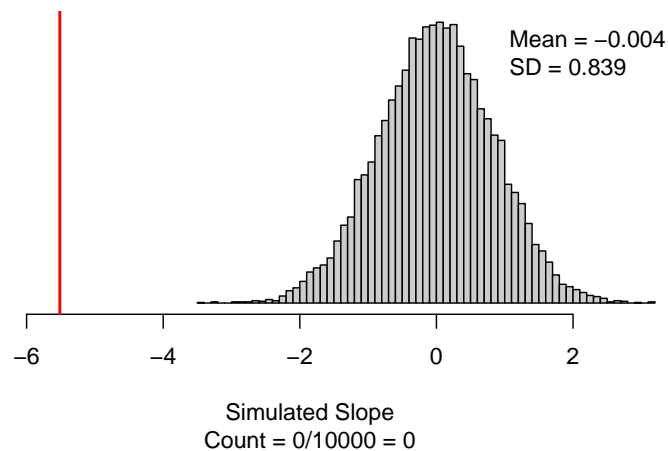
To test correlation:
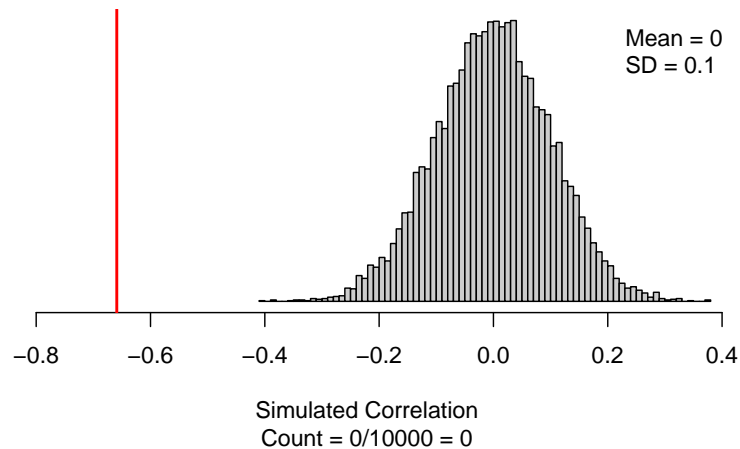
```r
set.seed(216)
regression_test(T_degC~Salnty, # response ~ explanatory
                data = water, # Name of data set
                direction = "less", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "correlation", # "slope" or "correlation"
                as_extreme_as = -0.659, # Observed slope or correlation
                number_repetitions = 10000) # Number of simulated samples for null distribution
```


Mean = 0
SD = 0.1

Simulated Correlation
Count = 0/10000 = 0

Explain why the null distribution is centered at the value of zero:

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
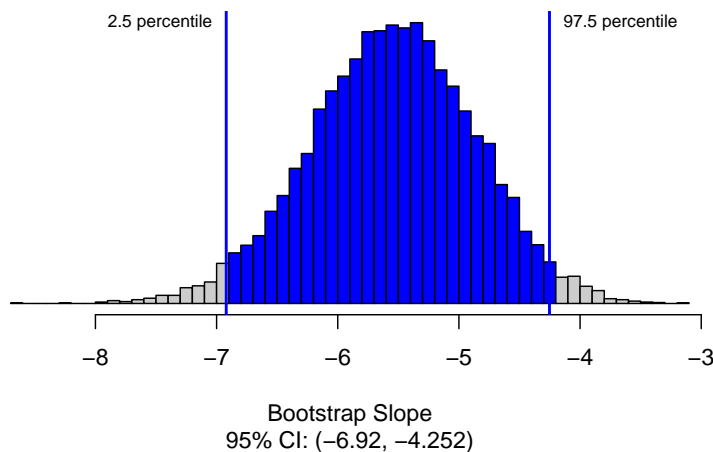
Confidence interval - Video 21.3

To estimate the true slope (or true correlation) we will create a confidence interval.

Simulation-based method

- Write the explanatory and response value pairs on cards

- Sample pairs with replacement n times

- Plot the resampled data points to find the least squares line of regression

- Calculate and plot the simulated slope (or correlation) from each simulation

- Repeat 10000 times (simulations) to create the bootstrap distribution

- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

Returning to the ocean example, we will estimate the true slope between salinity and temperature of the Pacific Ocean.

```
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
   data = water, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "slope", # Slope or correlation
   number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```


Bootstrap Slope
95% CI: (−6.92, −4.252)

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

Now we will estimate the true correlation between salinity and temperature of the Pacific Ocean.

```r
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
   data = water, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "correlation", # Slope or correlation
   number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```


Bootstrap Correlation
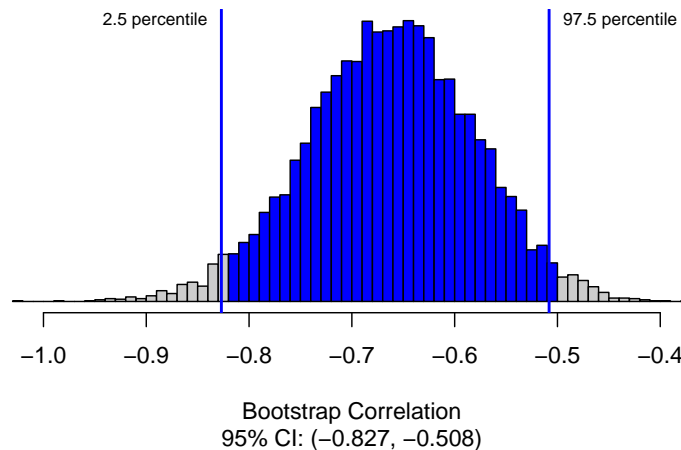95% CI: (−0.827, −0.508)

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

Theory-based method - Video 21.4to21.5TheoryTests

Conditions:

- Linearity (for both simulation-based and theory-based methods): the data should follow a linear trend.

  – Check this assumption by examining the _____ of the two variables, and _____. The pattern in the residual plot should display a horizontal line.

- Independence (for both simulation-based and theory-based methods)

    - One_____for an observational unit has no impact on

        _____.

- Constant variability (for theory-based methods only): the variability of points around the least squares
  line remains roughly constant

    - Check this assumption by examining the _____. The
      variability in the residuals around zero should be approximately the same for all fitted values.

- Nearly normal residuals (for theory-based methods only): residuals must be nearly normal

    - Check this assumption by examining a _____,
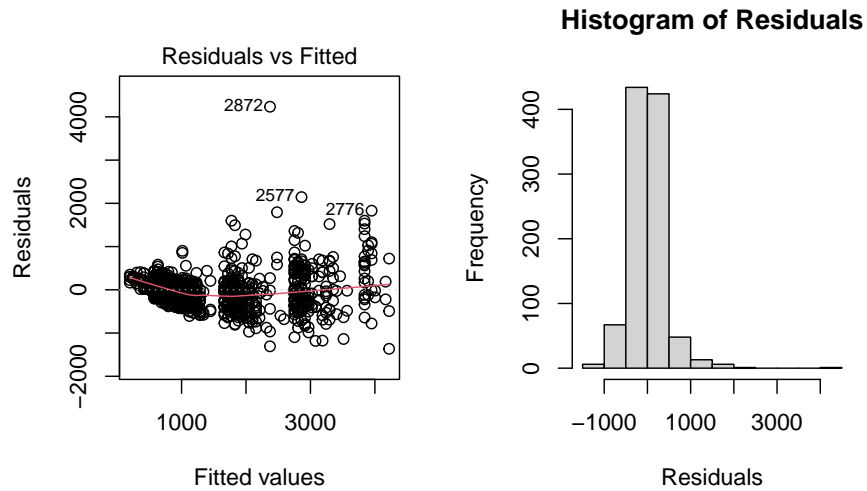   which should appear approximately normal

Example:

It is a generally accepted fact that the more carats a diamond has, the more expensive that diamond will be.
The question is, how much more expensive? Data on thousands of diamonds were collected for this data set. We
will only look at one type of cut ("Ideal") and diamonds less than 1 carat. Does the association between carat
size and price have a linear relationship for these types of diamonds? What can we state about the association
between carat size and price?

Scatterplot:

```
Diamonds %>% # Pipe data set into...
    ggplot(aes(x = carat, y = price))+  # Specify variables
    geom_point(alpha=0.5) +  # Add scatterplot of points
    labs(x = "carat",  # Label x-axis
      y = "price ($)",  # Label y-axis
      title = "Scatterplot of Diamonds Carats vs Price") +
            # Be sure to title your plots
    geom_smooth(method = "lm", se = FALSE)  # Add regression line
```


Scatterplot of Diamonds Carats vs Price
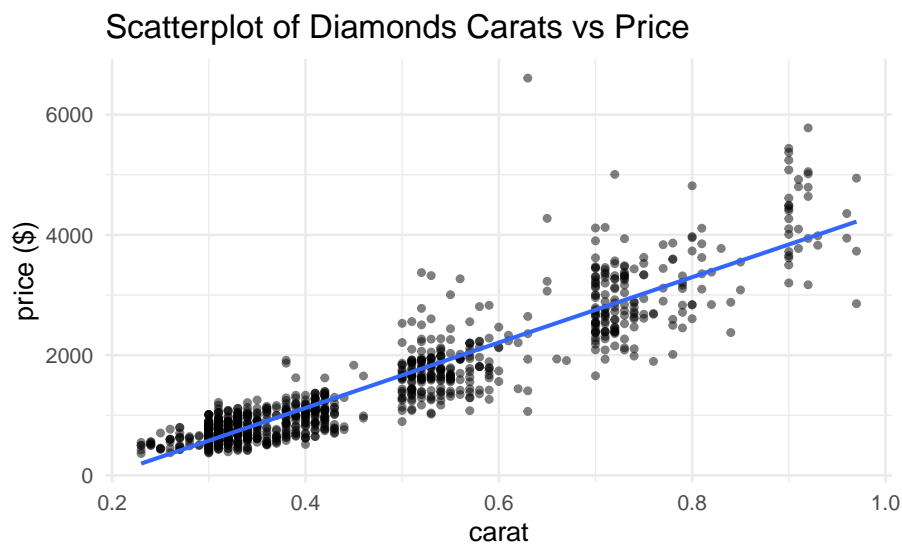
Diagnostic plots:

Check the conditions for the ocean data:

Scatterplot:

```
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "salinity (PSUs)",  # Label x-axis
       y = "temperature (C)",  # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
            # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```
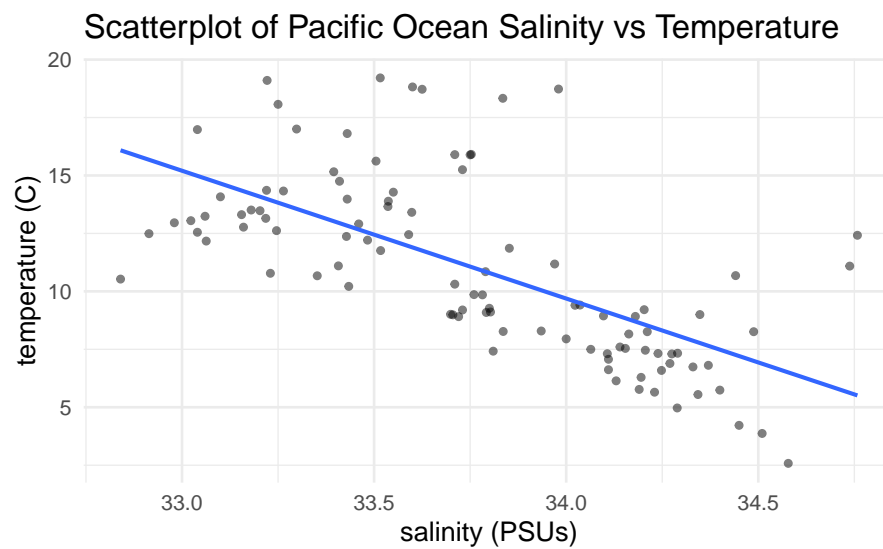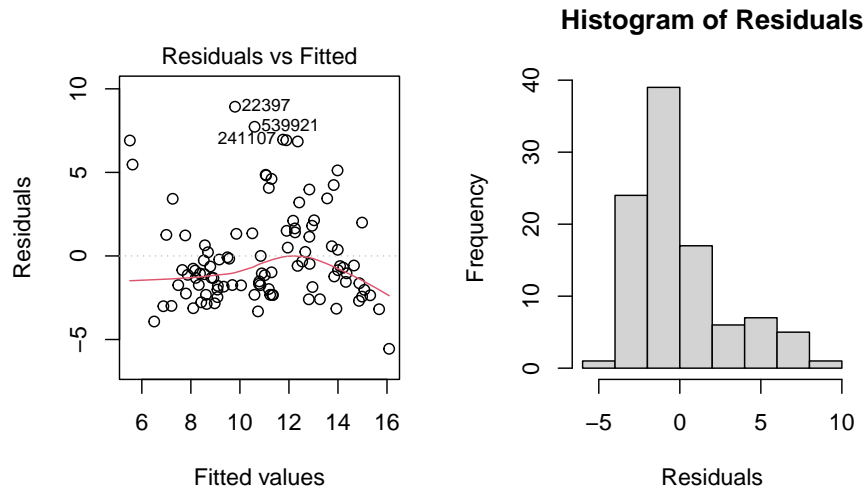
Diagnostic plots:

Like with paired data the t-distribution can be used to model slope and correlation.

- For two quantitative variables we use the _____-distribution with _____ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic
- Find the area under the t-distribution with $n-2$ df at least as extreme as the standardized statistic

Equation for the standardized slope:

Calculate the standardized slope for the ocean data

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients,3)
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18        0
#> Salnty        -5.514      0.636   -8.67        0
```

t–distribution with 98 df
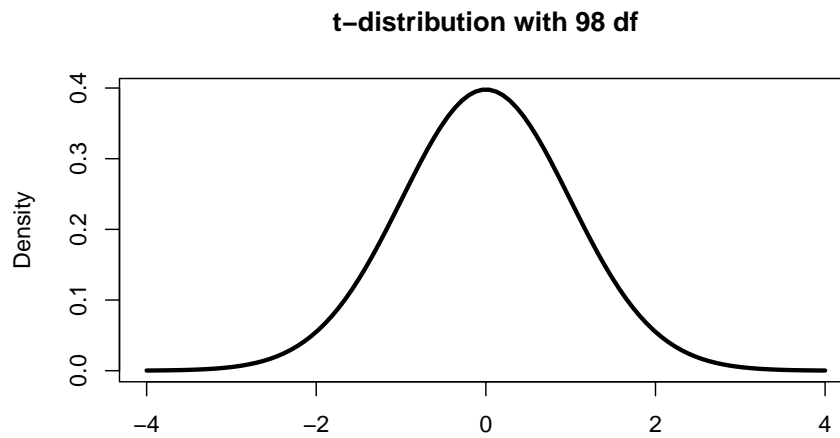
Interpret the standardized statistic:

