STAT 216 Coursepack

Summer 2024
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

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

Contents

Preface

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

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, reading guides are provided on D2L to aid in taking notes while you complete the required readings. 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!

The out-of-class activities will be completed outside of class, typically between the Monday and Wednesday classes. The additional 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.
  - 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*:
  - Fill in the lecture notes during class.
  - Work through the in-class activity or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
  - Complete any parts of the activity you did not complete in class.
  - Review the activity solutions in the Math and Stat Center, and take notes on key points.
  - Complete any remaining assigned readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Basics of Data

1.1 Activity 1: Intro to Data

1.1.1 Learning outcomes

- Creating a data set

1.1.2 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.

1.1.3 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.

1.1.4 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. *Weeks 3–4*
4. Use statistical analysis methods to draw inferences from the data. *Weeks 6–14*
5. Communicate the results and answer the research question. *Weeks 6–14*
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. One person from each group open the Google sheet linked in D2L and fill in the responses for the following questions for each group member. 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?

1.1.5 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.

1.1.6 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.

1.2 Lecture Notes Module 1: Intro to data

Read through Sections 1.2.1 – 1.2.5 in the course textbook prior to coming to class on Friday using the reading guides at the beginning of week 1 material.

Data basics: Sections 1.2.1 – 1.2.2

Data: _____ used to answer research questions

Observational unit or case: the people or things we _____ data from

Variable: what is measured on each _____ or _____.

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 for class discussion: 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 number which _____ an entire data set

- Also called the _____ _____

  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: Sections 1.2.3 – 1.2.5

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 for class discussion:

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 _____.

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 give units of measure for numerical values. Write down some issues found with the created class data set.

MODULE 2

Study Design

2.1 Lecture Notes Module 2: Study Design

Sampling Methods: Section 2.1 in the course textbook

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

- Response bias:

Example(s): 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.

Examples for class discussion

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?

Observational studies, experiments, and scope of inference: Sections 2.2 – 2.4 in the course textbook

- Review
  - Explanatory variable: the variable researchers think *may be* effecting 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:

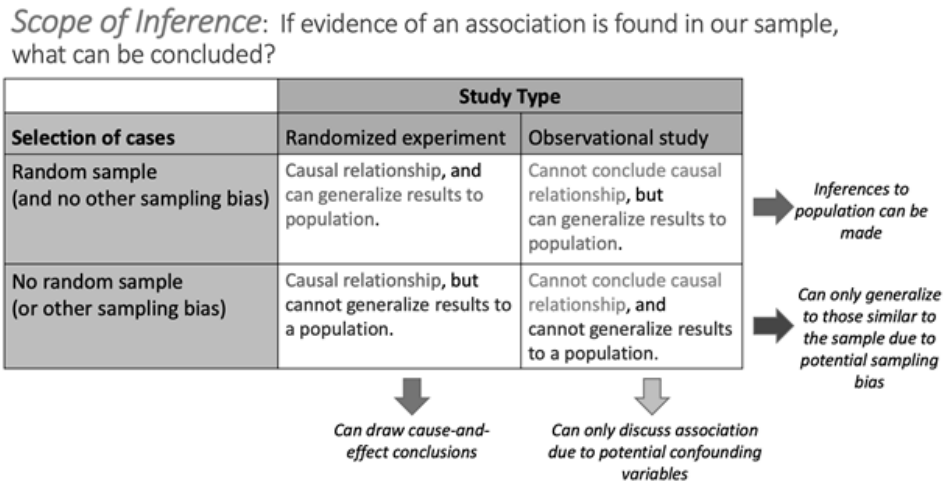
- Non-random sample with sampling bias:

2. What is the study design?

- Randomized experiment:

- Observational study:

Scope of Inference Table:



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?

Purpose of random assignment:

Purpose of random selection:

2.2 Out-of-Class Activity Module 2: American Indian Address

2.2.1 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.

2.2.2 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.

2.2.3 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.

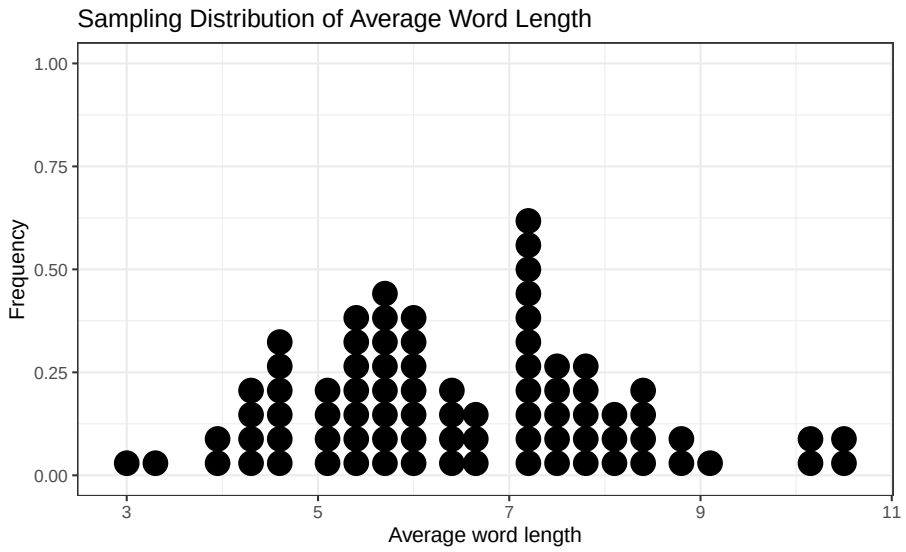
2. Fill in the table below with your selected words from the previous question and the length of each word (number of letters/digits in the word):

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?

A dot plot and summary statistics of the "by-eye" average word lengths from by-eye samples of size 10 from a Spring 2023 class with 82 students is provided.

```
#>   min  Q1 median  Q3  max     mean       sd  n missing
#> 1   3 5.3    6.3 7.5 10.6 6.479268 1.631309 82       0
```



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

5. 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?

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

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

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

9. 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?

10. 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?

Types of bias

11. 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.

12. A television station is interest in predicting whether or not a local referendum to legalize marijuana for adult use will pass. It 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.

13. To gauge the interest 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.

14. The Bozeman school district was interested in surveying parents of students about their opinions on returning to in-person classes following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division. Explain why selection bias would be present in this study design.

2.2.4 Take-home messages

1. There are three types of bias to be aware of when designing a sampling method: selection bias, non-response bias, and response bias.
2. When we use a biased method of selection, we will over or underestimate the parameter.
3. To see if a method is biased, 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.
4. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.

2.2.5 Additional notes

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

2.3 Activity 2: American Indian Address (continued)

2.3.1 Learning outcomes

- Explain the purpose of random selection and its effect on scope of inference.
- Select a simple random sample from a finite population using a random number generator.
- Explain why a sampling method is unbiased or biased.
- Explain the effect of sample size on sampling variability.

2.3.2 Terminology review

In today's 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 Section 2.1 in the textbook.

Random selection

Today we will return to the American Indian Address introduced in the out-of-class activity. 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.

1. 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".

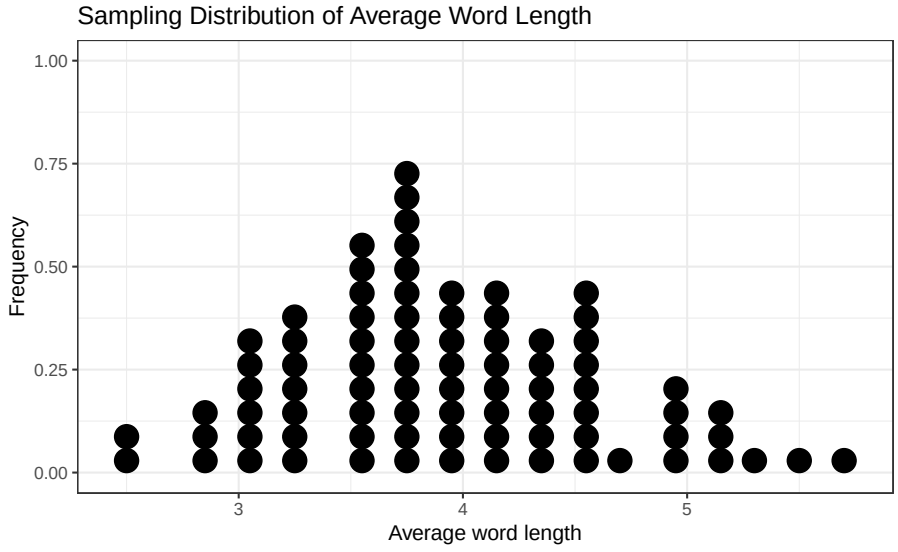
Fill in the table below with the random numbers selected and use the Becenti.csv data file found on D2L to determine each number's corresponding word and word length (number of letters/digits in the word):

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

2. Calculate the mean word length in your selected sample in question 1. Is this value a parameter or a statistic?

A dot plot and summary statistics of the average word lengths from random samples of size 10 from a Spring 2023 class of 82 students is provided.

```
#>   min  Q1 median    Q3 max     mean        sd  n missing
#> 1 2.5 3.5   3.85 4.375 5.7 3.926829 0.6856643 82       0
```



3. 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.

4. How does the plot given in this activity compare to the plot generated in the out-of-class activity?

   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.

- Navigate to the "One Variable with Sampling" Rossman/Chance web applet: http://www.rossmanchance.com/applets/2021/sampling/OneSample.html?population=gettysburg.
- Click "Clear" below the text box containing data from the Gettysburg address to delete that data set.
- Download the Becenti.csv file from D2L and open the spreadsheet on your computer.
- Copy and paste the population of word lengths (column C) into the applet from the data set provided making sure to include the header. Click "Use Data". Verify that the mean for the data set is 3.953 with a sample size of 359. If these are not the values you got, check with your instructor for help with copying in the data set correctly.
- Click the check-box for "Show Sampling Options"
- Select 1000 for "Number of samples" and select 10 for the "Sample size".
- Click "Draw Samples".

5. The plot labeled "Statistics" displays the 1,000 randomly generated sample mean word lengths. Sketch this plot below. Include a descriptive x-axis label and be sure to write down the provided mean and SD (standard deviation) of the distribution.

6. What is the center value (mean) of the distribution created in question 5?

7. 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".

8. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from question 5.

Effect of sample size

We will now consider the impact of sample size.

9. First, consider if each student had selected 20 words, instead of 10, by eye. Do you think this would make the plot from the out-of-class activity centered on 3.95 (the true mean word length)? Explain your answer.

10. Now we will select 20 words instead of 10 words at random.

- In the "One Variable with Sampling" Rossman/Chance web applet(http://www.rossmanchance.com/applets/2021/sampling/OneSample.html?population=gettysburg.), change the Sample size to 20.
- Click "Draw Samples".

The plot labeled "Statistics" displays the 1,000 randomly generated sample mean word lengths. Sketch this plot below. Include a descriptive x-axis label and be sure to write down the provided mean and SD (standard deviation) of the distribution.

11. Compare the distribution created in question 10 to the one created in question 5.

    Is the shape similar?

    Is the range (smallest to largest values) similar?

    Is the mean of the distribution similar?

12. Compare the values of the standard deviation of the plots in question 10 and in question 5. Which plot shows the smallest standard deviation?

13. 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

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

2.3.3 Take-home messages

1. Random selection is an unbiased method of selection.
2. 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.
3. Random selection eliminates selection bias. However, random selection will not eliminate response or non-response bias.
4. The larger the sample size, the more similar (less variable) the statistics will be from different samples.
5. 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.

2.3.4 Additional notes

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

2.4 Module 2 Lab: Study Design

2.4.1 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.

2.4.2 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.

2.4.3 General information labs

At the end of each week you will complete a lab. Questions are selected from each lab to be turned in on Gradescope. The questions to be submitted on Gradescope are bolded in the lab. As you work through the lab have the Gradescope lab assignment open so that you can answer those questions as you go.