To find the theory-based p-value:

```r
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients,3)

#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18        0
#> Salnty        -5.514      0.636   -8.67        0
```

or

```r
pt(-8.670, df = 98, lower.tail=TRUE)
#> [1] 4.623445e-14
```

Theory-based method

- Calculate the interval centered at the sample statistic

  statistic $\pm$ margin of error

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients, 3)
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18        0
#> Salnty        -5.514      0.636   -8.67        0
```

Using the ocean data, calculate a 95% confidence interval for the true slope.

- Need the t^* multiplier for a 95% confidence interval from a t-distribution with _____ df.

```
qt(0.975, df=98, lower.tail = TRUE)
```

```
#> [1] 1.984467
```

13.2.4   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. Explain why theory-based methods should not be used to analyze the salinity study?

2. What is the proper notation for the population slope? Population correlation?

13.3 Activity 26: Moneyball — Linear Regression

13.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.

- Use scatterplots to assess the relationship between two quantitative variables.

- Find the estimated line of regression using summary statistics and `R` linear model (`lm()`) output.

- Interpret the slope coefficient in context of the problem.

13.3.2 Terminology review

In today's activity, we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are:

- Scatterplot

- Least-squares line of regression

- Slope and *y*-intercept

- Residuals

To review these concepts, see Chapter 6 & 7 in the textbook.

13.3.3 Moneyball

The goal of a Major League baseball team is to make the playoffs. In 2002, the manager of the Oakland A's, Billy Bean, with the help of Paul DePodesta began to use statistics to determine which players to choose for their season. Based on past data, DePodesta determined that to make it to the playoffs, the A's would need to win at least 95 games in the regular season. In order to win more games, they would need to score more runs than they allowed. The Oakland A's won 20 consecutive games and a total of 103 games for the season. The success of this use of sports analytics was portrayed by the 2011 movie, Moneyball. In this study, we will see if there is evidence of a positive linear relationship between the difference in the number of runs scored minus the number of runs allowed (`RD`) and the number of wins for Major League baseball teams in the years before 2002. Some of the variables collected in the data set baseball consist of the following:

Variable	Description
RA	Runs allowed
RS	Runs scored
OBP	On-base percentage
SLG	Slugging percentage
BA	Batting average
OOBP	Opponent's on-base percentage
OSLG	Opponent's slugging percentage
W	Number of wins in the season
RD	Difference of runs scored minus runs allowed

```
moneyball <- read.csv("data/baseball.csv") # Reads in data set
moneyball$RD <- moneyball$RS - moneyball$RA
moneyball <-
    moneyball %>% # Pipe data set into
    subset(Year < 2002) # Select only years before 2002
```

Vocabulary review

- Use the provided R script file to create a scatterplot to examine the relationship between the difference in number of runs scored minus number of runs allowed and the number of wins by filling in the variable names (RD and W) for explanatory and response in line 14. Note, we are using the difference in runs scores minus runs allowed to predict the number of season wins.

- Highlight and run lines 1–20.

```
moneyball %>% # Data set pipes into...
    ggplot(aes(x = explanatory, y = response))+ # Specify variables
    geom_point() + # Add scatterplot of points
    labs(x = "Difference in number of runs", # Label x-axis
        y = "Number of Season wins", # Label y-axis
        title = "Scatterplot of Run Difference vs. Number of Season Wins for MLB Teams") +
# Be sure to tile your plots
geom_smooth(method = "lm", se = FALSE) # Add regression line
```

1. Assess the four features of the scatterplot that describe this relationship.

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

2. Based on the plot, does there appear to be an association between run difference and number of season wins? Explain your answer.

Slope

The linear model function in R (`lm()`) gives us the summary for the least squares regression line. The estimate for (`Intercept`) is the y-intercept for the line of least squares, and the estimate for `budget_mil` (the x-variable name) is the value of b_1, the slope.

- Run lines 24–25 in the R script file to reproduce the linear model output found in the coursepack.

```
# Fit linear model: y ~ x
moneyballLM <- lm(W~RD, data=moneyball)
round(summary(moneyballLM)$coefficients, 3) # Display coefficient summary
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   80.881      0.131 616.675        0
#> RD             0.106      0.001  81.554        0
```

3. Write out the least squares regression line using the summary statistics provided above in context of the problem.

4. Interpret the value of slope in context of the problem.

5. Using the least squares line from question 3, predict the number of season wins for a MLB team that has a run difference of -66 runs.

6. Predict the number of season wins for a MLB team that has a run difference of 400 runs.

7. The prediction in question 6 is an example of what?

Residuals

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the variability in the response that hasn't been modeled by the regression line.

$$\Longrightarrow \text{Residual} = \text{actual y value} - \text{predicted y value}$$

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

8. The MLB team *Florida Marlins* had a run difference of -66 runs and 79 wins for the season. Find the residual for this MLB team.

9. Did the line of regression overestimate or underestimate the number of wins for the season for this team?

Correlation

The following output shows a correlation matrix between several pairs of quantitative variables.

- Highlight and run lines 29–33 to produce the same table as below.

```
moveyball %>%  # Data set pipes into
  select(c("RD", "BA",
           "SLG", "W")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

```
#>        RD    BA   SLG     W
#> RD  1.000 0.442 0.428 0.939
#> BA  0.442 1.000 0.814 0.416
#> SLG 0.428 0.814 1.000 0.406
#> W   0.939 0.416 0.406 1.000
```

10. Report the value of correlation between the run difference and the number of season wins.

11. Calculate the coefficient of determination between the run difference and the number of season wins.

12. Interpret the value of coefficient of determination in context of the study.

13.3.4   Take-home messages

1. Two quantitative variables are graphically displayed in a scatterplot. The explanatory variable is on the x-axis and the response variable is on the y-axis. When describing the relationship between two quantitative variables we look at the form (linear or non-linear), direction (positive or negative), strength, and for the presence of outliers.

2. There are three summary statistics used to summarize the relationship between two quantitative variables: correlation (r), slope of the regression line (b_1), and the coefficient of determination (r^2).

13.3.5   Additional notes

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

13.4 Activity 27: IPEDS (continued)

13.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.

- Find the estimated line of regression using summary statistics and R linear model (`lm()`) output.

- Interpret the slope coefficient in context of the problem.

- Calculate and interpret r^2, the coefficient of determination, in context of the problem.

- Find the correlation coefficient from R output or from r^2 and the sign of the slope.

13.4.2 Terminology review

In today's activity, we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are:

- Least-squares line of regression

- Slope and y-intercept

- Residuals

- Correlation (r)

- Coefficient of determination (r-squared)

To review these concepts, see Chapter 6 in the textbook.

13.4.3 The Integrated Postsecondary Education Data System (IPEDS)

We will continue to assess the IPEDS data set collected on a subset of institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting

- United States only

- Title IV participating

- Not for profit

- 2-year or 4-year or above

- Has full-time first-time undergraduates

Some of the variables collected and their descriptions are below. Note that several variables have missing values for some institutions (denoted by "NA").

Variable	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Sector	whether public or private
LandGrant	Is this a land-grant institution (Yes/No)
Size	Institution size category based on total student enrolled for credit, Fall 2018: Under 1,000, 1,000 - 4,999, 5,000 - 9,999, 10,000 - 19,999, 20,000 and above
Cost_OutofState	Cost of attendance for full-time out-of-state undergraduate students

Variable	Description
Cost_InState	Cost of attendance for full-time in-state undergraduate students
Retention	Retention rate is the percent of the undergraduate students that re-enroll in the next year
Graduation_Rate	6-year graduation rate for undergraduate students
SATMath_75	75th percentile Math SAT score
ACT_75	75th percentile ACT score

The code below reads in the needed data set, IPEDS_2018.csv, and filters out the 2-year institutions.

- Highlight and run lines 1 – 11 to load the data set and filter out the 2-year institutions.

```r
IPEDS <- read.csv("https://www.math.montana.edu/courses/s216/data/IPEDS_2018.csv")
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") #Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") #Filters the data set to remove Private 2-year
IPEDS <- na.omit(IPEDS)
```

To create a scatterplot of the 75th percentile Math SAT score by retention rate for 4-year US Higher Education Institutions...

- Enter the variable `SATMath_75` for explanatory and `Retention` for response in line 16.

- Highlight and run lines 15–21.

```r
IPEDS %>% # Data sest pipes into...
    ggplot(aes(x = SATMath_75, y = Retention))+  # Specify variables
    geom_point(alpha=0.5) +  # Add scatterplot of points
    labs(x = "75th Percentile SAT Math Score",  # Label x-axis
        y = "Retention Rate (%)",  # Label y-axis
        title = "Scatterplot of SAT Math Score vs. Retention Rate for
        4-year US Higher Education Institutions") +
    # Be sure to title your plots with the type of plot, observational units, variable(s)
    geom_smooth(method = "lm", se = FALSE) + # Add regression line
    theme_bw()
```


Scatterplot of SAT Math Score vs. Retention Rate for 4–year US Higher Education Institutions
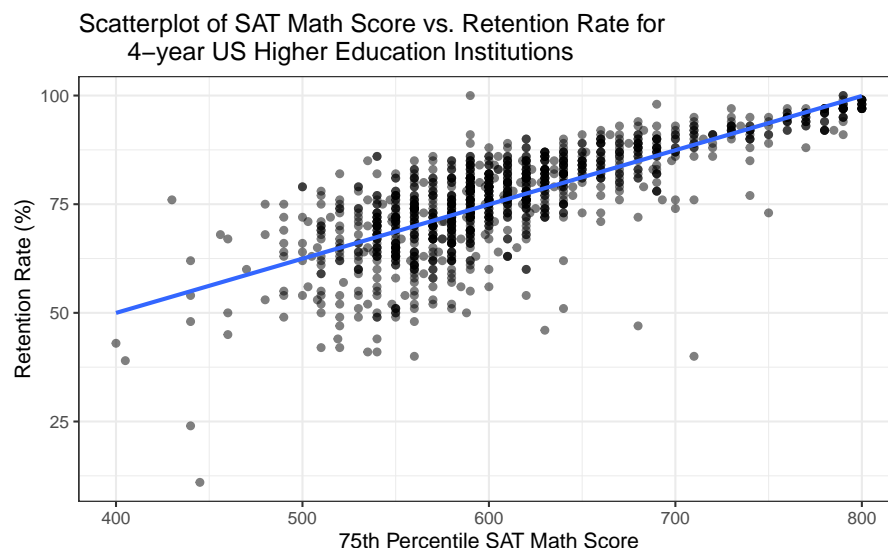
1. Describe the relationship between 75th percentile SAT Math score and retention rate.

Slope of the Least Squares Linear Regression Line

There are three summary measures calculated from two quantitative variables: slope, correlation, and the coefficient of determination. We will first assess the slope of the least squares regression line between 75th percentile SAT Math score and retention rate.

- Enter `Retention` for response and `SATMath_75` for explanatory in line 25

- Highlight and run lines 25–26 to fit the linear model.

```
# Fit linear model: y ~ x
IPEDSLM <- lm(Retention~SATMath_75, data=IPEDS)
round(summary(IPEDSLM)$coefficients,3) # Display coefficient summary
```

```
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    0.059      1.898   0.031    0.975
#> SATMath_75     0.125      0.003  40.485    0.000
```

2. Write out the least squares regression line using the summary statistics from the R output in context of the problem.

Slope Interpretation: An increase of one point in SAT Math 75th percentile score is associated with an increase in retention rate, on average, of 0.125 percentage points for 4-year higher education institutions.

3. Predict the retention rate for a 4-year US higher education institution with a 75th percentile SAT Math score of 440.

4. Calculate the residual for a 4-year US higher education institution with a 75th percentile SAT Math score of 440 and a retention rate of 24%.

Correlation

Correlation measures the strength and the direction of the linear relationship between two quantitative variables. The closer the value of correlation to +1 or −1, the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables.

The following output creates a correlation matrix between several pairs of quantitative variables.

```
IPEDS %>%  # Data set pipes into
  select(c("Retention", "Cost_InState",
           "Graduation_Rate", "Salary",
           "SATMath_75", "ACT_75")) %>%
```

```
cor(use="pairwise.complete.obs") %>%
  round(3)
```

```
#>                Retention Cost_InState Graduation_Rate Salary SATMath_75 ACT_75
#> Retention          1.000        0.388           0.832  0.698      0.767  0.768
#> Cost_InState       0.388        1.000           0.563  0.365      0.502  0.514
#> Graduation_Rate    0.832        0.563           1.000  0.683      0.817  0.833
#> Salary             0.698        0.365           0.683  1.000      0.747  0.706
#> SATMath_75         0.767        0.502           0.817  0.747      1.000  0.920
#> ACT_75             0.768        0.514           0.833  0.706      0.920  1.000
```

5. What is the value of correlation between SATMath_75 and Retention?

Coefficient of determination (squared correlation)

Another summary measure used to explain the linear relationship between two quantitative variables is the coefficient of determination (r^2). The coefficient of determination, r^2, can also be used to describe the strength of the linear relationship between two quantitative variables. The value of r^2 (a value between 0 and 1) represents the **proportion of variation in the response that is explained by the least squares line with the explanatory variable**. There are two ways to calculate the coefficient of determination:

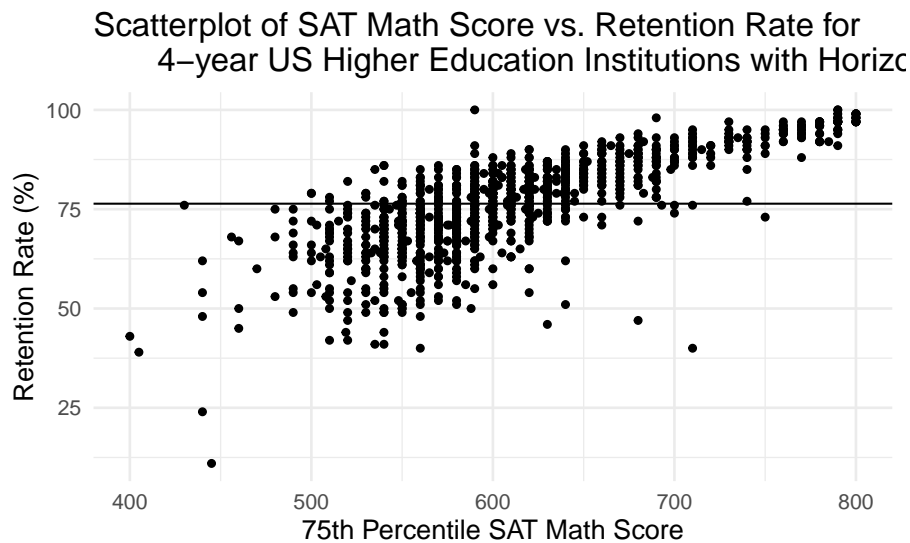Square the correlation coefficient: $r^2 = (r)^2$

Use the variances of the response and the residuals: $r^2 = \dfrac{s_y^2 - s_{RES}^2}{s_y^2} = \dfrac{SST - SSE}{SST}$

6. Use the correlation, r, found in question 5, to calculate the coefficient of determination between SATMath_75 and Retention, r^2.

7. The variance of the response variable, Retention in $MM, is $s_{Retention}^2 = 138.386 \ \%^2$ and the variability in the residuals is $s_{RES}^2 = 56.934 \ \%^2$. Use these values to calculate the coefficient of determination.

In the next part of the activity we will explore what the coefficient of determination measures.

In the first scatterplot, we see the data plotted with a horizontal line. Note that the regression line in this plot has a slope of zero; this assumes there is no relationship between SATMath_75 and Retention. The value of the y-intercept, 76.387, is the mean of the response variable when there is no relationship between the two variables. To find the sum of squares total (SST) we find the residual ($residual = y - \hat{y}$) for each response value from the horizontal line (from the value of 76.387). Each residual is squared and the sum of the squared values is calculated. The SST gives the **total variability in the response variable, Retention**.

Scatterplot of SAT Math Score vs. Retention Rate for
4–year US Higher Education Institutions with Horizo

The calculated value for the SST is 158451.8.

This next scatterplot, shows the plotted data with the best fit regression line. This is the line of best fit between budget and revenue and has the smallest sum of squares error (SSE). The SSE is calculated by finding the residual from each response value to the regression line. Each residual is squared and the sum of the squared values is calculated.

Scatterplot of SAT Math Score vs. Retention Rate for
4–year US Higher Education Institutions
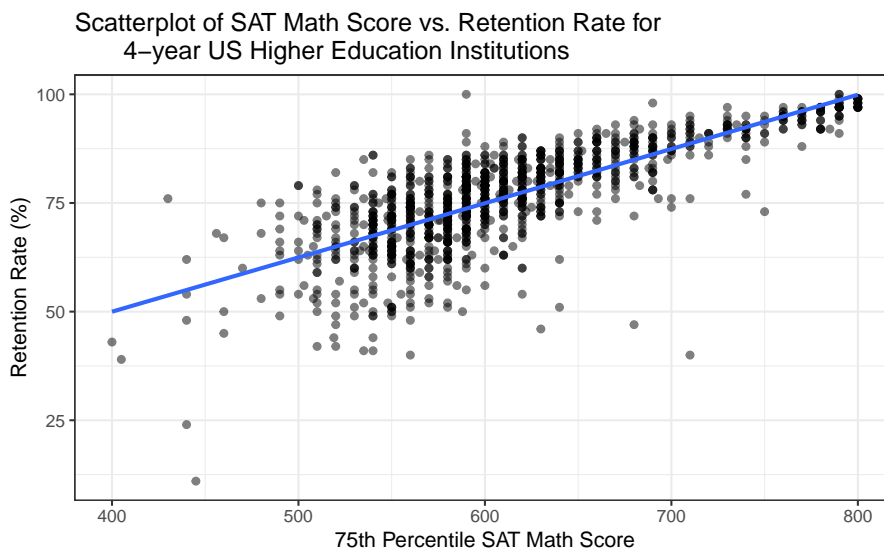
The calculated value for the SSE is 65133.022.

8. Calculate the value for r^2 using the values for SST and SSE provided below each of the previous graphs.

9. Write a sentence interpreting the coefficient of determination in context of the problem.

Multivariable plots

When adding another categorical predictor, we can add that variable as shape or color to the plot. In the following code we have added the variable `Sector`, whether the 4-year institution is public or private.

```
IPEDS %>% # Data sest pipes into...
    ggplot(aes(x = SATMath_75, y = Retention, shape = Sector, color=Sector))+  # Specify variables
    geom_point(alpha=0.5) +  # Add scatterplot of points
    labs(x = "75th Percentile SAT Math Score",  # Label x-axis
        y = "Retention Rate (%)",  # Label y-axis
        title = "Scatterplot of SAT Math Score vs. Retention Rate for
        4-year US Higher Education Institutions") +
    # Be sure to title your plots with the type of plot, observational units, variable(s)
    geom_smooth(method = "lm", se = FALSE) + # Add regression line
    scale_color_grey()
```
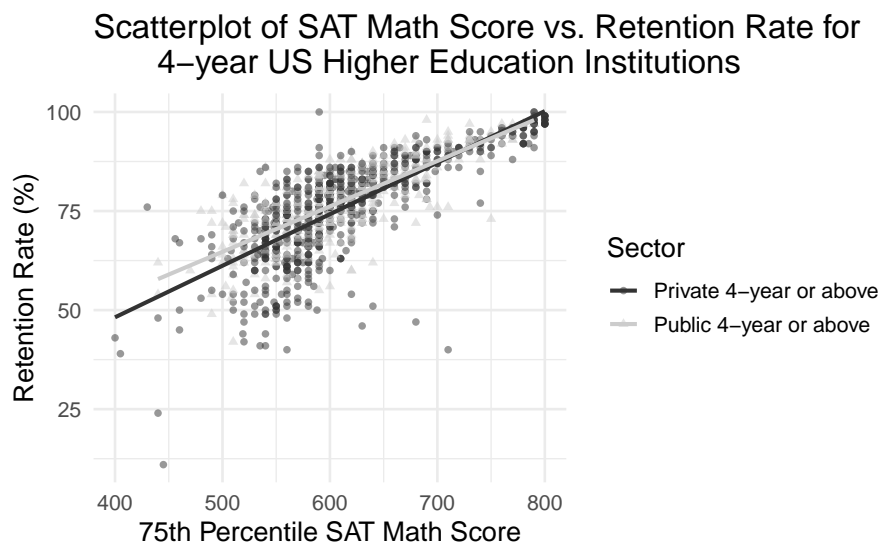

10. Does the relationship between 75th percentile SAT math score and retention rate of 4-year institutions change depending on the level of sector?

13.4.4   Take-home messages

1. The sign of correlation and the sign of the slope will always be the same. The closer the value of correlation is to -1 or $+1$, the stronger the linear relationship between the explanatory and the response variable.

2. The coefficient of determination multiplied by 100 ($r^2 \times 100$) measures the percent of variation in the response variable that is explained by the relationship with the explanatory variable. The closer the value of the coefficient of determination is to 100%, the stronger the relationship.

3. We can use the line of regression to predict values of the response variable for values of the explanatory variable. Do not use values of the explanatory variable that are outside of the range of values in the data set to predict values of the response variable (reflect on why this is true.). This is called **extrapolation**.

13.4.5   Additional notes

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

13.5 Activity 28: Prediction of Crocodilian Body Size

13.5.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for slope or correlation.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a slope or correlation.

- Use bootstrapping to find a confidence interval for the slope or correlation.

- Interpret a confidence interval for a slope or correlation.

13.5.2 Terminology review

In today's activity, we will use simulation-based methods for hypothesis tests and confidence intervals for a linear regression slope or correlation. Some terms covered in this activity are:

- Correlation

- Slope

- Regression line

To review these concepts, see Chapter 21 in the textbook.

13.5.3 Crocodilian Body Size

Much research surrounds using measurements of animals to estimate body-size of extinct animals. Many challenges exist in making accurate estimates for extinct crocodilians. The term crocodilians refers to all members of the family Crocodylidae ("true" crocodiles), family Alligatoridae (alligators and caimans) and family Gavialidae (gharial, Tomistoma). The researchers in this study (O'Brien 2019) state, "Among extinct crocodilians and their precursors (e.g., suchians), several methods have been developed to predict body size from suites of hard-tissue proxies. Nevertheless, many have limited applications due to the disparity of some major suchian groups and biases in the fossil record. Here, we test the utility of head width (HW) as a broadly applicable body-size estimator in living and fossil suchians." Data were collected on 76 male and female individuals of different species. Is there evidence that head width (measured in cm) is a good predictor of total body length (measured in cm) for crocodilians?

- Download the R script file from D2L and upload to the RStudio server

- Open the file and run lines 1 - 8 to load the dataset

```
# Read in data set
croc <- read.csv("https://math.montana.edu/courses/s216/data/Crocodylian_headwidth.csv")
croc <- croc %>%
    na.omit()
```

To create a scatterplot to examine the relationship between head width and total body length we will use `HW_cm` as the explanatory variable and `TL_cm` as the response variable.

- Enter the name of the explanatory and response variable in line 14

- Highlight and run lines 13 - 20

```
croc %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "head width (cm)",  # Label x-axis
       y = "total length (cm)",  # Label y-axis
       title = "Scatterplot of Crocodilian Head Width vs. Total Length") +
    # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

1. Describe the features of the plot, addressing all four characteristics of a scatterplot.

       If you indicated there are potential outliers, which points are they?

Hypotheses

When analyzing two quantitative variables we can either test regression slope or correlation. In both cases, we are testing that there is a linear relationship between variables.

2. Write the null hypothesis in words.

3. Write the null hypothesis to test slope in notation.

4. Write the null hypothesis to test correlation in notation.

5. Write the alternative hypothesis in words.

Summarize and visualize the data

To create the linear model output and find the value of correlation for the linear relationship...

- Enter the the name of the explanatory and response in line 25

- Highlight and run lines 25 - 27

```
#Linear model
lm.croc <- lm(response~explanatory, data=croc) #lm(response~explanatory)
round(summary(lm.croc)$coefficients, 5)
#Correlation
cor(croc$HW_cm, croc$TL_cm)
```

6. Using the output from the evaluated R code, write the equation of the regression line in the context of the problem using appropriate statistical notation.

7. Interpret the estimated slope in context of the problem.

Use statistical inferential methods to draw inferences from the data

In this activity, we will focus on using simulation-based methods for inference in regression.

Simulation-based hypothesis test

Let's start by thinking about how one simulation would be created on the null distribution using cards. First, we would write the values for the response variable, total length, on each card. Next, we would shuffle these y values while keeping the x values (explanatory variable) in the same order. Then, find the line of regression for the shuffled (x, y) pairs and calculate either the slope or correlation of the shuffled sample.

We will use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of shuffled slopes (or shuffled correlations) and compute a p-value. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `croc`), the summary measure for the test (either slope or correlation), number of repetitions, the sample statistic (value of slope or correlation), and the direction of the alternative hypothesis.

The response variable name is `TL_cm` and the explanatory variable name is `HW_cm` for these data.

8. What inputs should be entered for each of the following to create the simulation to test regression slope?

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

- Summary measure (choose `"slope"` or `"correlation"`):

- As extreme as (enter the value for the sample slope):

- Number of repetitions:

Using the R script file for this activity...

- Enter your answers for question 8 in place of the **xx**'s to produce the null distribution with 10000 simulations.

- Highlight and run lines 32–37.