2.4.4 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)?

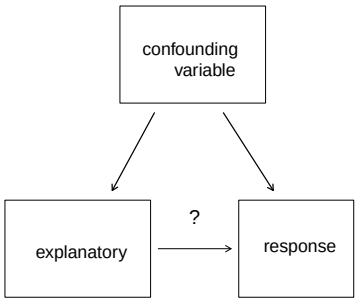
7. Notice the difference in the two proportions is shown as a dot in the plot at the bottom of the web page. Un-check the box for Animate above "Randomize" and click "Randomize" again. Did you get the same difference in proportion of males between the placebo and drug groups?

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?

	Study Type	
Selection of cases	**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

2.4.5 Study design

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:

2.4.6 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.

2.4.7 Additional notes

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

MODULE 3

Exploring Categorical and Quantitative Data

3.1 Lecture Notes Module 3: Exploratory Data Analysis

Summarizing categorical data

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

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

Notation used for the population proportion:

- Single categorical variable:

- Two categorical variables:

  - Subscripts represent the _____ variable groups

Notation used for the sample proportion:

- Single categorical variable:

- Two categorical variables

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

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

Example for class discussion: Gallatin Valley is the fastest growing county in Montana. You'll often hear Bozeman residents complaining about the 'out-of-staters' moving in. A local real estate agent recorded data on a random sample of 100 home sales over the last year at her company and noted where the buyers were moving from as well as the age of the person or average age of a couple buying a home. The variable age was binned

into two categories, "Under30" and "Over30." Additionally, the variable, state the buyers were moving from, was created as a binary variable, "Out" for a location out of state and "In" for a location in state.

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

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

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

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

- What are the observational units in this study?

- What type of variable is `Age`?

- What type of variable is `Age_Group`?

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

```
moving %>%
    count(From)
#>   From  n
#> 1   CA 12
#> 2   CO  8
#> 3   MT 61
#> 4   WA 19
```

Or a relative frequency table:

```
moving %>%
  count(From) %>%
  mutate(freq = n/sum(n))
#>   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?

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 _____ within each _____. 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 _____ within each _____. 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.

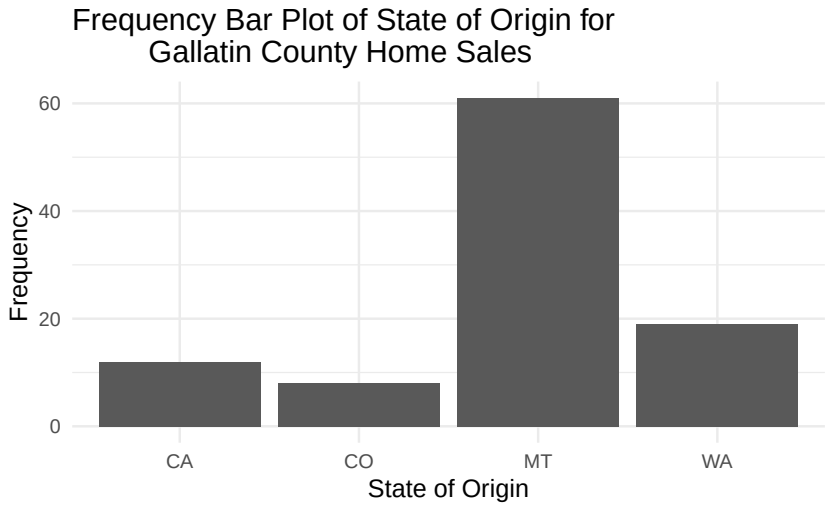
Displaying categorical variables

- Types of plots for a single categorical variable

- Types of plots for two categorical variables

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

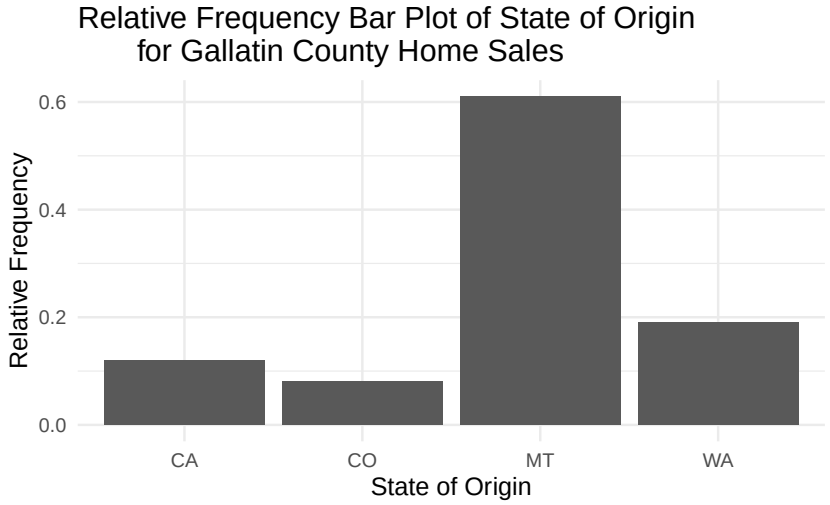
```
moving %>%
    ggplot(aes(x = From))+ #Enter the variable to plot
    geom_bar(stat = "count") +
    labs(title = "Frequency Bar Plot of State of Origin for
         Gallatin County Home Sales",
         #Title your plot (type of plot, observational units, variable)
       y = "Frequency", #y-axis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



- What can we see from this plot?

Additionally, we can create a relative frequency bar plot.

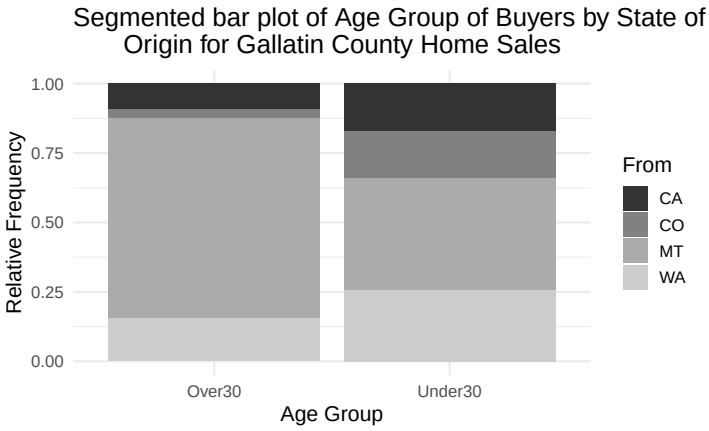
```
moving %>%
  ggplot(aes(x = From))+ #Enter the variable to plot
  geom_bar(aes(y = after_stat(prop), group = 1)) +
  labs(title = "Relative Frequency Bar Plot of State of Origin
       for Gallatin County Home Sales",
       #Title your plot
       y = "Relative Frequency", #y-axis label
       x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



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

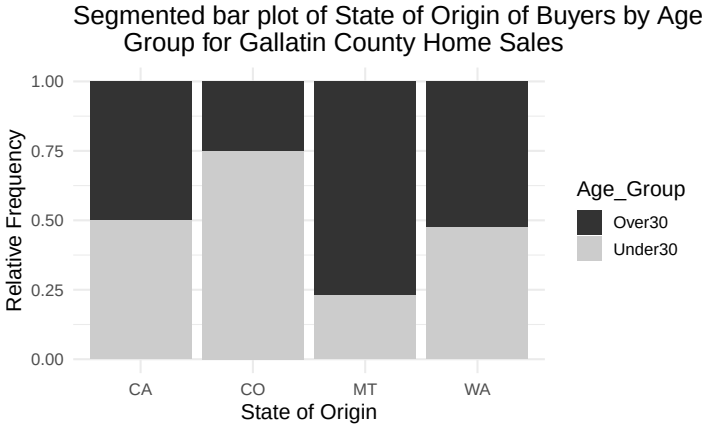
In a segmented bar plot, the bar for each category will sum to 1. In this first plot, we are plotting the row proportions calculated conditional on the age group.

```
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()
```



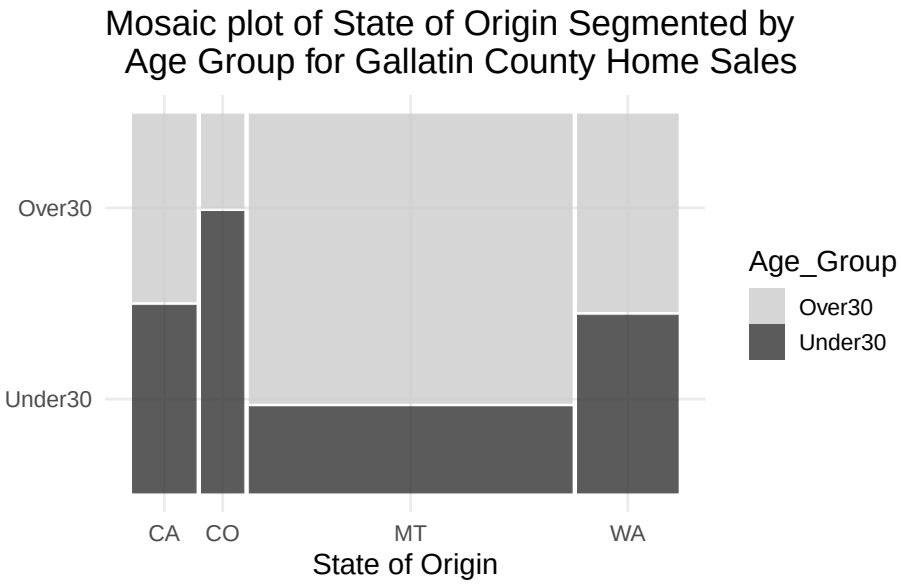
In this second plot, we are plotting the column proportions calculated conditional on the state of origin for the buyer.

```
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()
```



Mosaic plot:

```
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
```



- Why is the bar for MT the widest on the mosaic plot?

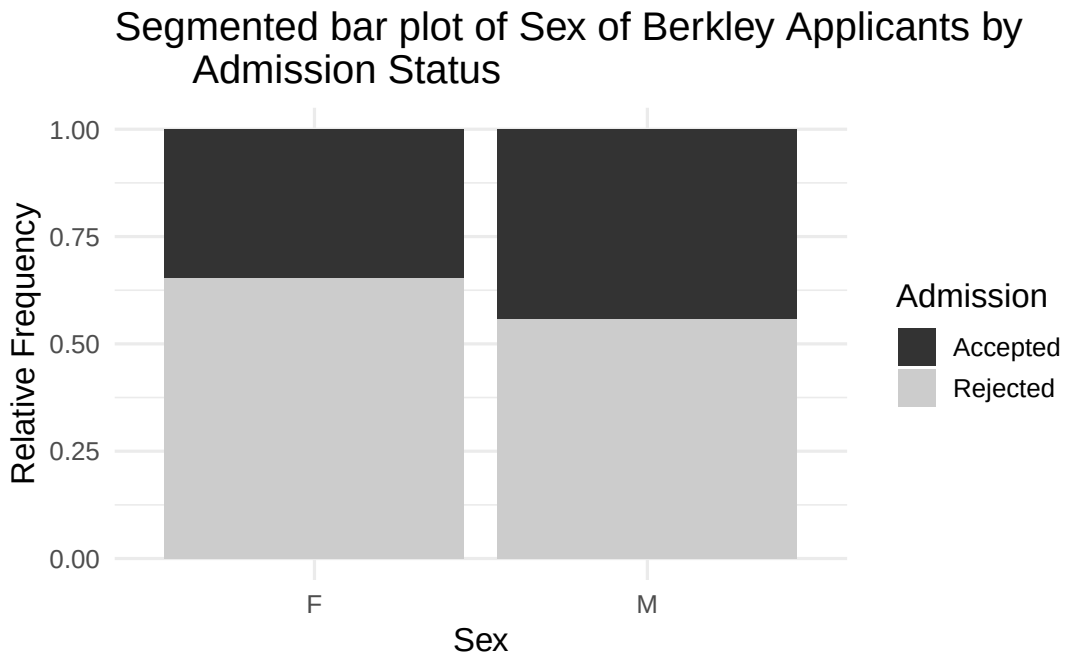
Simpson's paradox

- 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.