```
regression_test(TL_cm~HW_cm, # response ~ explanatory
                data = croc, # Name of data set
                direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "xx", # "slope" or "correlation"
                as_extreme_as = xx, # Observed slope or correlation
                number_repetitions = 10000) # Number of simulated samples for null distribution
```

9. Report the p-value from the R output.

10. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which two inputs in #8 would need to be changed to test for correlation? What inputs should you use instead?

11. Change the inputs in lines 32–37 to test for correlation instead of slope. Highlight and run those lines, then report the new p-value of the test.

12. The p-values from the test of slope (#9) and the test of correlation (#11) should be similar. Explain why the two p-values should match. *Hint: think about the relationship between slope and correlation!*

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample slopes (or sample correlations) and calculate a confidence interval.

- Fill in the missing values in the provided R script file to find a 95% confidence interval for slope.

- Highlight and run lines 42–46.

```
regression_bootstrap_CI(response~explanatory, # response ~ explanatory
  data = croc, # Name of data set
  confidence_level = xx, # Confidence level as decimal
```

```
    summary_measure = "xx", # Slope or correlation
    number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

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

14. Interpret the interval in question 13 in context of the problem. *Hint: use the interpretation of slope in your confidence interval interpretation.*

Communicate the results and answer the research question

15. Based on the p-value and confidence interval, write a conclusion in context of the problem.

13.5.4 Take-home messages

1. The p-value for a test for correlation should be approximately the same as the p-value for the test of slope. In the simulation test, we just change the statistic type from slope to correlation and use the appropriate sample statistic value.

2. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.

3. To create one simulated sample on the null distribution when testing for a relationship between two quantitative variables, hold the x values constant and shuffle the y values to new x values. Find the regression line for the shuffled data and plot the slope or the correlation for the shuffled data.

4. To create one simulated sample on the bootstrap distribution when assessing two quantitative variables, label n cards with the original (response, explanatory) values. Randomly draw with replacement n times. Find the regression line for the resampled data and plot the resampled slope or correlation.

13.5.5 Additional notes

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

13.6 Activity 29: Golf Driving Distance

13.6.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Assess the conditions to use the normal distribution model for a slope.

- Find the T test statistic (T-score) for a slope based off of `lm()` output in R.

- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a slope.

- Create and interpret a theory-based confidence interval for a slope.

- Use a confidence interval to determine the conclusion of a hypothesis test.

13.6.2 Terminology review

In this week's in-class activity, we will use theory-based methods for hypothesis tests and confidence intervals for a linear regression slope. Some terms covered in this activity are:

- Slope

- Regression line

To review these concepts, see Chapter 21 in the textbook.

13.6.3 Golf driving distance

In golf the goal is to complete a hole with as few strokes as possible. A long driving distance to start a hole can help minimize the strokes necessary to complete the hole, as long as that drive stays on the fairway. Data were collected on 354 PGA and LPGA players in 2008 ("Average Driving Distance and Fairway Accuracy" 2008). For each player, the average driving distance (yards), fairway accuracy (percentage), and sex was measured. Use these data to assess, "Does a professional golfer give up accuracy when they hit the ball farther?"

- Download the R script file from D2L and open in the RStudio server

```
# Read in data set
golf <- read.csv("https://math.montana.edu/courses/s216/data/golf.csv")
```

Plot review.

To create a scatterplot showing the relationship between the driving distance and percent accuracy for professional golfers:

- Enter the name of the explanatory and response in line 10

- Highlight and run lines 1 - 16

```
golf %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "Driving Distance (yards)",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       title = "Scatterplot of Driving Distance by Percent Accuracy
       for Professional Golfers") +
               # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

Conditions for the least squares line

When performing inference on a least squares line, the follow conditions are generally required:

- *Independent observations* (for both simulation-based and theory-based methods): individual data points must be independent.
  - Check this assumption by investigating the sampling method and determining if the observational units are related in any way.
- *Linearity* (for both simulation-based and theory-based methods): the data should follow a linear trend.
  - Check this assumption by examining the scatterplot of the two variables, and a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The pattern in the residual plot should display a horizontal line.
- *Constant variability* (for theory-based methods only): the variability of points around the least squares line remains roughly constant
  - Check this assumption by examining a scatterplot of the residuals (on the y-axis) versus the fitted values (on the x-axis). The variability in the residuals around zero should be approximately the same for all fitted values.
- *Nearly normal residuals* (for theory-based methods only: residuals must be nearly normal.
  - Check this assumption by examining a histogram of the residuals, which should appear approximately normal.

The scatterplot generated earlier and the residual plots shown below will be used to assess these conditions for approximating the data with the t-distribution.
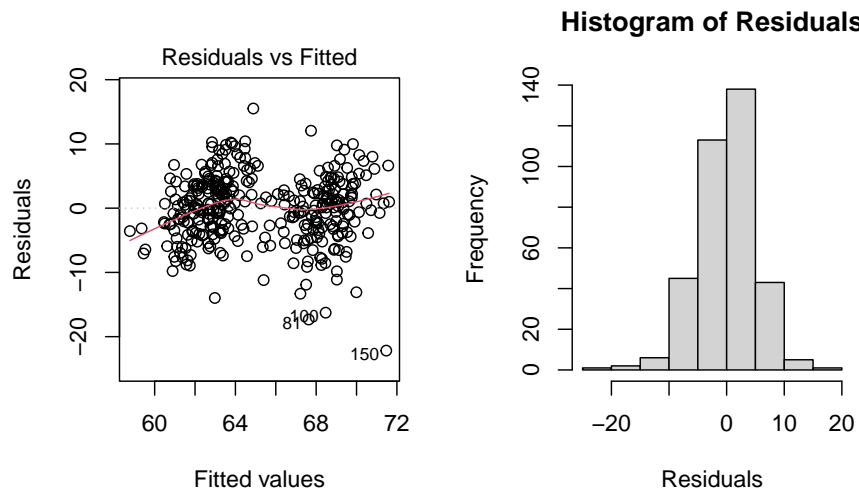
1. Are the conditions met to use the t-distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.

Ask a research question

2. Write out the null hypothesis in words to test the slope.

3. Using the research question, write the alternative hypothesis in notation to test the slope.

Summarize and visualize the data

The linear model output for this study is shown below.

```
lm.golf <- lm(Percent_Accuracy~Driving_Distance, data=golf) # lm(response~explanatory)
round(summary(lm.golf)$coefficients, 3)
```

```
#>                  Estimate Std. Error t value Pr(>|t|)
#> (Intercept)       103.586      3.329  31.119        0
#> Driving_Distance   -0.142      0.012 -11.553        0
```

4. Report the summary statistic (sample slope) for the linear relationship between driving distance and percent accuracy of golfers. Use proper notation.

Use statistical inferential methods to draw inferences from the data

Hypothesis test   To find the value of the standardized statistic to test the slope we will use,

$$T = \frac{\text{slope estimate} - nullvalue}{SE} = \frac{b_1 - 0}{SE(b_1)}.$$

We will use the linear model R output above to get the estimate for slope and the standard error of the slope.

5. Calculate the standardized statistic for slope. Identify where this calculated value is in the linear model R output.

6. The p-value in the linear model R output is the two-sided p-value for the test of significance for slope. Report the p-value to answer the research question.

7. Based on the p-value, how much evidence is there against the null hypothesis?

Confidence interval Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^* \times SE(\text{estimate}).$$

When the point estimate is a regression slope, this formula becomes

$$b_1 \pm t^* \times SE(b_1).$$

The t^* multiplier comes from a t-distribution with $n-2$ degrees of freedom. The sample size for this study is 354 so we will use the degrees of freedom 352 $(n-2)$.

- Enter the percentile needed to find the multiplier for a 95% confidence interval for xx

- Enter the degrees of freedom for yy

- Highlight and run line 35

```
qt(xx, yy, lower.tail = TRUE) # 95% t* multiplier
```

8. Calculate the 95% confidence interval for the true slope.

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

Communicate the results and answer the research question

10. Write a conclusion to answer the research question in context of the problem.

Multivariable plots

Another variable that may affect the percent accuracy is the which league the golfer is part of. We will look at how this variable may change the relationship between driving distance and percent accuracy.

```
golf %>%
  ggplot(aes(x = Driving_Distance, y = Percent_Accuracy, color=League))+  # Specify variables
  geom_point(aes(shape = League), size = 2, alpha=0.5) +  # Add scatterplot of points
  labs(x = "Driving Distance (yards)",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       color = "League", shape = "League",
       title = "Scatterplot of Golf Driving Distance and Percent
       Accuracy by League for Professional Golfers") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
    scale_color_grey()
```

Scatterplot of Golf Driving Distance and Percent Accuracy by League for Professional Golfers
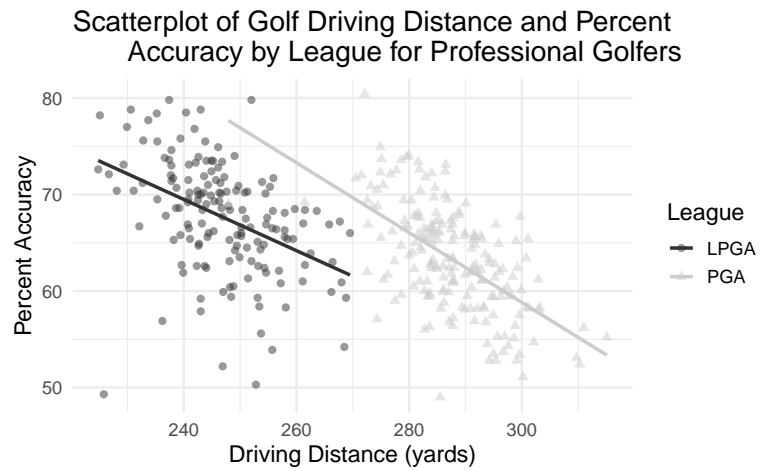
11. Does the association between driving distance and percent accuracy change depending on which league the golfer is a part of? Explain your answer.

12. Explain the association between league and each of the other two variables. Use the following plots in addition to the scatterplot from Q11 to explain your answer.

Side–by–side Box Plot of Percent Accuracy by League for Professional Golfers

Side–by–side Box Plot of Driving Distance by League for Professional Golfers
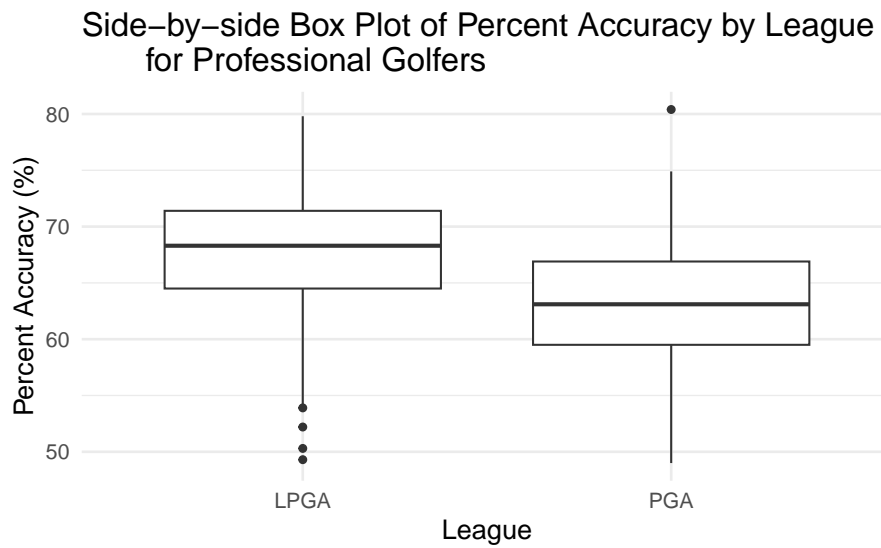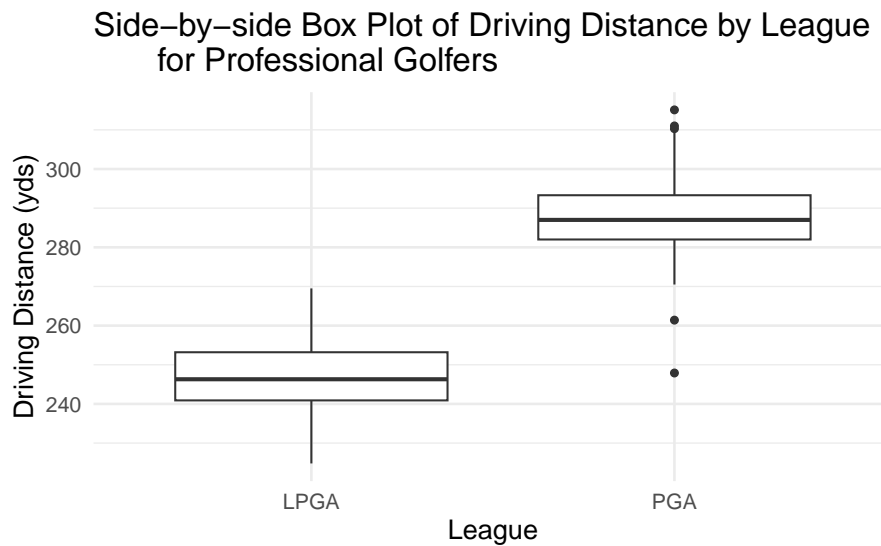
13.6.4   Take-home messages

1. To check the validity conditions for using theory-based methods we must use the residual diagnostic plots to check for normality of residuals and constant variability, and the scatterplot to check for linearity.

2. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.

3. Use the explanatory variable row in the linear model R output to obtain the slope estimate (`estimate` column) and standard error of the slope (`Std. Error` column) to calculate the standardized slope, or T-score. The calculated T-score should match the `t value` column in the explanatory variable row. The standardized slope tells the number of standard errors the observed slope is above or below 0.

4. The explanatory variable row in the linear model R output provides a **two-sided** p-value under the `Pr(>|t|)` column.

5. The standardized slope is compared to a t-distribution with $n-2$ degrees of freedom in order to obtain a p-value. The t-distribution with $n-2$ degrees of freedom is also used to find the appropriate multiplier for a given confidence level.

13.6.5   Additional notes

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

13.7 Module 13 Lab: Big Mac Index

13.7.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Assess the conditions to determine in theory or simulation-based methods should be used.

- Find, interpret, and evaluate the p-value for a hypothesis test for a slope or correlation.

- Create and interpret a confidence interval for a slope or correlation.

13.7.2 Big Mac Index

Can the relative cost of a Big Mac across different countries be used to predict the Gross Domestic Product (GDP) per person for that country? The log GDP per person and the adjusted dollar equivalent to purchase a Big Mac was found on a random sample of 55 countries in January of 2022. The cost of a Big Mac in each country was adjusted to US dollars based on current exchange rates. Is there evidence of a positive relationship between Big Mac cost (`dollar_price`) and the log GDP per person (`log_GDP`)?

- Upload and open the R script file for Week 13 lab.

- Upload the csv file, `big_mac_adjusted_index_S22.csv`.

- Enter the name of the data set for datasetname in the R script file in line 9.

- Highlight and run lines 1–9 to load the data.

```
# Read in data set
mac <- read.csv("datasetname")
```

Summarize and visualize the data

- To find the correlation between the variables, `log_GDP` and `dollar_price` highlight and run lines 13–16 in the R script file.

```
mac %>%
  select(c("log_GDP", "dollar_price")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

1. Report the value of correlation between the variables.

2. **Calculate the value of the coefficient of determination between `log_GDP` and `dollar_price`.**

3. Interpret the value of the coefficient of determination in context of the problem.

In the next part of the activity we will assess the linear model between Big Mac cost and log GDP.

- Enter the variable `log_GDP` for `response` and the variable `dollar_price` for `explanatory` in line 22.

- Highlight and run lines 22–23 to get the linear model output.

```
# Fit linear model: y ~ x
bigmacLM <- lm(response~explanatory, data=mac)
round(summary(bigmacLM)$coefficients,3) # Display coefficient summary
```

4. Give the value of the slope of the regression line. Interpret this value in context of the problem.

Conditions for the least squares line

5. Is there independence between the responses for the observational units? Justify your answer.

- Highlight and run lines 28–33 to create the scatterplot to check for linearity.

```
#Scatterplot
mac %>% # Pipe data set into...
  ggplot(aes(x = dollar_price, y = log_GDP))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "Big Mac Cost",  # Label x-axis
       y = "log GDP",  # Label y-axis
       title = "Scatterplot of Big Mac Cost vs. log GDP per person
       for Countries in 2022") +  # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

6. Is the linearity condition met to use regression methods to analyze the data? Justify your answer.

- Highlight and run lines 38–42 to produce the diagnostic plots needed to assess conditions to use theory-based methods.

```
#Diagnostic plots
bigmacLM <- lm(log_GDP~dollar_price, data = mac) # Fit linear regression model
par(mfrow=c(1,2)) # Set graphics parameters to plot 2 plots in 1 row
plot(bigmacLM, which=1) # Residual vs fitted values
hist(bigmacLM$resid, xlab="Residuals", ylab="Frequency",
     main = "Histogram of Residuals") # Histogram of residuals
```

7. **Are the conditions met to use the t-distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.**

Ask a research question

8. Write out the null and alternative hypotheses in notation to test *correlation* between Big Mac cost and log GDP.

H_0 :

H_A :

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of sample **correlations** and compute a p-value. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `mac`), the summary measure used for the test, number of repetitions, the sample statistic (value of correlation), and the direction of the alternative hypothesis.

The response variable name is `log_GDP` and the explanatory variable name is `dollar_price`.

9. What inputs should be entered for each of the following to create the simulation to test correlation?

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

- Summary measure (choose `"slope"` or `"correlation"`):

- As extreme as (enter the value for the sample correlation):

- Number of repetitions:

Using the R script file for this activity, enter your answers for question 9 in place of the **xx**'s to produce the null distribution with 10000 simulations.

- Highlight and run lines 47–53.

- **Upload a copy of your plot showing the p-value to Gradescope for your group.**

```
regression_test(log_GDP~dollar_price, # response ~ explanatory
                data = mac, # Name of data set
                direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure  = "xx", # "slope" or "correlation"
                as_extreme_as = xx, # Observed slope or correlation
                number_repetitions = 10000) # Number of simulated samples for null distribution
```

10. Report the p-value from the R output.

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample **correlations** and calculate a confidence interval.

- Fill in the **xx**'s in the the provided R script file to find a 90% confidence interval.

- Highlight and run lines 58–62.

```r
regression_bootstrap_CI(log_GDP~dollar_price, # response ~ explanatory
    data = mac, # Name of data set
    confidence_level = xx, # Confidence level as decimal
    summary_measure = "xx", # Slope or correlation
    number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

11. Report the bootstrap 90% confidence interval in interval notation.

Communicate the results and answer the research question

12. Using a significance level of 0.1, what decision would you make?

13. What type of error is possible?

14. Interpret this error in context of the problem.

15. Write a paragraph summarizing the results of the study as if you are reporting these results in your local newspaper. **Upload a copy of your paragraph to Gradescope for your group.** 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)?

MODULE 14

Unit 3 Review

The following section contains both a list of key topics covered in Unit 3 as well as Module Review Worksheets.

14.0.1 Key Topics

Review the key topics for Unit 3 to review prior to the exams. All of these topics will be covered in Modules 11–13.

14.0.2 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 3 Review folder after the study sessions.

14.1 Key Topics Exam 3

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 the study design (observational study or randomized experiment).

6. Identify the sampling method and potential types of sampling bias (non-response, response, selection).

7. Determine the scope of inference (causation/association and generalizability) of the study.

8. Calculate and interpret the mean difference from paired data.

9. Calculate and interpret the difference in means from independent data.

10. Identify the slope of the repression line from R output and interpret.

11. Identify the correlation from R output and describe the strength and direction.

12. Calculate and interpret coefficient of determination.

Hypothesis testing

13. Identify which of the three scenarios applies to the study: paired data, independent groups, or two quantitative variables.

14. Write the parameter of interest in words and correct notation.

15. Find the value of the observed statistic (point estimate, summary statistic). Use correct notation.

16. State the null and alternative hypotheses in words and in correct notation.

17. Verify the validity condition is met to use simulation-based methods to find a p-value.

18. Verify the validity conditions are met to use theory-based methods to find a p-value from the theoretical distribution.

19. In a simulation-based hypothesis test, describe how to create one dot on a dotplot of the null distribution using cards.

20. Explain where the null distribution is centered and why.

21. Describe and illustrate how R calculates the p-value for a simulation-based test.

22. Describe and illustrate how R calculates the p-value for a theory-based test.

23. Type of theoretical distribution (t-distribution and appropriate degrees of freedom) used to model the standardized statistic in a theory-based hypothesis test.

24. Calculate and interpret the standard error of the statistic using the correct formula on the Golden ticket.

25. Calculate and interpret the appropriate standardized statistic using the correct formula on the Golden ticket.

26. Interpret the p-value in context of the study: it is the probability of _____, assuming _____.

27. Evaluate the p-value for strength of evidence against the null: how much evidence does the p-value provide against the null?

28. Write a conclusion about the research question based on the p-value.

29. Given a significance level, what decision can be made about the research question based on the p-value.

Confidence interval

30. Describe how to simulate one bootstrapped sample using cards.

31. Explain where the bootstrap distribution is centered and why.

32. Find an appropriate percentile confidence interval using a bootstrap distribution from R output.

33. Verify the validity condition is met to use simulation-based methods to find the confidence interval.

34. Verify the validity conditions are met to use theory-based methods to calculate a confidence interval.

35. Describe and illustrate how the bootstrap distribution is used to find the confidence interval for a given confidence level.

36. Describe and illustrate how the t-distribution is used to find the multiplier for a given confidence level.

37. Calculate the appropriate margin of error and confidence interval using theory-based methods.

38. Interpret the confidence interval in context of the study.

39. 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.

40. Interpret the confidence level in context of the study. What does "confidence" mean?

41. Describe which features of the study have an effect on the width of the confidence interval and how.

14.2 Module 11 Review - Paired Data

Students in an introductory statistics class were asked to participate in an experiment to answer this question. Each student flipped a coin to determine which exercise to complete first. If the coin landed on heads the student would do jumping jacks for 30 seconds and then measure their heart rate in beats per minute (bpm). After a 2 minute break the student would do bicycle kicks for 30 seconds and then record their heart rate. If the coin landed on tails the student would complete bicycle kicks first followed by jumping jacks using the same times as above. For this study we will use the order of subtraction jumping jacks – bicycle kicks. Which exercise, jumping jacks or bicycle kicks will raise your heart rate more?

```
#>   min Q1 median Q3 max     mean       sd  n missing
#> 1 -16 -4      8 15  57 7.604651 15.91666 43       0
```

The following code created the boxplot of differences.


Boxplot of the Differences in Heartrates
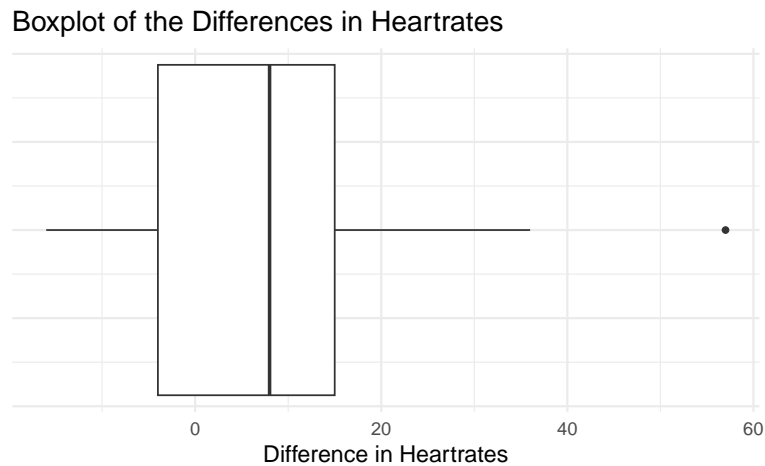
1. What is the study design (observational or randomized experiment)?

2. Is this paired study or two independent samples?

3. What are the roles and types of each variable?

4. What is the scope of inference for this study?

5. Write the parameter of interest for this study.

6. Write the null hypothesis in notation.

7. Write the alternative hypothesis in words.

We will start with simulation methods.

8. Calculate the difference $\mu_0 - \bar{x}_d$. Will we need to shift the data up or down?

```
set.seed(216)
paired_test(data = heartrate$Diff,    #Vector of differences or data set with column for each group
            shift = -6.429,   #Shift needed for bootstrap hypothesis test
            as_extreme_as = 6.429,   #Observed statistic
            direction = "two-sided",   #Direction of alternative
            number_repetitions = 10000,   #Number of simulated samples for null distribution
            which_first = 1)   #Not needed when using calculated differences
```


Mean = 1.135
SD = 2.387

Simulated Mean Difference
Count = 158/10000 = 0.0158
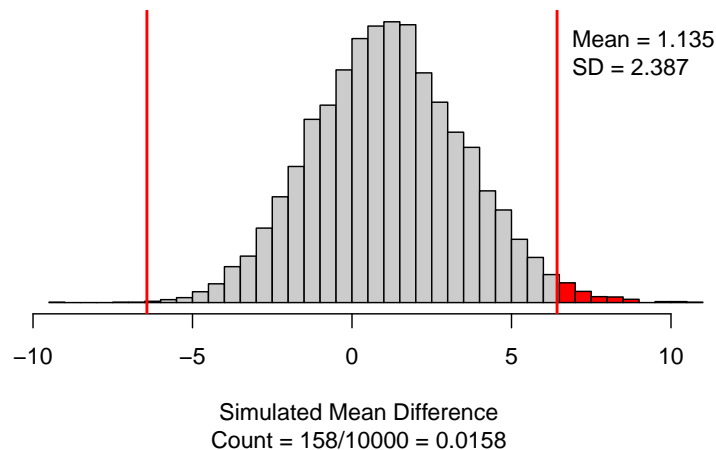
9. Based on the p-value for this study, which of the following are true?