```
discrim <- read.csv ("https://waf.cs.illinois.edu/discovery/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()
```

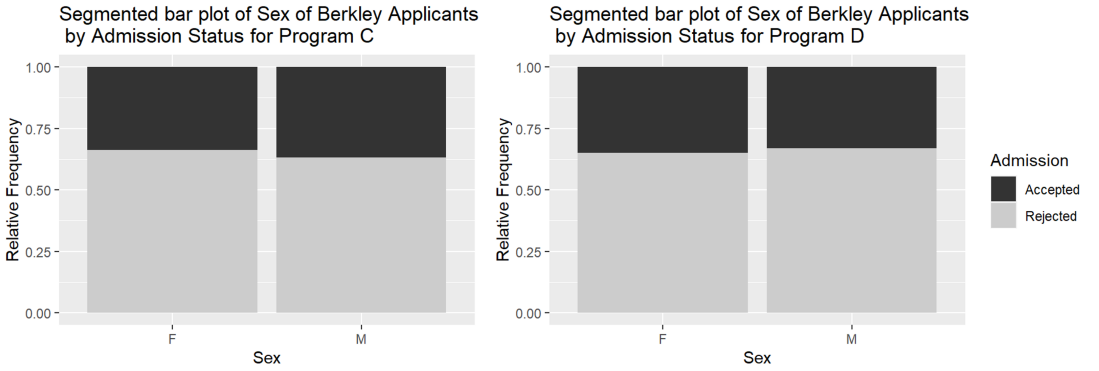


The data showed that 44% of male applicants were accepted and 35% of female applicants were accepted. Does it appear that the female students are discriminated against?

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.

Summarizing quantitative data

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: On average _____ each data point if 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

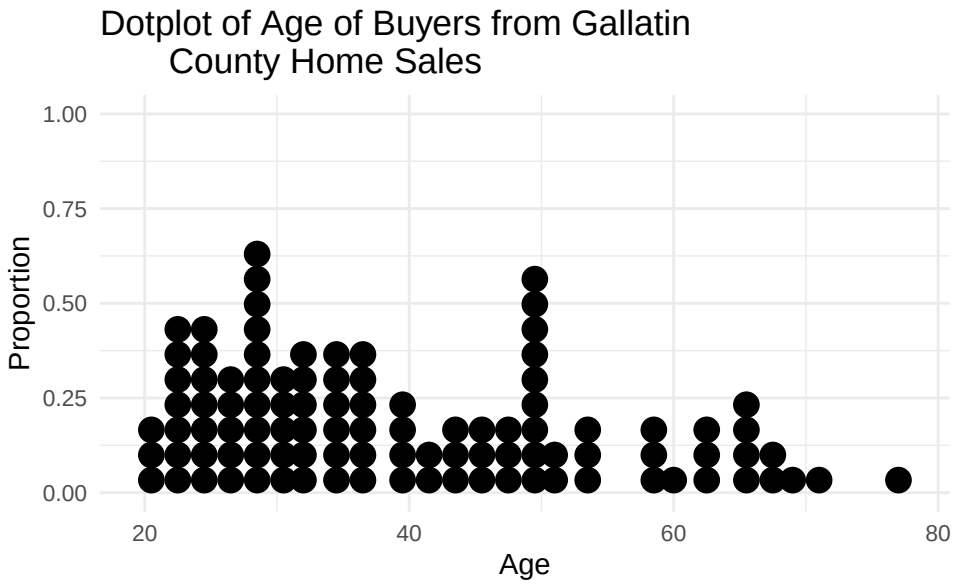
  - _____ % of values are at and _____ the value of Q1

Types of plots

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

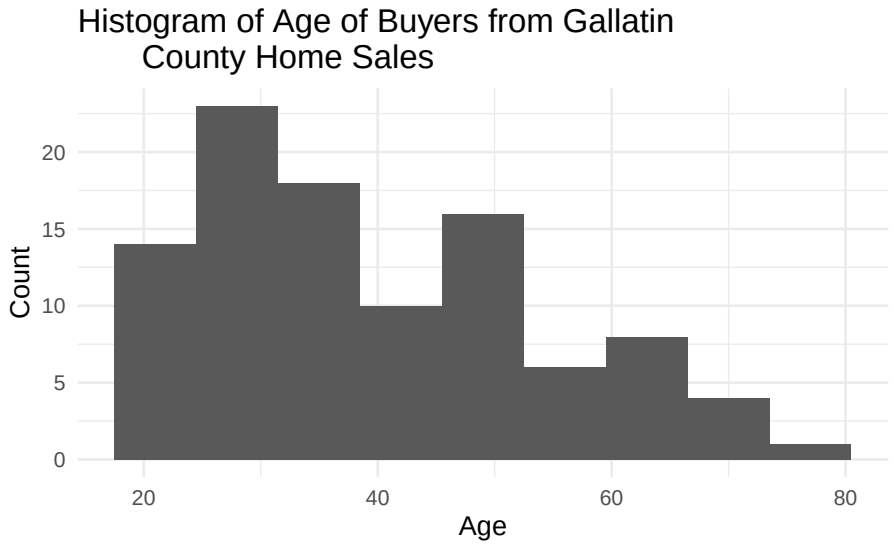
- Dotplot:

```
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
```



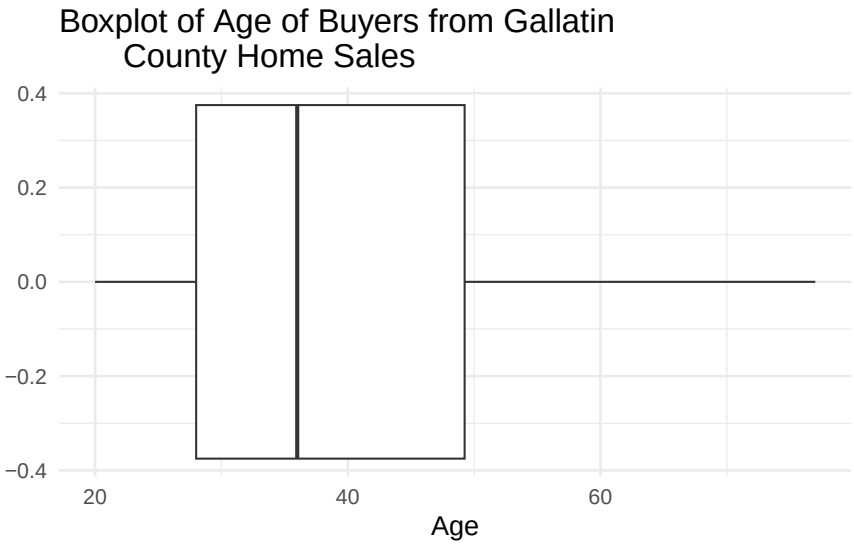
- Histogram

```
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")
```



- Boxplot
  - 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
```



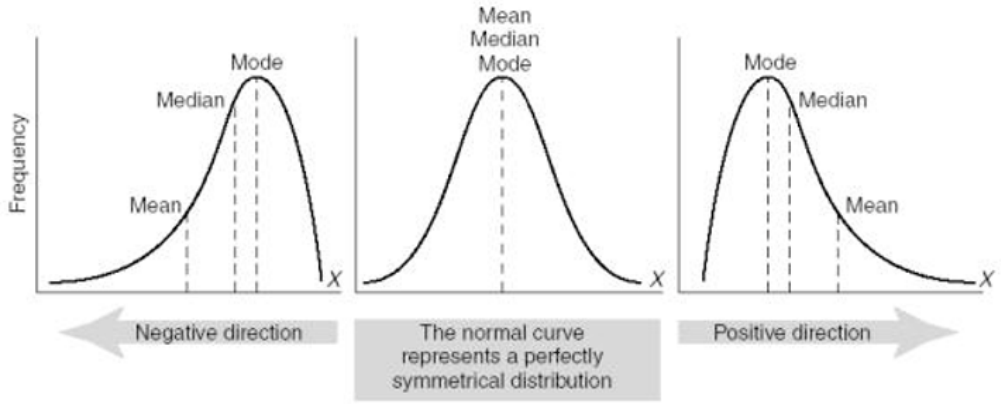

```
favstats(moving$Age)
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

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



  - 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?

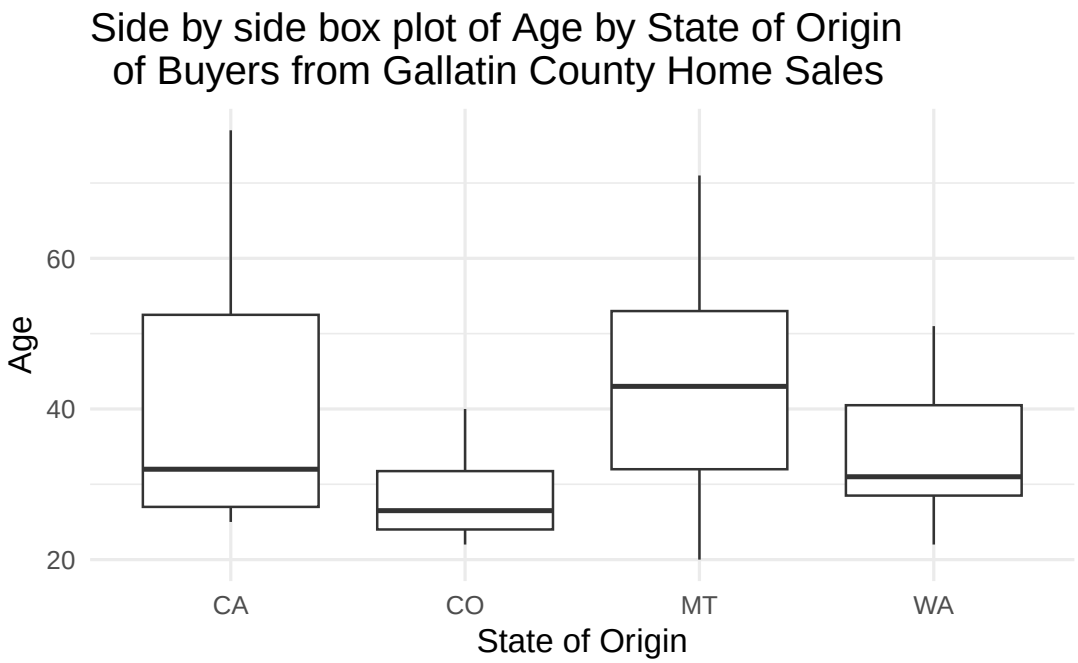
  values $< Q_1 - 1.5 \times IQR$

  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.

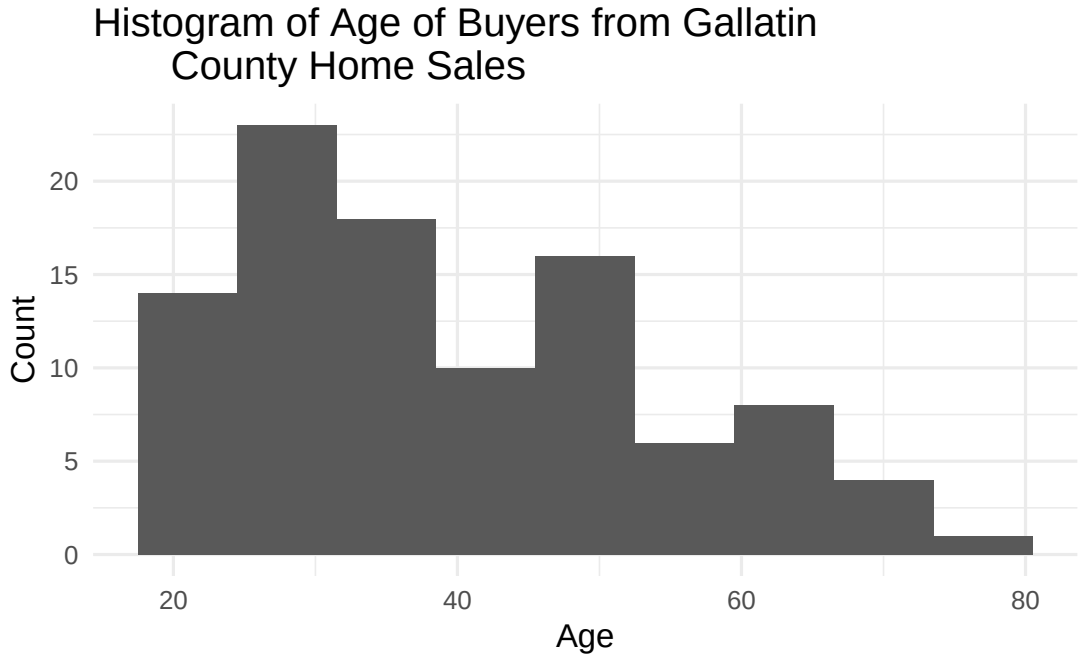
```
moving %>%  # Data set piped into...
  ggplot(aes(y = Age, x = From))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of Age by State of Origin
  of Buyers from Gallatin County Home Sales",  # Title
       x = "State of Origin",     # x-axis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



- 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

Let's review the summary statistics and histogram of age of buyers from sampled 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 _____ measure of center.
- The _____ is a _____ measure of spread.
- Robust means not _____ by.