- There is very strong evidence that there is a true difference in heart rates for students who did jumping jacks and bicycle kicks (jumping jacks – bicycle kicks), on average.

- If there is no true mean difference in heart rates for students who did jumping jacks and bicycle kicks, in 1 out of 10000 simulated samples, we would observe a sample mean difference in heart rates of 6.429 bpm or more extreme.

- The 95% confidence interval would be entirely positive.

346

- There could be a potential Type I error.

- We would conclude that there is evidence of a difference in heart rates between exercises, on average, when in fact there is not.

Bootstrap CI simulation to create a 95% confidence interval

```
paired_bootstrap_CI(data = heartrate$Diff, #Enter vector of differences
                    number_repetitions = 10000, #Number of bootstrap samples for CI
                    confidence_level = 0.95,  #Confidence level in decimal form
                    which_first = 1)  #Not needed when entering vector of differences
```


Bootstrap Mean Difference
95% CI: (3.07, 12.488)
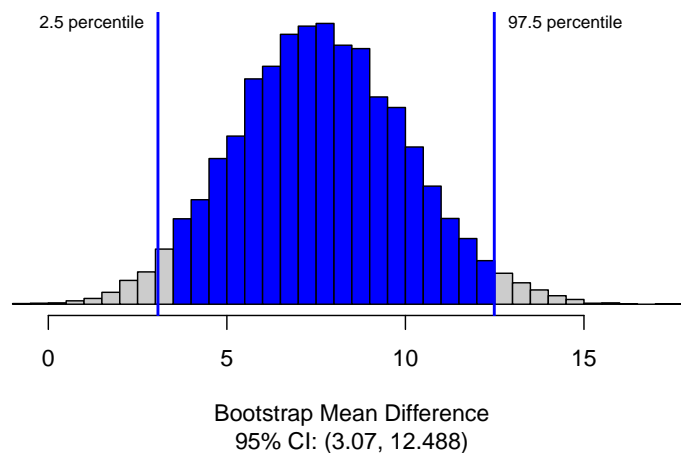
10. Interpret the 95% confidence interval in context of the study.

11. Interpret the confidence level in context of the study. What does confidence mean?

Next we will use theory-based methods.

The sampling distribution for $\bar{x}$ based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a normal distribution when certain conditions are met.

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

347

- **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!)

- **Normality Condition**: either the sample differences come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:

  - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the distribution of differences, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

  - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers in the differences of differences, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal to satisfy the condition

  - $n \geq 100$: If the sample size is greater than 100 then we typically assume the sampling distribution of $\bar{x}$ is nearly normal to satisfy the condition, even if the underlying distribution of individual observations is not.

12. Are the conditions met to model the data with theory-based methods?

To find the standardized statistic for the paired differences we will use the following formula:

$$T = \frac{\bar{x}_d - \text{null value}}{SE(\bar{x}_d)},$$
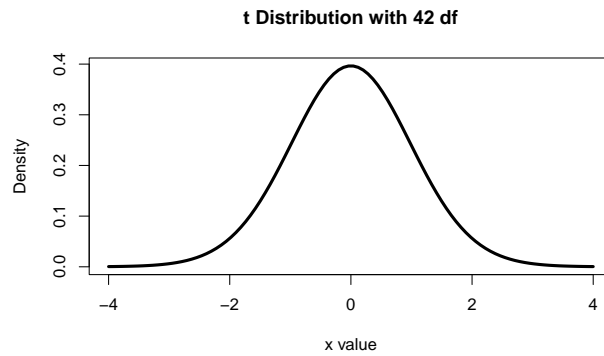
where the standard error of the sample mean difference is:

$$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}.$$

13. Calculate the standard error of the mean difference.

14. Calculate the standardized mean difference.

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

t Distribution with 42 df


x value

P-value for the test:

```
2*pt(2.957, df=41, lower.tail=FALSE)
#> [1] 0.005134632
```

To calculate the 95% theory-based confidence interval for the paired mean difference, use the following formula:

$$\bar{x}_d \pm t^* SE(\bar{x}_d).$$

We will need to find the t^* multiplier using the function `qt()`. For a 95% confidence level, we are finding the t^* value at the 97.5th percentile with $\mathrm{df} = n_d - 1 = 42 - 1 = 41$.

```
qt(0.975, df = 42, lower.tail=TRUE)
#> [1] 2.018082
```

t Distribution with 42 df

x value
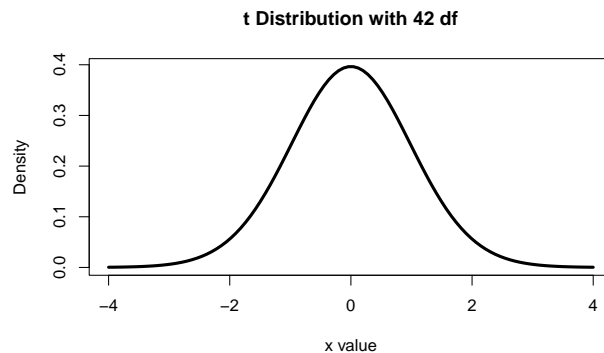
16. Calculate the 95% confidence interval.

17. Write a conclusion to the research question.

14.3 Module 12 Review - Independent Samples

The "social brain hypothesis," or the "social intelligence hypothesis," suggests that living in socially cohesive groups with differentiated relationships requires a higher cognitive load, in turn resulting in higher brain volume. There is evidence this hypothesis holds for primates and some other mammal groups, but it hasn't been explored in birds, as most birds typically have temporary social groupings that lack clear relationships. However, woodpeckers have a wide range of clearly differing social relationships while also having the benefit of being physiologically and environmentally similar across species. Researchers want to know if the "social brain hypothesis" holds true for woodpeckers: is the average brain volume (in mm^3) smaller for woodpeckers that tend to be solitary compared to woodpeckers that tend to live in pairs or groups? For the purpose of this study, "solitary" birds are classified as those that only pair-bond to breed, and otherwise are solitary for more than half a year each year. "Group-living" birds are those that spend more than half the year in communal groups or flocks. Researchers examined 61 species of woodpeckers. Use solitary - group living as the order of subtraction

The summary of the data and boxplots are given below:

```
woodpeckers <- read.csv("data/woodpeckers.csv")
# Summary statistics
woodpeckers %>%
    reframe(favstats(Volume~SocialCategory))
#>   SocialCategory min      Q1 median     Q3  max     mean       sd  n missing
#> 1   group_living 292 1623.75 2731.5 4562.5 7856 3179.900 2062.236 20       0
#> 2        solitary 600 1450.00 2106.0 3060.0 7700 2483.927 1539.478 41       0
```

```
# Side-by-side box plots
woodpeckers %>%
ggplot(aes(x = SocialCategory, y = Volume)) +
    geom_boxplot() +
    labs(title = "Boxplots of Brain Volume for Woodpeckers \n by Social Grouping",
        x = "Social Grouping",
        y = "Brain Volume (mm^3")
```
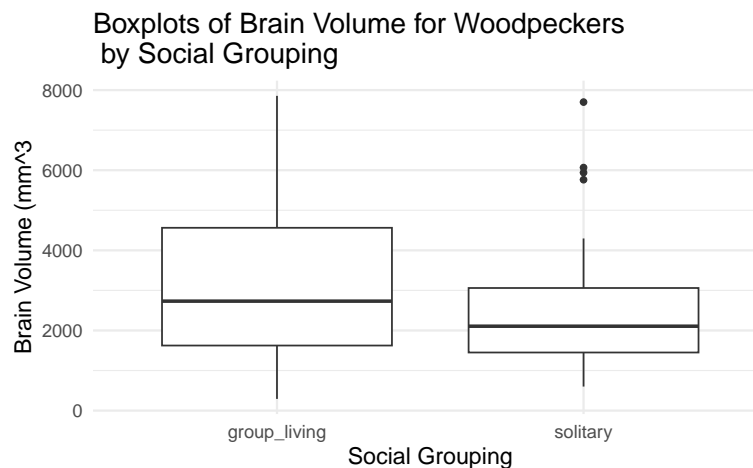

1. Write out the parameter of interest in context of the problem. Use proper notation.
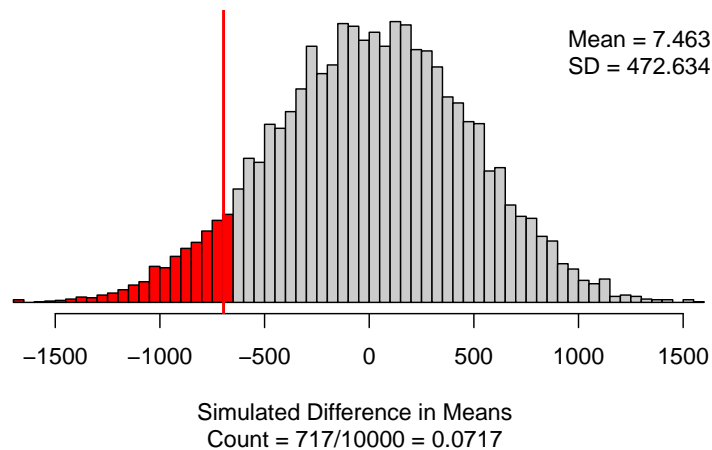
2. Write the null hypothesis in notation.

3. Write the alternative hypothesis in words.

4. Calculate the summary statistic. Use proper notation.

Simulation Methods

Hypothesis Testing

In the two_mean_test function, enter the response~explanatory variable names in for the formula (response~explanatory) and the name of the data set (woodpeckers) for data. Since the order of subtraction is `solitary - group_living` enter `solitary` for first_in_subtraction. Enter the summary statistic in for as_extreme_as and choose the direction to match the alternative hypothesis.

```
set.seed(216)
two_mean_test(Volume~SocialCategory, data = woodpeckers,  #Variables and data
                    first_in_subtraction = "solitary", #First value in order of subtraction
                    number_repetitions = 10000,  #Number of simulations
                    as_extreme_as = -695.973, #Observed statistic
                    direction = "less")  #Direction of alternative: "greater", "less", or "two-sided"
```


Mean = 7.463
SD = 472.634

Simulated Difference in Means
Count = 717/10000 = 0.0717

5. Based on the p-value for this study, explain why each of the following are false.

A. There is strong evidence that there is a true mean difference in brain volume for species of woodpeckers that live solo and those that live in groups (solitary - group).

B. If the difference in true mean brain volume for species of woodpeckers that live solo and that live in groups is less than zero, in 717 out of 10000 samples, we would observe a sample difference in mean brain volume of -695.973 mm^3 or less.

C. The 99% confidence interval would not include the value of zero.

D. We could conclude that the brain volume for species of woodpeckers that live solo is less than for those that live in groups when in fact there is no difference in brain volume for species of woodpeckers that live solo and that live in groups.

Bootstrap Confidence Interval To find the 99% confidence interval for the true difference in mean brain volume for species of woodpeckers that live in groups and species of woodpeckers that live solo use the two_mean_bootstrap_CI. The inputs are similar as to what we used in the two_mean_test.

```
set.seed(216)
two_mean_bootstrap_CI(Volume~SocialCategory, data = woodpeckers,  #Variables and data
                    first_in_subtraction = "solitary", #First value in order of subtraction
                    number_repetitions = 10000,  #Number of simulations
                    confidence_level = 0.99)
```


Bootstrap Difference in Means
99% CI: (−2026.254, 539.059)
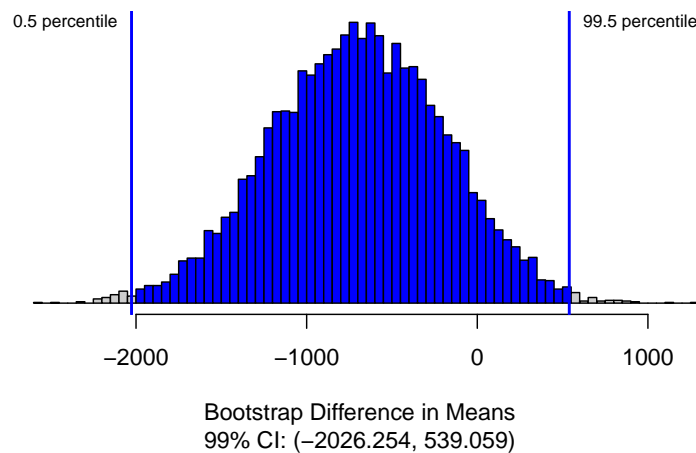
6. Interpret the confidence interval in context of the problem.

7. Write a conclusion to the research question.

Hypothesis testing using theory-based methods

Standardized Statistic:

$T = \frac{\bar{x}_1 - \bar{x}_2 - \text{null value}}{SE(\bar{x}_1 - \bar{x}_2)}$

$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

8. Calculate the standard error of the difference in means.

9. Calculate the standardized difference in sample mean.

Enter the t score into the pt function using a df = minimum(n - 1) = 20 - 1 = 19, and lower.tail = FALSE.

```
pt(-1.338, df=19, lower.tail=TRUE)
#> [1] 0.09834555
```

10. Why do we use lower.tail=TRUE to find the p-value?

Confidence interval using theory-based methods

To calculate the 99% confidence interval we use the formula:

$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$ we will need to find the t^* multiplier using the function qt.