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.

3.2 Out-of-Class Activity Module 3: Summarizing Categorical Variables

3.2.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question involving categorical variables.
- Plots for a single categorical variable: bar plot.
- Plots for association between two categorical variables: segmented bar plot, mosaic plot.

3.2.2 Terminology review

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

- Proportions
- Bar plots
- Segmented bar plots
- Mosaic plots

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

3.2.3 Graphing categorical variables

For this out-of-class activity we will walk through how to use the statistical package R to analyze data through the IDE (integrated development environment) RStudio. Even though the completed code is provided for you in this activity, we recommend that you login to the RStudio server and follow along to see how to create the plots and get the summary statistics for categorical data.

For almost all activities and labs it will be necessary to upload the provided R script file from D2L for that day. Follow these steps to upload the necessary R script file for this activity:

- Download the Myopia 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 Myopia 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 three lines 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. 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

Throughout these 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 6 in the R script file) reads in the data set and names the data set `myopia`.

- Highlight and run line 6 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")
```

- 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!

There are two variables in this data set. `Light` with three possible outcomes: `Full Light`, `Nightlight`, and `No Light` and `Sight` with three possible outcomes: `High Myopia`, `Myopia`, and `No Myopia`.

Displaying a single categorical variable

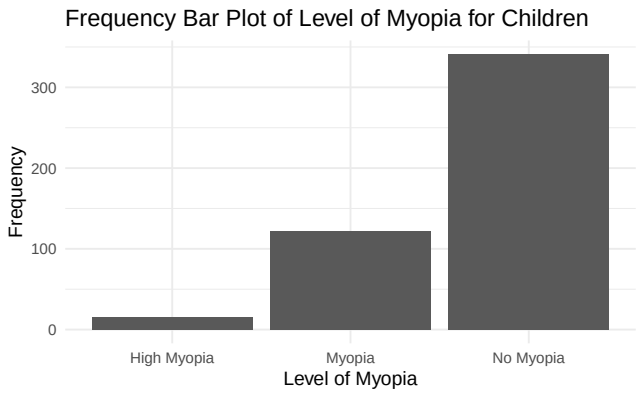
If we wanted to know how many children in our data set were in each level of myopia, we could create a frequency bar plot of the variable `Sight`.

- In the R code below (and the provided R script file), we will enter the variable name, `Sight` (*note the capital S*), for `variable` into the `ggplot` code. This is in line 12 in the R script file.
- Highlight and run lines 11–16 to create the plot. Note: this is a **frequency** bar plot plotting counts (the number of children in each level of sight is displayed on the y-axis).

```
myopia %>% # Data set piped into...
ggplot(aes(x = varible)) +   # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia for Children",
       # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Frequency")  # Label the y axis
```

Frequency bar plot of sight:

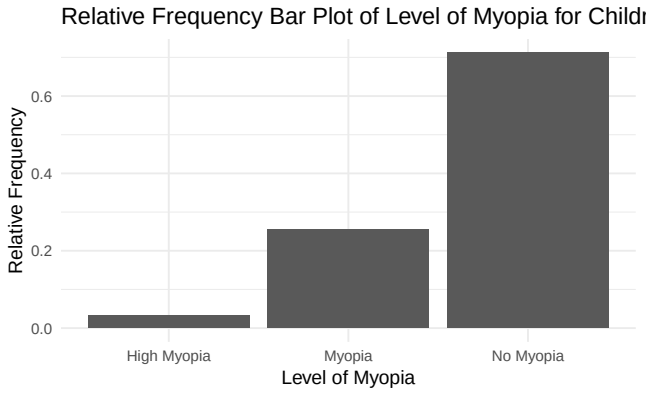
```
myopia %>% # Data set piped into...
ggplot(aes(x = Sight)) +   # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia for Children",
       # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Frequency")  # Label the y axis
```



2. Using the bar chart created, estimate how many children have some level of myopia.

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 myopia is divided by the sample size. This calculation is the sample proportion for each level of myopia. Notice that in this code we told R to create a bar plot with proportions.

```
myopia %>% # Data set piped into...
ggplot(aes(x = Sight)) +   # 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 Level of Myopia for Children",  # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



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

Displaying 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 29 in the R script file.
- Highlight and run line 29 to get the counts for each combination of levels of variables.

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

Here is the completed code and R output:

```
myopia %>% group_by(Light) %>% count(Sight) %>% print(n=9)
#> # A tibble: 9 x 3
#> # Groups:   Light [3]
#>   Light      Sight           n
#>   <chr>      <chr>       <int>
#> 1 Full Light High Myopia     5
#> 2 Full Light Myopia         42
#> 3 Full Light No Myopia      38
#> 4 Nightlight High Myopia     9
#> 5 Nightlight Myopia         65
#> 6 Nightlight No Myopia     149
#> 7 No Light   High Myopia     2
#> 8 No Light   Myopia         15
#> 9 No Light   No Myopia     154
```

4. 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.

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

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

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

8. 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.

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

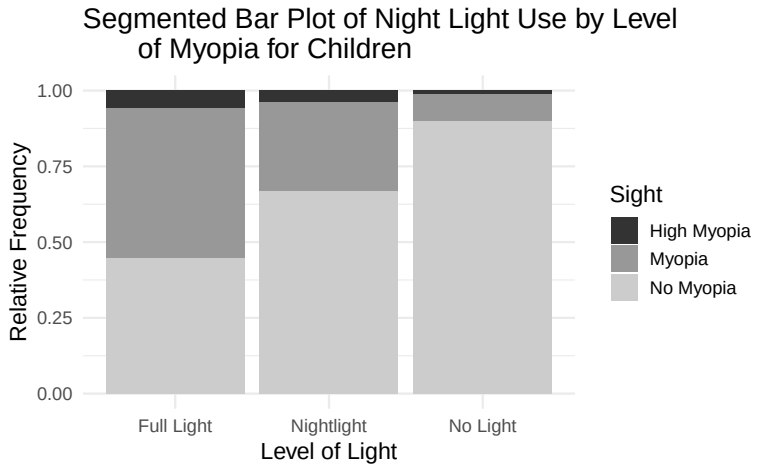
Two types of plots can be created to display two categorical variables. To examine the differences in level of myopia for the level of light, we will first create a segmented bar plot of `Light` segmented by `Sight`.

- To create the segmented bar plot enter the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in line 35 in the R script file.
- Highlight and run lines 34–40.

```
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 for Children",  # Make sure to title your plot
       x = "Level of Light",   # x axis label
       y = "Relative Frequency")  + # y axis label
  scale_fill_viridis_d()  # Make figure color
```

Segmented bar plot of Night Light Use by Level of Myopia:

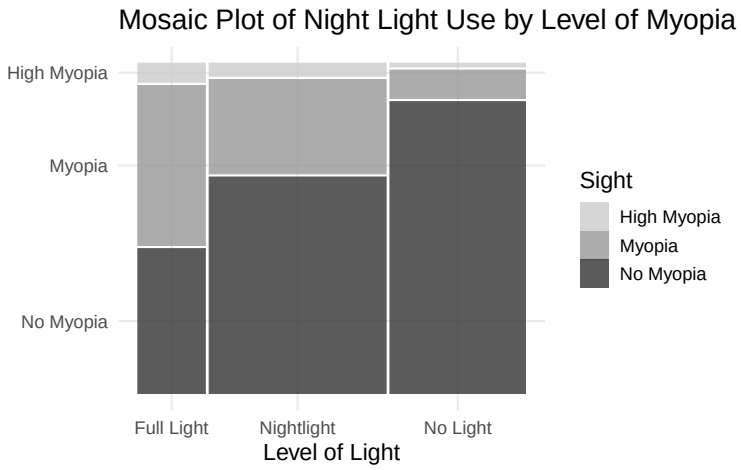
```
myopia %>% # Data set piped into...
ggplot(aes(x = Light, fill = Sight)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Night Light Use by Level
       of Myopia for Children",  # Make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



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

We could also create a mosaic plot of the data, 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 color
```



11. What is similar and what is different between the segmented bar chart and the mosaic bar chart?

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

3.2.4 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. This allows assessment of the relative sizes of the levels of one variable in assessing changes in relative distributions of the levels of the other variable and the proportions of the totals in each combination of levels.

3.2.5 Additional notes

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

3.3 Activity 3: IMDb Movie Reviews — Displaying Quantitative Variables

3.3.1 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.

3.3.2 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)
- Types of graphs: box plots, dot plots, histograms
- Identify and create appropriate summary statistics and plots given a data set or research question for a single categorical and a single quantitative variable.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

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

3.3.3 Movies released in 2016

A data set was collected on movies released in 2016 ("IMDb Movies Extensive Dataset" 2016). Here is a list of some of the variables collected on the observational units, movies released in 2016.

Variable	Description
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

Summarizing a single quantitative variable

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 `imdb_score`. 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 IMDb score.

- Highlight and run lines 1 – 9 in the provided `R` script file to load the data set. Check that the summary statistics match the output given in the coursepack.

```
# Read in data set
movies <- read.csv("https://math.montana.edu/courses/s216/data/Movies2016.csv")
movies %>% # Data set piped into...
  summarise(favstats(imdb_score)) # Apply favstats function to imdb_score
```

```
#>   min   Q1 median  Q3 max     mean       sd  n missing
#> 1 3.4 5.65    6.4 7.1 8.2 6.309783 1.086689 92       0
```

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

2. Calculate the interquartile range (IQR = Q3 − Q1) of IMDb scores.

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

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 IMDb scores would plot a dot for the IMDb score for each movie released in 2016.

We will create both a histogram and a boxplot of the variable `IMDb`.

- Enter the name of the variable in both line 16 and line 23 for `variable` in the R script file.
- Replace the word title for each plot (lines 18 and 25) 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 15 – 27 to create a histogram and boxplot.

Notice that the **bin width** for the histogram is 0.5. For example the first bin consists of the number of movies in the data set with an IMDb score of 3.25 to 3.75. It is important to note that a movie with a IMDb score on the boundary of a bin will fall into the bin above it; for example, 4.75 would be counted in the bin 4.75–5.25.

```
movies %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 0.5) +  # Create histogram with specified binwidth
  labs(title = "Title", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

```r
movies %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_boxplot() +  # Create histogram with specified binwidth
  labs(title = "Title", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "") # Remove y axis label
```

4. What is the shape of the distribution of IMDb scores?

5. Which range of IMDb scores have the highest frequency?

6. Sketch the boxplot created and identify the values of the 5-number summary (minimum value, Q1, median, Q3, 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 found.

$$\text{Lower Fence: values} \leq \text{Q1} - 1.5 \times \text{IQR}$$

$$\text{Upper Fence: values} \geq \text{Q3} + 1.5 \times \text{IQR}$$

Summary statistics for a single categorical and single quantitative Variable

Is there an association between content rating and budget for movies released in 2016? To use the `favstats()` function in the mosaic package with two variables, we will enter the variables as a formula, response~explanatory. This function will give the summary statistics for budget for each content rating.

- Highlight and run lines 31–33 in the provided `R` script file and check that the summary statistics match those provided in the coursepack.

```r
movies %>% # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  reframe(favstats(budget_mil~content_rating)) # Find the summary measures for each content rating
```

```
#>   content_rating min    Q1 median      Q3 max     mean       sd  n missing
#> 1             PG 0.5 11.00   74.0 151.250 175 86.54167 71.52795 12       0
#> 2          PG-13 0.0 17.25   33.5 138.750 250 74.17500 74.15190 46       0
#> 3              R 0.0  7.75   19.5  29.625  60 21.09375 16.99926 32       0
```

7. Which content rating has the largest IQR?

8. Report the mean budget amount for the PG rating. Use appropriate notation.

9. Report the mean budget amount for the R rating. Use appropriate notation.

10. Calculate the difference in mean budget amount for movies released in 2016 with a PG rating minus those with a R rating. Use appropriate notation with informative subscripts.

11. Interpret the difference in means calculated in question 10 in context of the problem.

Displaying a single categorical and single quantitative variable

The boxplot of movie budgets (in millions) by content rating is plotted using the code below.

- Enter the variable `budget_mil` for `response` and the variable `content_rating` for explanatory at line 40.
- Highlight and run code lines 38–44. This plot compares the budget for different levels of content rating.

```
movies %>%  # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side-by-side Box Plot of Budget by Content Rating for Movies Released in 2016",
       # Title
       x = "Content Rating",    # x-axis label
       y = "Budget (in Millions)")  # y-axis label
```

12. Sketch the box plots created using the `R` code.

13. Answer the following questions about the box plots created.

    a. Which content rating has the highest center?

    b. Which content rating has the largest spread?

    c. Which content rating has the most skewed distribution?

    d. Fifty percent of movies released in 2016 with a PG-13 content rating fall below what value? What is the name of this value?

    e. What is the value for the first quartile (Q1) for the PG-13 rating? Interpret this value in context.

3.3.4 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. Values in the data set that are less than $Q_1 - 1.5 \times IQR$ and 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. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, IQR = Q3 − Q1.

3.3.5 Additional notes

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

3.4 Module 3 Lab: IPEDs

3.4.1 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.

3.4.2 Terminology review

In today's lab, 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)
- Types of graphs: box plots, dot plots, histograms
- Identify and create appropriate summary statistics and plots given a data set or research question for a single categorical and a single quantitative variable.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

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

3.4.3 The Integrated Postsecondary Education Data System (IPEDS)

Upload and open the provided R script file for the week 3 lab to answer the following questions. **Remember bolded questions will be answered on Gradescope for your group.**

These data are 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
- Note that several variables have missing values for some institutions (denoted by "NA").

Variable Name	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Sector	• Public 2-year • Private 2-year • Public 4-year or higher • Private 4-year or higher
LandGrant	Is this a land-grant institution? (Yes/No)
Size	Institution size category based on total students 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, first-time degree/certificate seeking out-of-state undergraduate students living on campus for academic year 2018-2019. It includes out-of-state tuition and fees, books and supplies, on campus room and board, and other campus expenses.
Cost_InState	Cost of attendance for full-time, first-time degree/certificate seeking in-state undergraduate students living on campus for academic year 2018-2019. It includes in-state tuition and fees, books and supplies, on campus room and board, and other campus expenses.
Retention	The full-time retention rate is the percent of the (fall full-time cohort from the prior year minus exclusions from the fall full-time cohort) that re-enrolled at the institution as either full- or part-time in the current year.
Percent_InState	Percent of first-time degree/certificate seeking undergraduate students who reside in the same state of the institution.
Enrollment	Total number of people enrolled for credit in the fall of the academic year.
Graduation_Rate	Graduation rate of first-time, full-time degree or certificate-seeking students – 2012 cohort (4-year institutions) and 2015 cohort (less-than-4-year institutions). This rate is calculated as the total number of completers within 150% of normal time divided by the revised cohort minus any allowable exclusions.
Percent_FinancialAid	Percentage of all full-time, first-time degree/certificate-seeking undergraduate students who were awarded any financial aid.

Summary statistics for a single quantitative variable

Look through the provided chart above showing the description of variables measured. The UnitID and Name are identifiers for each observational unit, *US degree granting institutions in 2018*.

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

In the previous activities this week, the code was provided to import the data set needed directly from the Stat 216 website. Follow these steps to upload and import the data set for today's lab.

- Download the provided data set `IPED_Data_2018` from D2L
- Upload the data set `IPEDS_Data_2018` to the RStudio server using the same steps to upload the R script file.
- Click on "Import Dataset" in the Environment tab in the upper right hand corner.
- Choose "From Text(base)" in the drop-down menu and select the correct csv file.
- Be sure that "Yes" is selected next to "Heading" in the pop-up screen. Click "Import".
- To view the data set, click on the data set name (`IPEDS_Data_2018`). Verify that that column names match the first column in the chart on the previous page. If the columns are named V1, V2, V3…etc, you did not select "Yes" for "Heading".

Teams will also need the R script file for this week's lab.

- Download the provided R script file from D2L
- Upload the R script file to the RStudio server
- Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 9.
- We will look at the retention rates for the 4-year institutions only. Enter the variable name `Retention` for `variable` in line 15.
- Highlight and run lines 1 – 15. **Note that the two lines of code (lines 10 and 12) are filtering to remove the 2-year institutions so we are only assessing Public 4-year and Private 4-year institutions.**

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. The summary statistics give the two measures of center and two measures of spread for retention rates for 4-year institutions.

```
IPEDS <- datasetname #Creates the object IPEDS
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 %>%
  summarise(favstats(variable)) #Gives the summary statistics
```

2. Identify the observational units for this study after removing the 2-year institutions.

3. **Report the value for quartile 3 and interpret this value in context of the study.**

4. Report and interpret the value of the standard deviation.

5. How many missing values are there? What does this indicate?

Robust Statistics

Let's examine how the presence of outliers affect the values of center and spread.

6. Report the two measures of center (mean and median) for retention rates given in the R output.

7. Report the value of the standard deviation and calculate the value of the IQR (two measures of spread) for retention rates from the R output.

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%.

- Highlight and run lines 19–20 in the R script file.

```
IPEDS %>% # Data set piped into...
  summarise(favstats(Retention_Inc))
```

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

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

10. **Which measure of center is robust to (not affected by) outliers? Explain your answer.**

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

Summarizing a single categorical and single quantitative variable

Is there a difference in retention rates for public and private 4-year institutions? In the next part of the activity we will compare retention rates for public and private 4-year institutions. Note that this variable (public or private) is labelled `Sector` in the data set.

12. **Based on the research question, which variable will we treat as the explanatory variable? Response variable?**

- To assess the research question described before question 12, enter the name of the explanatory variable and the name of the response variable in lines 28 and 31 of the R script file. Remember that the variable name must be typed in EXACTLY as it is written in the data set.
- Replace the word title (line 33) 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 27 – 35 to find the summary statistics and create side by side boxplots of the data.
- **Export and upload the side-by-side box plot to Gradescope for your group.**
  - To export the graph: in the bottom right corner in the Plots tab, click on `Export`
  - Choose `Save as Image.` Save the image as a png. This will save your graph to the server.
  - In the `Files` tab, click on the box next to your saved image file, click `More` and choose `Export`. This will save your file to your downloads folder on your computer.

```
IPEDS %>%  # Data set piped into...
  reframe(favstats(response~explanatory)) # Summary statistics for retention rates by sector
```

```
IPEDS %>%  # Data set piped into...
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Create box plot
  labs(title = "title",  # Give your plot a title
       x = "Sector",    # x-axis label
       y = "Retention Rates")  # y-axis label
```

13. **Compare the two boxplots.**

    Which type of university has the highest center?

    Largest spread?

    What is the shape of each distribution?

    Does either distribution have potential outliers?

14. Report the difference in mean retention rates for private and public universities. Use private minus public as the order of subtraction. Use the appropriate notation.

15. Does there appear to be an association between retention rates and type of university? Explain your answer using the boxplots and summary statistic.

Summarizing two categorical variables

Are private 4-year institutions smaller than public one? The following set of code will create a segmented bar plot of size of the institution by sector.

- Enter the variable `Sector` for explanatory and `Size` for response in line 41.
- Highlight and run lines 40 - 46 in the R script file.

```
IPEDS %>%
  ggplot(aes(x=explanatory, fill = response)) + # Enter the explanatory and response variables
  geom_bar(stat = "count", position = "fill") + # Create a segmented bar plot
  labs(title = "Segmented Bar Plot of Sector by Size for
       4-year Institutions", # Title
       x = "Sector", # x-axis label
       y = "Relative Frequency") # y-axis label
```

16. Does there appear to be an association between sector and size of 4-year institutions? Explain your answer using the plot.

MODULE 4

Exploring Multivariable Data

4.1 Lecture Notes Module 4: Regression and Correlation

Summary measures and plots for two quantitative variables

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 summary measures for two quantitative variables are:

- _____
- _____
- _____

- 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 $\widehat{\text{response}}$ is
- b_0 is
- b_1 is
- x or explanatory is
- 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 =

  - To find SSE, the _____ for each data point is found, squared and all the squared residuals are summed together

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 coefficient of determination can be found by squaring the value of correlation, using the _____ 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.

Notation:

- Population slope:
- Population correlation:
- Sample slope:

- Sample correlation:

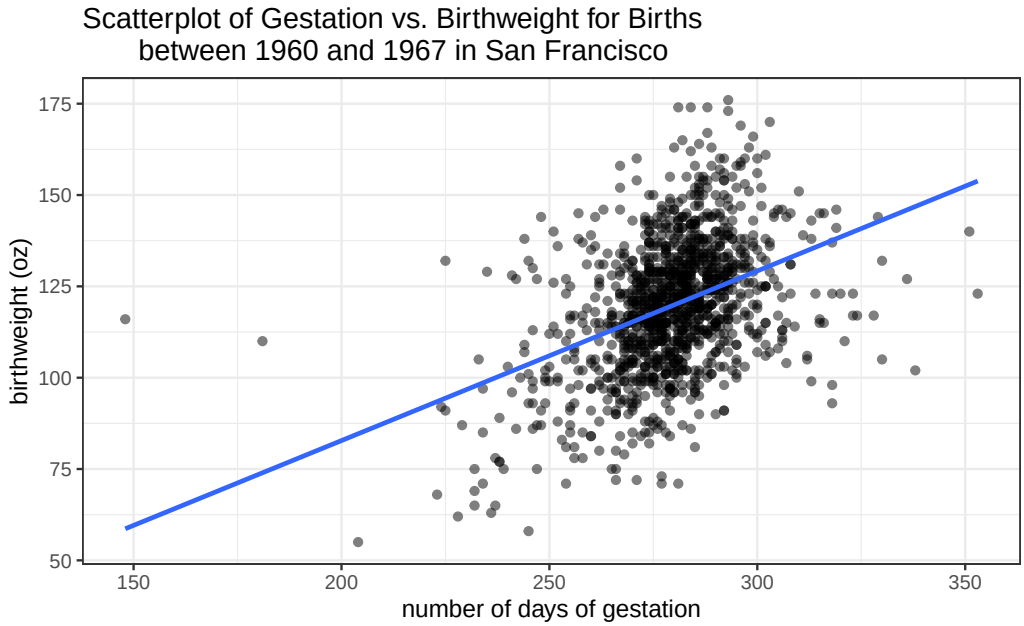
Example for class discussion: Data were collected from 1236 births between 1960 and 1967 in the San Francisco East Bay area to better understand what variables contributed to child birthweight, as children with low birthweight often suffer from an array of complications later in life ("Child Health and 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.

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()
```



Describe the scatterplot using the four characteristics of a scatterplot.