For a 95% confidence interval we are finding the t^* value at the 99.5th percentile with df = minimum(n - 1) = 20 - 1 = 19.

```
qt(0.995, df = 19, lower.tail=TRUE)
#> [1] 2.860935
```

11. Calculate the 99% confidence interval.

12. Why do the simulation and theory based methods not give the same results?

14.4   Module 13 Review - Regression

Spotify created a list of the top songs around the world for the past 10 years and several different audio features of those songs. Among the variables measured on these songs, we will look at the relationship between Valance and Danceability. Valance measures the positive mood of a song; the higher the point value the more positive the mood of the song. Danceability measures how easy it is to dance to a song; the higher the point value the easier it is to dance to the song. Is there evidence that songs with a higher valance value are more danceable, on average?

```r
songs <- read.csv("data/top10s.csv") #Reads in data set
songs %>% #Data set pipes into...
ggplot(aes(x = Valance, y = Danceability))+  #Specify variables
  geom_point() +  #Add scatterplot of points
  labs(x = "valance",  #Label x-axis
       y = "danceability",  #Label y-axis
       title = "Scatterplot of Valance vs. Danceability for Top songs") + #Be sure to title your plot
  theme_bw()  #Add regression line
```


Scatterplot of Valance vs. Danceability for Top songs
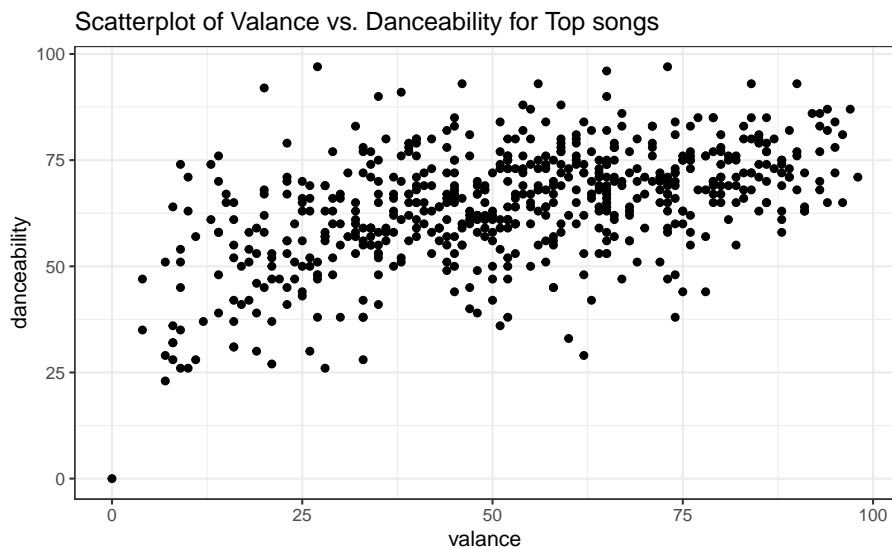
1. Identify the explanatory variable and the response variable.

2. Describe the scatterplot using the four charactistics of scatterplots.

The linear model output is given below with the correlation coefficient.

```
# Fit linear model: y ~ x
songsLM <- lm(Danceability~Valance, data=songs)
round(summary(songsLM)$coefficients, 5) # Display coefficient summary
```

```
#>             Estimate Std. Error  t value Pr(>|t|)
#> (Intercept) 48.80920    1.19239 40.93393        0
#> Valance      0.29814    0.02097 14.21805        0
```

```
cor(songs$Danceability, songs$Valance)
```

```
#> [1] 0.5016962
```

3. Write the least squares equation of the regression line in context of the problem.

4. Interpret the slope in context of the problem.

5. Write the null hypothesis, in words, in context of the problem.

6. Write the alternative hypothesis, in notation, to test slope, in context of the problem.

Simulation Methods

The following code creates the null distribution for this study.

```
# Simulation-based test for slope
regression_test(Danceability~Valance, # response ~ explanatory
                data = songs, # name of data set
                direction = "greater", # sign in alternative ("greater", "less", "two-sided")
                summary_measure = "slope",
                as_extreme_as = 0.298, #observed slope
                number_repetitions = 10000) #Number of simulated samples for null distribution
```

Mean = 0
SD = 0.024

Simulated Slope
Count = 0/10000 = 0
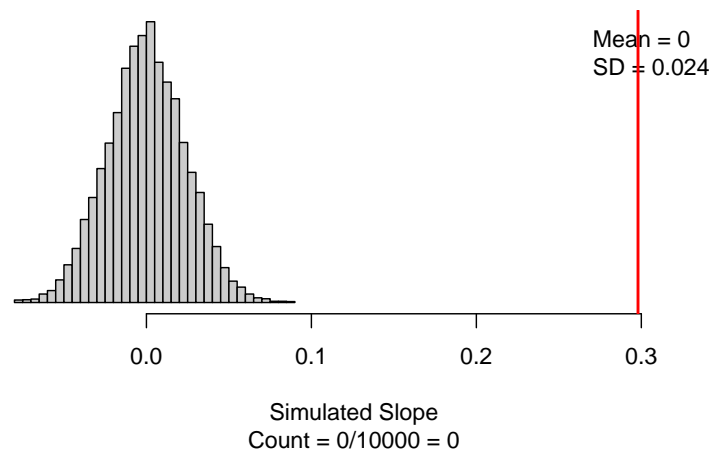
7. Report the value of the p-value. Interpret this value in context of the problem.

8. Based on the p-value, write a conclusion in context of the problem.

Now let's estimate the true regression slope for the relationship between valance and danceability of songs.

```
# Bootstrap CI for slope
regression_bootstrap_CI(Danceability~Valance, # response ~ explanatory
                        data = songs, # name of data set
                        confidence_level = 0.99, # confidence level as decimal
                        summary_measure =  "slope", # slope or correlation
                        number_repetitions = 10000) #Number of simulated samples for null distribution
```

0.5 percentile   99.5 percentile

0.25   0.30   0.35

Bootstrap Slope
99% CI: (0.24, 0.359)

9. Interpret the 99% confidence interval in context of the problem.

Now let's test correlation.

10. How will the null and alternative hypotheses change?

```
# Simulation-based test for correlation
regression_test(Danceability~Valance, # response ~ explanatory
                data = songs, # name of data set
                direction = "greater", # sign in alternative ("greater", "less", "two-sided")
                summary_measure =  "correlation",
                as_extreme_as = 0.502, # observed correlation
                number_repetitions = 10000) #Number of simulated samples for null distribution
```
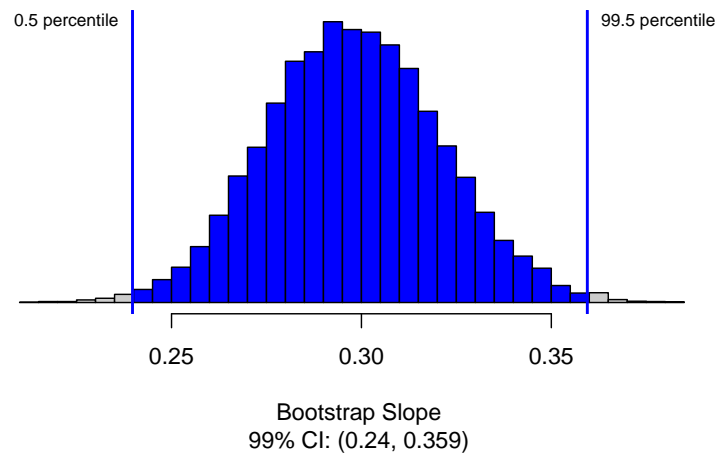
Mean = 0
SD = 0.041

Simulated Correlation
Count = 0/10000 = 0

```
# Bootstrap CI for correlation
regression_bootstrap_CI(Danceability~Valance, # response ~ explanatory
                        data = songs, # name of data set
                        confidence_level = 0.99, # confidence level as decimal
                        summary_measure =  "correlation",
                        number_repetitions = 10000) #Number of simulated samples for null distribution
```
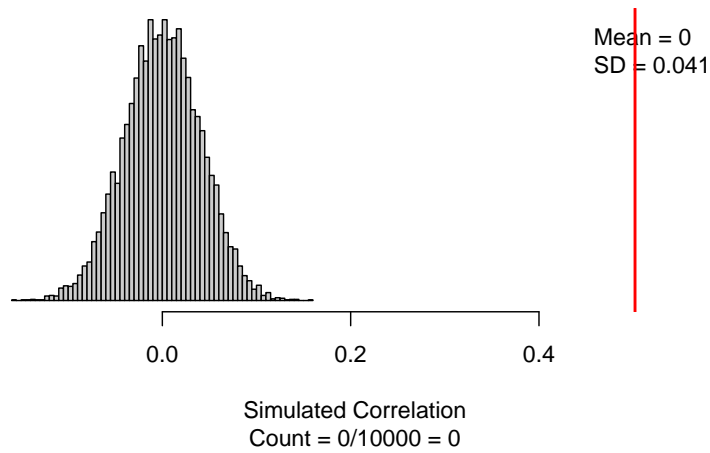

0.5 percentile   99.5 percentile

Bootstrap Correlation
99% CI: (0.401, 0.601)
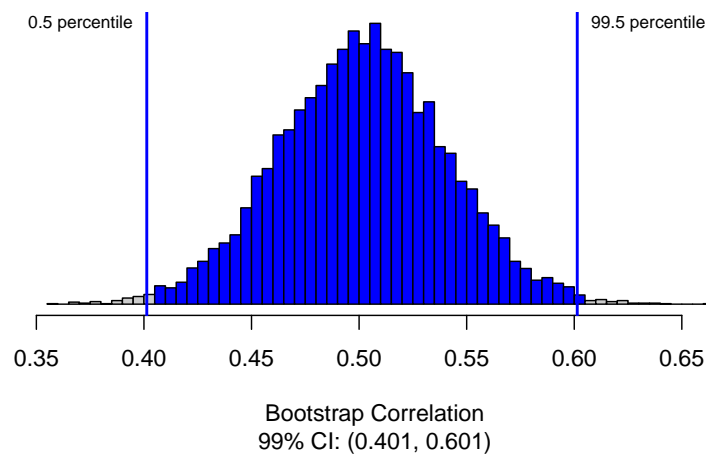
11. Interpret the 99% confidence interval for the true correlation between valance and danceability.

Theory-based Methods

When performing inference on a least squares line, the follow conditions are generally required

- Linearity: the data should follow a linear trend

- Nearly normal residuals: residuals must be nearly normal

- Constant variability: the variability of points around the least squares line remains roughly constant

- Independent observations: individual data points must be independent

The scatterplot and the residual plots will be used to assess the conditions for approximating the data with the t-distribution.
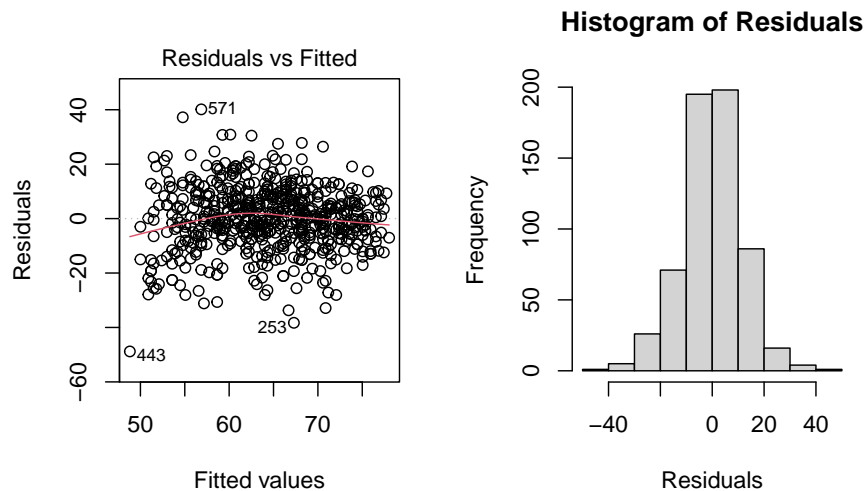
12. Are the conditions met to use the t-distribution to approximate the sampling distribution of our test statistic?

To find the value of the test statistic to test the slope we will use,

$$T = \frac{b_1 - \text{null value}}{SE(b_1)}$$

We will use the linear model output above to get the estimate for slope and standard error.

13. Calculate the standardized slope.

14. Using the linear model output, report the p-value for the test of significance.

15. Based on the p-value, how much evidence is there against the null hypothesis?

Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^* \times SE(estimate)$$

$$b_1 \pm t^* \times SE(b_1)$$

The t^* multiplier comes from the t-distribution with $n-2$ df. Recall for a 99% confidence interval, use the 99.5% percentile (99% of the distribution is in the middle, leaving 0.5% in each tail). The sample size is 603 so the df is 601.

```
qt(0.995, 601) #95% t* multiplier
```

```
#> [1] 2.584034
```

16. Calculate the 99% confidence interval for the true slope.

Semester Review

15.1 Group Final Exam Review

Use the provided data set from the Islands (Bulmer, n.d.) (FinalExamReviewData.csv) and the appropriate Exam Review R script file to answer the following questions. Each adult (>21) islander was selected at random from all adult islanders. Note that some islanders choose not to participate in the study. These islanders that did not consent to be in the study are removed from the dataset before analysis. Variables and their descriptions are listed below. Here is some more information about some of the variables collected. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measured after listening to music for each Islander. Heart rate and blood glucose levels were both measured before and then after drinking a caffeinated beverage.