The linear model output for this study is given below:

```
# Fit linear model: y ~ x
babiesLM <- lm(bwt ~ gestation, data=babies)
summary(babiesLM)$coefficients # Display coefficient summary
#>                Estimate Std. Error   t value     Pr(>|t|)
#> (Intercept) -10.0641842 8.32220357 -1.209317 2.267751e-01
#> gestation     0.4642626 0.02974366 15.608793 3.224362e-50
```

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.

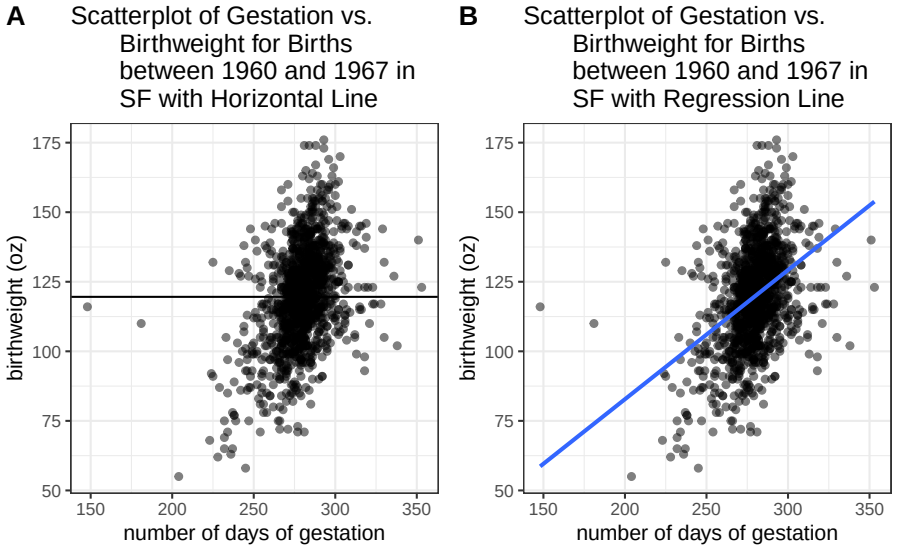
Calculate the residual for a birth of a baby with a birthweight of 151 ounces and born at 310 days gestation.

Is this value (310, 151) above or below the line of regression? Did the line of regression overestimate or underestimate the birthweight?

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.



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.

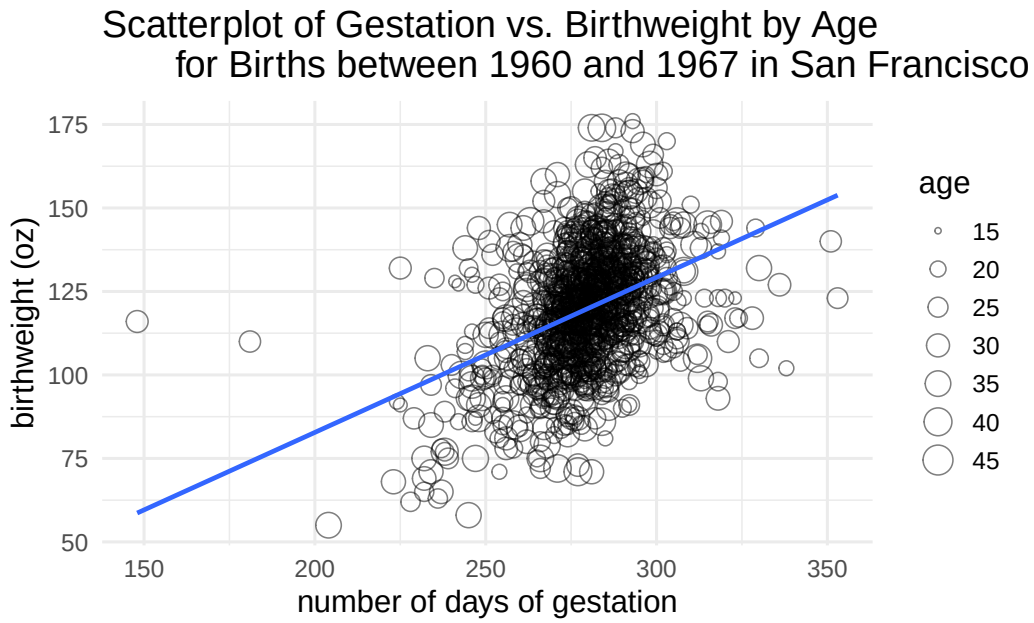
Multivariable plots

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.

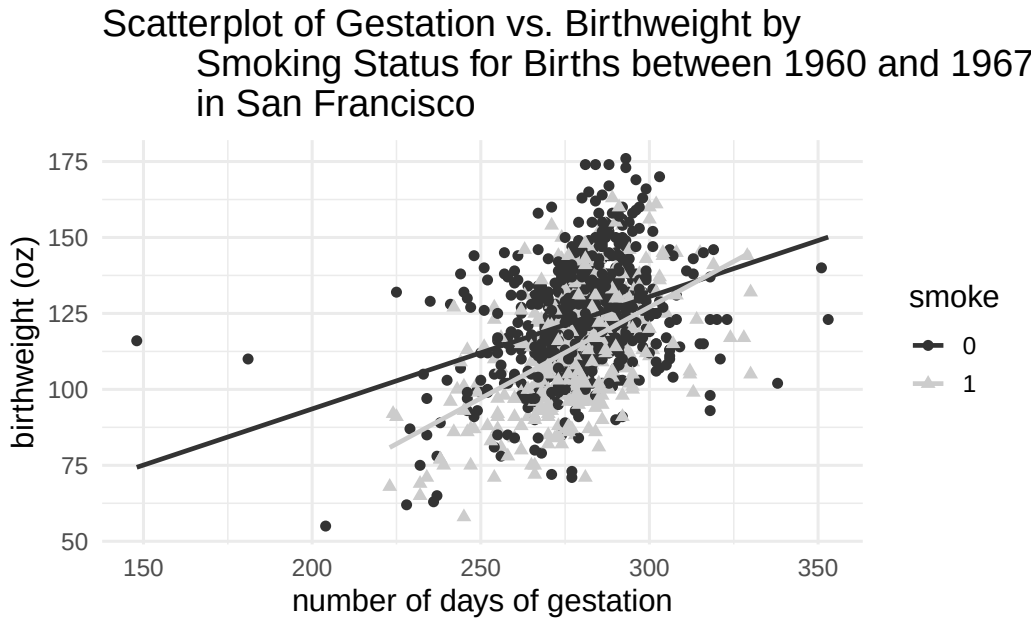
```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+  # Specify variables
  geom_point(alpha=0.5, shape=1, aes(size=age)) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot 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
```



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.

4.2 Out-of-Class Activity Module 4: Movie Profits — Correlation and Coefficient of Determination

4.2.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- 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.

4.2.2 Terminology review

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

- Correlation (r)
- Coefficient of determination (r-squared)

To review these concepts, see Chapter 6 in the textbook.

4.2.3 Movies released in 2016

A data set was collected on movies released in 2016 ("IMDb Movies Extensive Dataset" 2016). Here is a list of some of the variables collected on the observational units, movies released in 2016. (Note: both budget and revenue are measured in "millions of dollars" ($MM).)

Variable	Description
`budget_mil`	Amount of money ($MM) budgeted for the production of the movie
`revenue_mil`	Amount of money ($MM) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

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.

```
movies <- read.csv("https://math.montana.edu/courses/s216/data/Movies2016.csv") # Reads in data set
movies %>%  # Data set pipes into
  select(c("budget_mil", "revenue_mil",
           "duration", "imdb_score",
           "facebook_likes")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

```
#>                budget_mil revenue_mil duration imdb_score facebook_likes
#> budget_mil          1.000       0.686    0.463      0.292          0.678
#> revenue_mil         0.686       1.000    0.227      0.398          0.723
#> duration            0.463       0.227    1.000      0.261          0.438
#> imdb_score          0.292       0.398    0.261      1.000          0.309
#> facebook_likes      0.678       0.723    0.438      0.309          1.000
```

1. Explain why the correlation values on the diagonal are equal to 1.

2. Using the output above, ignoring the values of 1, which pair of variables have the *strongest* correlation? What is the value of this correlation?

3. What is the value of correlation between budget and revenue?

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: $r^2 = (r)^2$

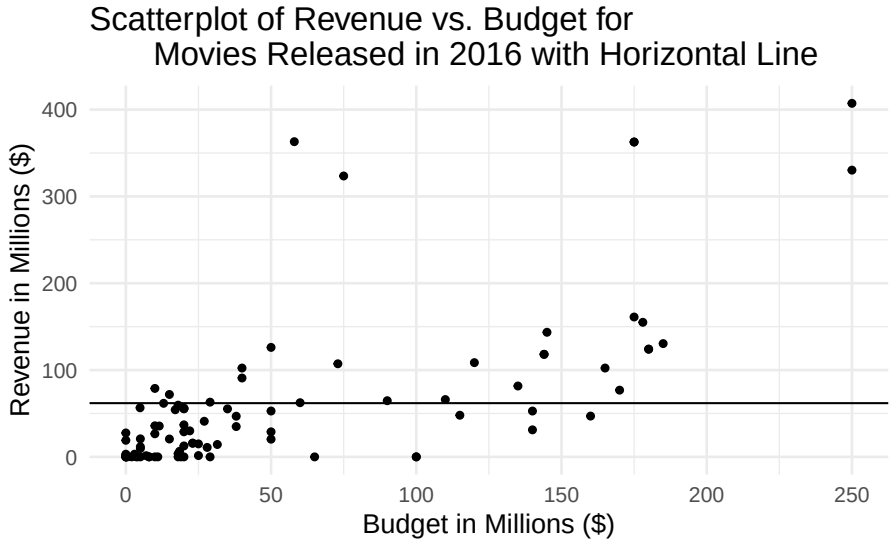
Use the variances of the response and the residuals: $r^2 = \dfrac{s_y^2 - s_{RES}^2}{s_y^2} = \dfrac{SST - SSE}{SST}$

4. Use the correlation, r, found in question 3 of the activity, to calculate the coefficient of determination between budget and revenue, r^2.

5. The variance of the response variable, revenue in \$MM, is about $s^2_{revenue} = 8024.261$ \$MM^2 and the variability in the residuals is about $s^2_{RES} = 4244.832$ \$MM^2. Use these values to calculate the coefficient of determination. Verify that your answers to 4 and 5 are the same.

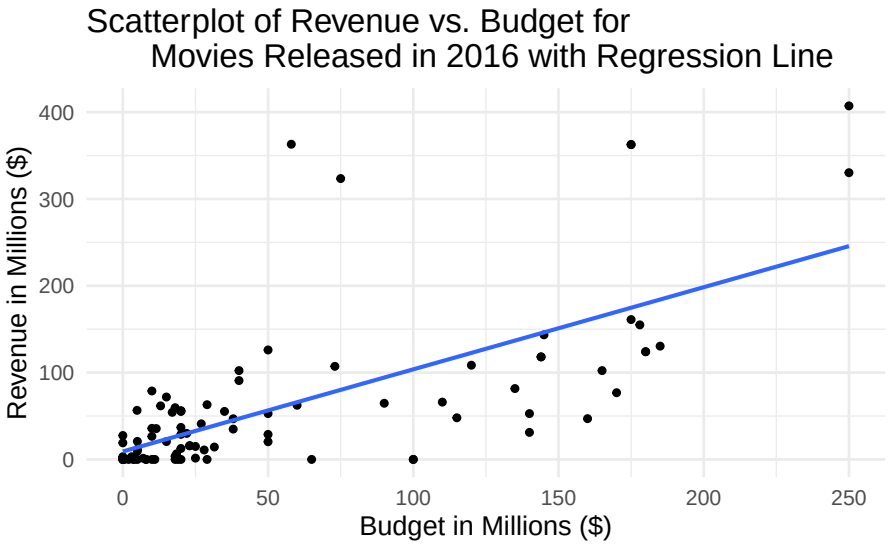
In the next part of the activity we will explore what the coefficient of determination measures.

In the scatterplot below, 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 budget and revenue. The value of the y-intercept, 61.87, 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 61.87). Each residual is squared and the sum of the squared values is calculated. The SST gives the **total variability in the response variable, revenue**.



The calculated value for the SST is 730207.72.

This next scatterplot, shows the plotted data with the best fit regression line. We will learn more about the regression line in the next class. 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.



The calculated value for the SSE is 386279.71.

6. Calculate the value for r^2 using the values for SST and SSE provided below each of the previous graphs.

7. Write a sentence interpreting the coefficient of determination in context of the problem.

4.2.4 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.

4.2.5 Additional notes

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

4.3 Activity 4: Movie Profits — Linear Regression

4.3.1 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.

4.3.2 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.

4.3.3 Movies released in 2016

We will revisit the movie data set collected on Movies released in 2016 ("IMDb Movies Extensive Dataset" 2016) to further explore the relationship between budget and revenue. Here is a reminder of the variables collected on these movies. (Note: both budget and revenue are measured in "millions of dollars" ($MM).)

Variable	Description
`budget_mil`	Amount of money ($MM) budgeted for the production of the movie
`revenue_mil`	Amount of money ($MM) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