Variable	Description
Island	Name of Island that the Islander resides on
City	Name of City in which the Islander resides
Population	Population of the City
Name	Name of Islander
Consent	Whether the Islander consented to be in the study (`Declined`, `Consented`)
Gender	Gender of Islander (`M` = male, `F` = Female)
Age	Age of Islander
Married	Marital status of Islander (`yes`, `no`)
Smoking_Status	Whether the Islander is a current smoker (`nonsmoker`, `smoker`)
Children	Whether the Islander has children (`yes`, `no`)
weight_kg	Weight measured in kg
height_cm	Height measured in cm
respiratory_rate	Breaths per minute
Type_of_Music	Music type Islander was randomly assigned to listen to (`Classical`, `Heavy Metal`)
After_PuzzleCube	Time to complete puzzle cube (minutes) after listening to assigned music
Education_Level	Highest level of education completed (`highschool`, `university`)
Balance_Test	Time balanced measured in seconds with eyes closed
Blood_Glucose_before	Level of blood glucose (mg/dL) before consuming assigned drink
Heart_Rate_before	Heart rate (bpm) before consuming assigned drink
Blood_Glucose_after	Level of blood glucose (mg/dL) after consuming assigned drink
Heart_Rate_after	Heart rate (bpm) after consuming assigned drink

1. Use the appropriate Exam 2 Review R script file and analyze the following research question: "Is there evidence that adult Islanders with a University degree are less likely to smoke than those with a High School degree?"

a. Parameter of Interest:

b. Null Hypothesis:

Notation:

Words:

c. Alternative Hypothesis:

Notation:

Words:

d. Use the R script file to get the counts for each level and combination of variables. Fill in the following table with the variable names, levels of each variable, and counts using the values from the R output.

	Explanatory Variable		
Response variable	Group 1	Group 2	Total
Success			
Failure			
Total			

e. Calculate the value of the summary statistic to answer the research question. Give appropriate notation.

f. Interpret the value of the summary statistic in context of the problem:

g. Assess if the following conditions are met:

Independence (needed for both simulation and theory-based methods):

Success-Failure (must be met to use theory-based methods):

h. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

j. Write a conclusion to the research question based on the p-value.

k. Using a significance level of $\alpha = 0.05$, what statistical decision will you make about the null hypothesis?

l. Use the provided R script file to find a 95% confidence interval.

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

n. Regardless to your answer in part g, calculate the standardized statistic.

o. Interpret the value of the standardized statistic in context of the problem.

p. Use the provided R script file to find the theory-based p-value.

q. Use the provided R script file to find the appropriate z* multiplier and calculate the theory-based confidence interval.

r. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

s. What is the scope of inference for this study?

2. Use the appropriate Final Exam Review R script file to analyze the following research question, "Is there evidence that adult Islander's heart rates increase after drinking a caffeinated beverage compared to before drinking a caffeinated beverage, on average?" Use before − after as the order of subtraction.

a. Parameter of Interest:

b. Null Hypothesis:

   Notation:

   Words:

c. Alternative Hypothesis:

   Notation:

   Words:

d. Use the R script file to get the summary statistics. Fill in the following table with the variable names, levels of each variable, and values from the R output.

Summary value	Variable
Mean	
Standard deviation	
Sample size	

e. Write the summary statistic to answer the research question with appropriate notation.

f. Interpret the value of the summary statistic in context of the problem:

g. Assess if the following conditions are met:

   Independence (needed for both simulation and theory-based methods):

Normality:

h. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

j. Write a conclusion to the research question based on the p-value.

k. Using a significance level of $\alpha = 0.1$, what statistical decision will you make about the null hypothesis?

l. Use the provided R script file to find a 90% confidence interval.

m. Interpret the 90% confidence interval in context of the problem.

n. Regardless to your answer in part g, calculate the standardized statistic.

o. Interpret the value of the standardized statistic in context of the problem.

p. Use the provided R script file to find the theory-based p-value.

q. Use the provided R script file to find the appropriate t* multiplier and calculate the theory-based confidence interval.

r. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

s. What is the scope of inference for this study?

3. Use the appropriate Final Exam Review R script file to analyze the following research question: "Is there evidence that adult Islanders who listen to classical music take less time, on average, to complete the puzzle cube after listening to the music than for Islanders that listen to heavy metal music?" Use - classical - heavy metal as the order of subtraction.

a. Parameter of Interest:

b. Null Hypothesis:

Notation:

Words:

c. Alternative Hypothesis:

Notation:

Words:

d. Use the R script file to get the summary statistics for each level of the explanatory variable. Fill in the following table with the variable names, levels of each variable, and values from the R output.

	Explanatory Variable	
Summary value	Group 1	Group 2
Mean		
Standard deviation		
Sample size		

e. Calculate the value of summary statistic to answer the research question. Give appropriate notation.

f. Interpret the value of the summary statistic in context of the problem:

g. Assess if the following conditions are met:

Independence (needed for both simulation and theory-based methods):

Normality:

h. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

j. Write a conclusion to the research question based on the p-value.

k. Using a significance level of $\alpha = 0.05$, what statistical decision will you make about the null hypothesis?

l. Use the provided R script file to find a 95% confidence interval.

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

n. Regardless to your answer in part g, calculate the standardized statistic.

o. Interpret the value of the standardized statistic in context of the problem.

p. Use the provided R script file to find the theory-based p-value.

q. Use the provided R script file to find the appropriate t* multiplier and calculate the theory-based confidence interval.

r. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

s. What is the scope of inference for this study?

4. Use the appropriate Final Exam Review R script file to analyze the following research question: "Is there evidence that height of adult Islanders can be used to predict their balance time?"

a. Parameter of Interest:

b. Null Hypothesis:

   Notation:

   Words:

c. Alternative Hypothesis:

   Notation:

   Words:

d. Use the R script file to get the summary statistics for this data. Fill in the following table using the values from the R output:

	y-intercept	slope	correlation
Summary value			

e. Interpret the value of slope in context of the problem.

f. Assess if the following conditions are met:

   Independence (needed for both simulation and theory-based methods):

   Linearity (needed for both simulation and theory-based methods):

   Constant Variance:

Normality of Residuals:

g. Use the provided R script file to find the simulation p-value to assess the research question. Report the p-value.

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

i. Write a conclusion to the research question based on the p-value.

j. Using a significance level of $\alpha = 0.01$, what statistical decision will you make about the null hypothesis?

k. Use the provided R script file to find a 99% confidence interval.

l. Interpret the 99% confidence interval in context of the problem.

m. Regardless to your answer in part g, calculate the standardized statistic.

n. Interpret the value of the standardized statistic in context of the problem.

o. Use the provided R script file to find the theory-based p-value.

p. Use the provided R script file to find the appropriate t* multiplier and calculate the theory-based confidence interval.

q. Does the theory-based p-value and CI match those found using simulation methods? Explain why or why not.

r. What is the scope of inference for this study?

15.2 Golden Ticket to Descriptive and Inferential Statistical Methods

In this course, we have covered descriptive (summary statistics and plots) and inferential (hypothesis tests and confidence intervals) methods for five different scenarios:

- one categorical response variable
- two categorical variables
- one quantitative response variable or paired differences in a quantitative variable
- two quantitative variables
- one quantitative response variable and one categorical explanatory variable

The "golden ticket" shown on the next page presents a visual summary of the similarities and differences across these five scenarios.

Scenario	One Categorical Response	Two Categorical Variables	One Quantitative Response or Paired Differences	Quant. Response and Categ. Explanatory (independent samples)	Two Quantitative Variables
Type of plot	Bar plot	Segmented bar plot, Mosaic plot	Dotplot, histogram, boxplot	Side-by-side boxplots, Stacked dotplots or histograms	Scatterplot
Summary measure	Proportion	Difference in proportions	Mean or Mean difference	Difference in means	Slope or correlation
Parameter notation	π	$\pi_1 - \pi_2$	μ or μ_d	$\mu_1 - \mu_2$	β_1 or ρ
Statistic notation	$\hat{p}$	$\hat{p}_1 - \hat{p}_2$	$\bar{x}$ or $\bar{x}_d$	$\bar{x}_1 - \bar{x}_2$	b_1 or r
Null hypothesis	$H_0: \pi = \pi_0$	$H_0: \pi_1 - \pi_2 = 0$	$H_0: \mu = \mu_0$ or $H_0: \mu_d = 0$	$H_0: \mu_1 - \mu_2 = 0$	$H_0: \beta_1 = 0$ or $H_0: \rho = 0$
Conditions for simulation-based methods	Independent cases	Independent cases (within and between groups)	Independent cases	Independent cases (within and between groups)	Independent cases; Linear form
Simulation test (how to generate a null distn)	Spin spinner with probability equal to π_0, n times or draw with replacement n times from a deck of cards created to reflect π_0 as probability of success. Plot the proportion of successes. Repeat 10000 times. Centered at π_0.	Label cards with response values from original data; mix cards together; shuffle into two new groups of sizes n_1 and n_2. Plot difference in proportion of successes. Repeat 10000 times. Centered at 0.	Shift the original data by adding $(\mu_0 - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Plot sample mean or sample mean difference. Repeat 10000 times. Centered at 0.	Label cards with response variable values from original data; mix cards together; shuffle into two new groups of sizes n_1 and n_2. Plot difference in means. Repeat 10000 times. Centered at 0.	Hold the x values constant; shuffle new y's to x's. Find the regression line for the shuffled data; plot the slope or the correlation for the shuffled data. Repeat 10000 times. Centered at 0.
p-value = proportion of null simulations at or beyond (H_A direction) the observed statistic					
Bootstrap CI (how to generate a boot. distn)	Label n cards with the original responses. Randomly draw n times with replacement. Plot the resampled proportion of successes. Repeat 10000 times. Centered at $\hat{p}$.	Label n_1 cards with the original responses from group 1 and n_2 cards with the original responses from group 2. Randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Plot the resampled difference in proportion of successes. Repeat 10000 times. Centered at $\hat{p}_1 - \hat{p}_2$.	Label n cards with the original responses. Randomly draw with replacement n times. Plot the resampled mean difference. Repeat 10000 times. Centered at $\bar{x}_d$.	Label n_1 cards with the original responses from group 1 and n_2 cards with the original responses from group 2. Keep groups separate. Randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Plot the resampled difference in means. Repeat 10000 times. Centered at $\bar{x}_1 - \bar{x}_2$.	Label n cards with the original (explanatory, response) pairs. Randomly draw with replacement n times. Plot the resampled slope or correlation. Repeat 10000 times. Centered at b_1 or r.
X% CI: $\left(\dfrac{1-X}{2}\%tile,\ \left(X + \dfrac{1-X}{2}\right)\%tile\right)$					
Theory-based distribution	Standard Normal	Standard Normal	t-distribution with $n-1$ df	t-distribution with min of $n_1 - 1$ or $n_2 - 1$ df	t-distribution with $n-2$ df
Conditions for theory-based hypothesis tests and confidence intervals	Independent cases; Number of successes and number of failures in sample both at least 10.	Independence (within and between groups); Number of successes and number of failures in EACH sample all at least 10. (All four cell counts at least 10.)	Independent cases; $n < 30$ with no clear outliers OR $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$	Independent cases (within and between groups); In each sample, $n < 30$ with no clear outliers OR $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$	Linear form; Independent cases; Nearly normal residuals; Variability around the regression line is roughly constant.
Theory-based standardized statistic (test statistic)	$Z = \dfrac{\hat{p} - \pi_0}{SE_0(\hat{p})}$ $= \sqrt{\dfrac{\pi_0 \times (1 - \pi_0)}{n}}$	$Z = \dfrac{\hat{p}_1 - \hat{p}_2 - 0}{SE_0(\hat{p}_1 - \hat{p}_2)}$ $SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\widehat{p_{pool}} \times (1 - \widehat{p_{pool}}) \times \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$	$T = \dfrac{\bar{x} - 0}{SE(\bar{x})}\ OR\ T = \dfrac{\bar{x}_d - 0}{SE(\bar{x}_d)}$ $SE(\bar{x}) = \dfrac{s}{\sqrt{n}}, SE(\bar{x}_d) = \dfrac{s_d}{\sqrt{n}}$	$T = \dfrac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	$T = \dfrac{b_1 - 0}{SE(b_1)}$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.
Theory-based confidence interval	$\hat{p} \pm z^* \times SE(\hat{p})$ $SE(\hat{p}) = \sqrt{\dfrac{\hat{p} \times (1 - \hat{p})}{n}}$	$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$ $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\dfrac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$	$\bar{x} \pm t^* \times SE(\bar{x})$ $\bar{x}_d \pm t^* \times SE(\bar{x}_d)$ $SE(\bar{x}) = \dfrac{s}{\sqrt{n}}, SE(\bar{x}_d) = \dfrac{s_d}{\sqrt{n}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	$b_1 \pm t^* \times SE(b_1)$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.

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).

LUETKEMEIER, et al., M. 2017. "Skin Tattoos Alter Sweat Rate and Na+ Concentration." *Medicine and Science in Sports and Exercise* 49(7).

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-Confirmed 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.*

Watson, et al., N. 2015. "Recommended Amount of Sleep for a Heathy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society." *Sleep* 38(6).

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).