```
movies <- read.csv("https://math.montana.edu/courses/s216/data/Movies2016.csv") # Reads in data set
```

Vocabulary review

To look at the relationship between two quantitative variables we will create a scatterplot with the explanatory variable on the x-axis and the response variable on the y-axis.

We will look at the relationship between budget and revenue for movies released in 2016.

- Upload and open the provided R script file.
- Enter the explanatory variable name, `budget_mil`, for `explanatory` and the response variable name, `revenue_mil`, for `response` at line 9 in the R script file to create the scatterplot.
- Highlight and run lines 1–14.

```
movies %>% # Data set pipes into...
ggplot(aes(x = explanatory, y = response))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Budget in Millions ($)",  # Label x-axis
       y = "Revenue in Millions ($)",  # Label y-axis
       title = "Scatterplot of Revenue vs. Budget for Movies
       Released in 2016") +
    # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

1. Sketch the scatterplot created from the code.

2. Assess the four features of the scatterplot that describe this relationship.

- Form (linear, non-linear)
- Direction (positive, negative)
- Strength
- Unusual observations or outliers

3. Based on the plot, does there appear to be an association between budget and revenue? Explain.

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 18–19 in the R script file to reproduce the linear model output found in the coursepack.

```
# Fit linear model: y ~ x
revenueLM <- lm(revenue_mil ~ budget_mil, data=movies)
summary(revenueLM)$coefficients # Display coefficient summary
```

```
#>              Estimate Std. Error  t value     Pr(>|t|)
#> (Intercept) 9.1693054  9.0175499 1.016829 3.119606e-01
#> budget_mil  0.9460001  0.1056786 8.951670 4.339561e-14
```

4. Write out the least squares regression line using the summary statistics provided above in context of the problem.

You may remember from middle and high school that slope $= \frac{\text{rise}}{\text{run}}$.

Using b_1 to represent slope, we can write that as the fraction $\frac{b_1}{1}$.

Therefore, the slope predicts how much the line will *rise* for each *run* of +1. In other words, as the x variable increases by 1 unit, the y variable is predicted to change (increase/decrease) by the value of slope.

5. Interpret the value of slope in context of the problem.

6. Using the least squares line from question 4, predict the revenue for a movie with a budget of 165 \$MM.

7. Predict the revenue for a movie with a budget of 500 \$MM.

8. The prediction in question 7 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.

$$\implies \text{Residual} = \text{actual y value} - \text{predicted y value}$$

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

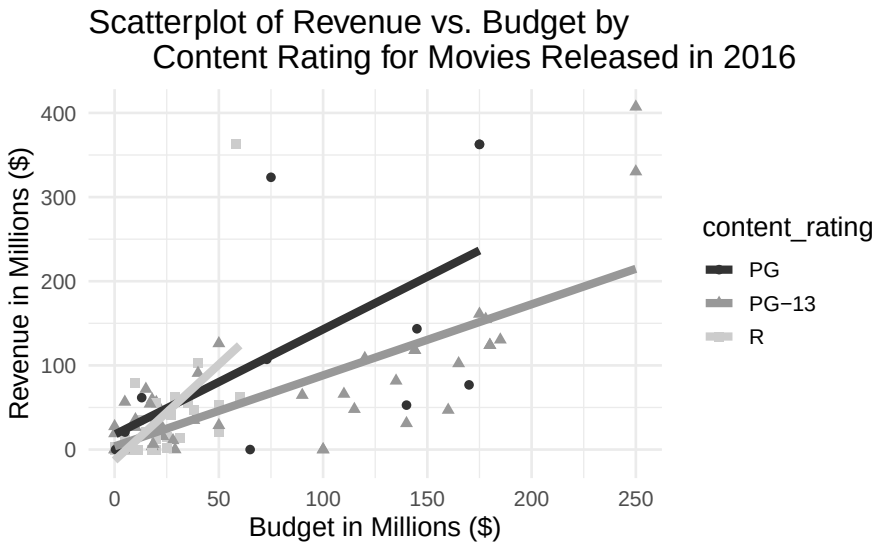
9. The movie *Independence Day: Resurgence* had a budget of 165 \$MM and revenue of 102.315 \$MM. Find the residual for this movie.

10. Did the line of regression overestimate or underestimate the revenue for this movie?

Multivariable plots

What if we wanted to see if the relationship between movie budget and revenue differs if we add another variable into the picture? The following plot visualizes three variables, creating a **multivariable** plot.

```
movies %>% # Data set pipes into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(x = budget_mil, y = revenue_mil, color = content_rating)) +  # Specify variables
  geom_point(aes(shape = content_rating), size = 2) +  # Add scatterplot of points
  labs(x = "Budget in Millions ($)",  # Label x-axis
       y = "Revenue in Millions ($)",  # Label y-axis
       color = "content_rating",  # Label legend
       title = "Scatterplot of Revenue vs. Budget by
       Content Rating for Movies Released in 2016") +
    # Be sure to title your plot
  geom_smooth(method = "lm", se = FALSE, lwd = 2) + # Add regression lines
  scale_color_grey() # Make black and white
```



11. Identify the three variables plotted in this graph.

12. Does the *relationship* between movie budget and revenue differ among the different content ratings? Explain.

4.3.4 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).
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**.

4.3.5 Additional notes

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

4.4 Module 4 Lab: Penguins

4.4.1 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.
- 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.

Penguins

The Palmer Station Long Term Ecological Research Program sampled three penguin species on islands in the Palmer Archipelago in Antarctica. Researchers took various body measurements on the penguins, including bill depth and body mass. The researchers were interested in the relationship between bill depth and body mass and wondered if bill depth could be used to accurately predict the body mass of these three penguin species.

- Upload and import the `Antarctica_Penguins` csv file
- Upload and open the provided R script file for week 4 lab.
- Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 5.

First we will create a scatterplot of the bill depth and body mass. Notice that we are using bill depth (mm) to predict body mass (g). This makes bill depth the explanatory variable.

- **Make sure to give your plot a descriptive title between the quotations in line 16. Remember that the title should include the type of plot, the observational units, and the variable(s) plotted.**
- Highlight and run lines 1–17 in the R script file.
- **Upload a copy of your scatterplot to Gradescope.**

```
penguins <- datasetname %>% #Creates the object penguins
    na.omit() #Removes data points without values
penguins %>%
  ggplot(aes(x = bill_depth_mm, y = body_mass_g))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "bill depth (mm)",  # Label x-axis
       y = "body mass (g)",  # Label y-axis
       title = "Title") + # Be sure to title your plot
  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

To create the correlation matrix…

- Highlight and run lines 20–24 in the R script file.

```
penguins %>%  # Data set pipes into
  select(c("bill_length_mm", "bill_depth_mm",
           "flipper_length_mm", "body_mass_g")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

2. Using the R output, report the value of correlation between bill depth and body mass.

3. Using the value of correlation found in question 2, calculate the value of the coefficient of determination.

4. **Interpret the coefficient of determination in context of the problem.**

To get the linear model output…

- Enter the variable name `body_mass_g` for `response` and the variable name `bill_depth_mm` for `explanatory` in line 29 in the R script file.
- Highlight and run lines 29–30.

```
# Fit linear model: y ~ x
penguinsLM <- lm(response~explanatory, data=penguins)
summary(penguinsLM)$coefficients # Display coefficient summary
```

5. Write out the least squares regression line using the summary statistics from the R output in context of the problem.

6. **Interpret the value of slope in context of the problem.**

7. **Using the least squares regression line from question 5, predict the body mass for a penguin with a bill depth of 19.6 mm.**

8. One penguin had a bill depth of 19.6 mm and a body mass of 4675 g. Find the residual for this penguin.

9. Did the line of regression overestimate or underestimate the body mass for this penguin?

Does species change the relationship between bill depth and body mass?

- Highlight and run lines 34 - 43 to get the multivariable plot.

```
penguins %>%
  ggplot(aes(x = bill_depth_mm, y = body_mass_g, color=species))+  # Specify variables
  geom_point(aes(shape = species), size = 2, alpha=0.5) +  # Add scatterplot of points
  labs(x = "bill depth (mm)",  # Label x-axis
       y = "body mass (g)",  # Label y-axis
       color = "species",
       shape = "species",
       title = "Scatterplot of Bill Depth and Body Mass by Penguin Species") +
    # Enter the title for the plot between the quotations
  geom_smooth(method = "lm", se = FALSE) +  # Add regression line
  scale_color_viridis_d(end=0.8)
```

10. What three variables are plotted on this plot?

11. **Do species and bill depth appear to interact when predicting body mass of Antarctic penguins? Explain your answer.**

12. Explain the association between species and each of the other two variables.

13. Notice that the slope of the line between bill depth and body mass for each species is positive while the slope for the line not accounting for species is negative. What phenomena is this an example of?

MODULE 5

Group Exam 1 Review

Use the provided data set from the Islands (Bulmer, n.d.) (Exam1ReviewData.csv) and the appropriate Exam 1 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
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. What are the observational units?

2. In the table above, indicate which variables are categorical (C) and which variables are quantitative (Q).

3. What type of bias may be present in this study? Explain.

Complete questions 4a, 4b, 5a, 5b, 6a, and 6b. Then choose the scenario for each research question and use the appropriate Exam 1 Review R script file to find the summary statistic(s) and graphical display of the research question.

4. Use the appropriate Exam 1 Review R script file to find the summary statistic and graphical display of the data to assess the following research question, "Is there a difference in proportion of Islanders who have children for those who completed high school and those that completed university?" Use high school − university as the order of subtraction.

a. What is the name of the explanatory variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

b. What is the name of the response variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

c. 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			

d. Calculate the value of the summary statistic to answer the research question. Give appropriate notation.

e. Interpret the value of the summary statistic in context of the problem:

f. What type of graph(s) would be appropriate for this research question?

g. Using the provided R file create a graph of the data. Sketch the graph below:

h. Does there appear to be an association between the two variables? Clearly explain your answer using the graph and calculated summary statistic.

i. Is this an observational study or a randomized experiment? Explain your answer.

j. What is the scope of inference for this study?

5. Use the appropriate Exam 1 Review R script file to find the appropriate summary statistic and graphical display of the data to assess the following research question: "Do Islanders who listen to classical music take less time 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. What is the name of the explanatory variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

b. What is the name of the response variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

c. 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 name, levels of the variable, and the summary statistics from the R output.

	Explanatory Variable	
Summary value	Group 1	Group 2
Mean		
Standard deviation		
Sample size		

d. Calculate the value of the summary statistic to answer the research question. Give appropriate notation.

e. Interpret the value of the summary statistic in context of the problem:

f. What type of graph(s) would be appropriate for this research question?

g. Using the provided R file create a graph of the data. Sketch the graph below:

h. Does there appear to be an association between the two variables? Clearly explain your answer using the graph and calculated summary statistic.

i. Compare the two plots using the four characteristics to describe plots of quantitative variables.

Shape:

Center:

Spread:

Outliers:

j. Is this an observational study or a randomized experiment? Explain your answer.

k. What is the scope of inference for this study?

6. Use the appropriate Exam 1 Review R script file to find the appropriate summary statistic and graphical display of the data to assess the following research question: "Do Islanders who are heavier tend to take more breaths per minute?"

a. What is the name of the explanatory variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

b. What is the name of the response variable to be assessed in this research question?

What type of variable (categorical or quantitative) is the variable you identified?

c. 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			

d. Interpret the value of slope in context of the problem.

e. Calculate the value of the coefficient of determination.

f. Interpret the coefficient of determination in context of the problem.

g. What type of graph(s) would be appropriate for this research question?

h. Using the provided R file create a graph of the data. Sketch the graph below:

i. Does there appear to be an association between the two variables? Clearly explain your answer using the graph and calculated summary statistic.

j. Describe the plot using the four characteristics to describe scatterplots.

Form:

Direction:

Strength:

Outliers:

k. Is this an observational study or a randomized experiment? Explain your answer.

l. What is the scope of inference for this study?

MODULE 6

Inference for a Single Categorical Variable: Simulation-based Methods

6.1 Lecture Notes Module 6: Inference for One Categorical Variable using Simulation-based Methods

Hypothesis Testing

Purpose of a hypothesis test:

- Use data collected on a _____ to give information about the _____
- 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.

Hypotheses

- Two possible outcomes:
  - Either the _____ hypothesis is true and the _____ occurred by _____ chance.
  - Or the null hypothesis is _____ and the sample provides _____ against the _____.
- 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

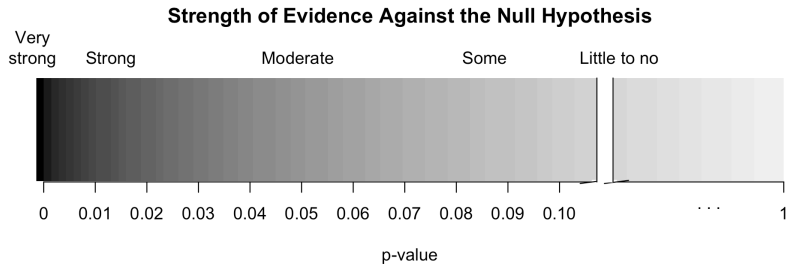
- Find the proportion of _____ at least as extreme as the observed sample _____
- The null distribution estimates the sample to sample _____ expected in the population

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 next week*

P-value

- What does the p-value measure?
  - Probability of observing the sample _____ or more _____ assuming the _____ hypothesis is _____.
- How much evidence does the p-value provide against the null hypothesis?



- The _____the p-value, the _____ the evidence against the null hypothesis.

- Write a conclusion based on the p-value.

    - Answers the _____ question.

    - Amount of _____ in support of the _____ hypothesis.

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

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

    - p-value $\leq \alpha$

        - Strong enough evidence against the null hypothesis

        - Decision:

        - Results are _____ significant.

    - p-value $> \alpha$

        - Not enough evidence against the null hypothesis

        - Decision:

        - Results are not _____ significant.

One proportion test

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

Notation:

- Population proportion:
- Sample proportion:

Parameter of Interest:

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

π :

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 alternative hypothesis.

Example for class discussion: 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 words:

H_0 :

H_A :

In notation:

H_0 :

H_A :

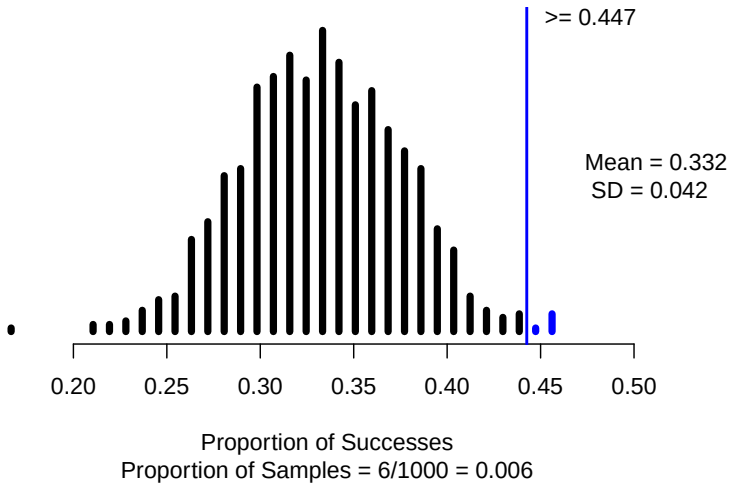
Statistic:

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

Simulation-based method

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

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



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

Interpretation of the p-value:

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

Conclusion:

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

Generalization:

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

Confidence interval

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

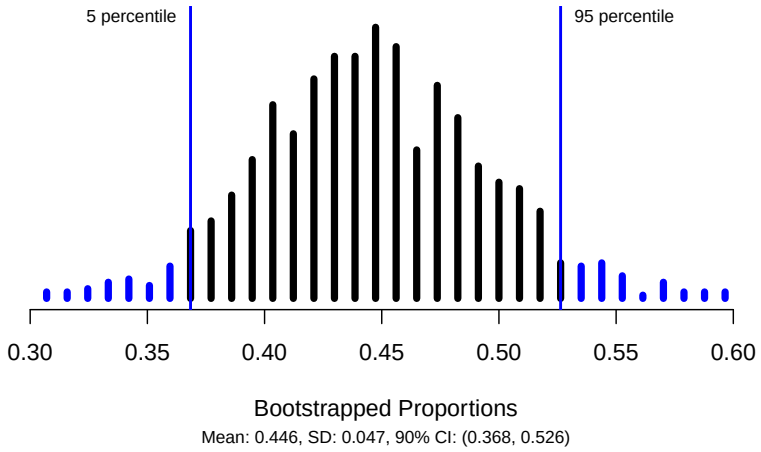
What is bootstrapping?

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

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

Bootstrap distribution:

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



Confidence interval interpretation:

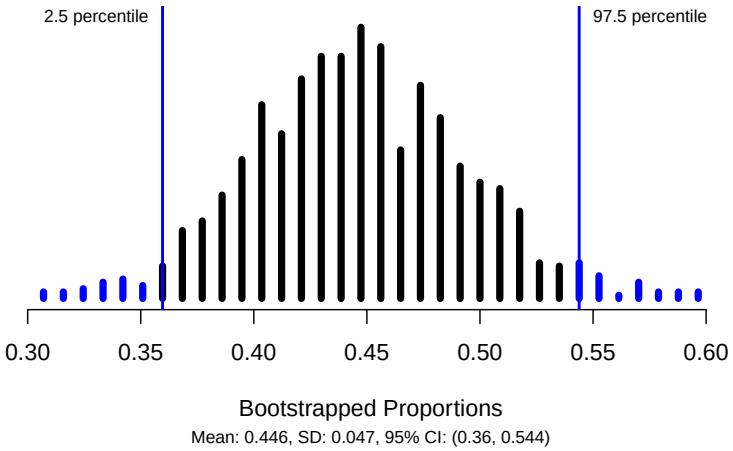
- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

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

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

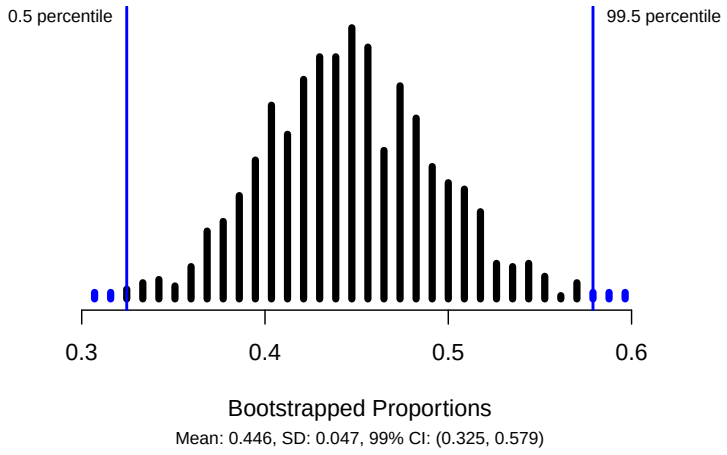
95% Confidence Interval:

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



99% Confidence Interval:

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



6.2 Out-of-Class Activity Module 6: Helper-Hinderer — Simulation-based Confidence Interval and Hypothesis Test

6.2.1 Learning outcomes

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

6.2.2 Terminology review

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

- Parameter of interest
- Bootstrapping
- Confidence interval
- Null hypothesis
- Alternative hypothesis
- Simulation

To review these concepts, see Chapters 9, 10 & 14 in your textbook.

6.2.3 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.

6.2.4 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 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.

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.

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.

1. Are the cases independent? Justify your answer.

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

```
 # 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
```

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

2. 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. This value is also called the **point estimate.**

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**.

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

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

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

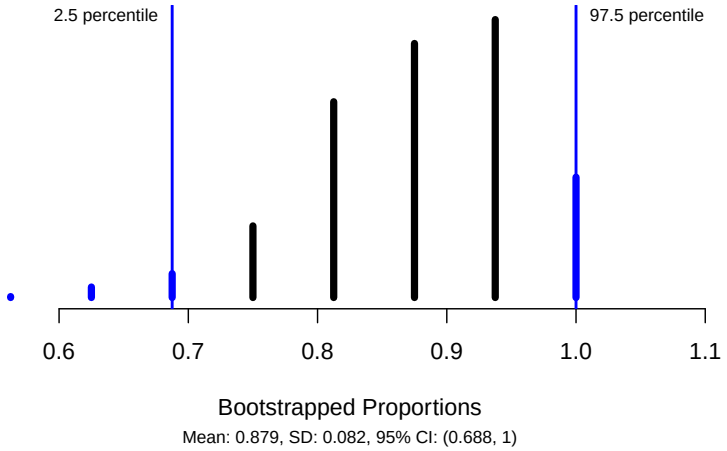
- "sample size" (the number of observational units or cases in the sample),
- "number of successes" (the number of cases that choose the helper character),
- "number of repetitions" (the number of samples to be generated), and
- the "confidence level" (which level of confidence are we using to create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

4. 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. Check that your answers to question 4 match what is entered in the R code.

```
set.seed(216)
one_proportion_bootstrap_CI(sample_size = 16, # Sample size
                    number_successes = 14, # Observed number of successes
                    number_repetitions = 1000, # Number of bootstrap samples to use
                    confidence_level = 0.95) # Confidence level as a decimal
```



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

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

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

8. Interpret the 95% confidence interval in context.

Use statistical analysis methods to draw inferences from the data

In the next part of the activity, we will perform a hypothesis test to assess the research question.

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

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.

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.

10. Using the parameter of interest given prior to question 3, 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, $H_0 : \pi = \pi_0$.

11. 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.

12. Based on the research question from question 9, are we testing that the parameter is greater than 0.5, less than 0.5 or different than 0.5?

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

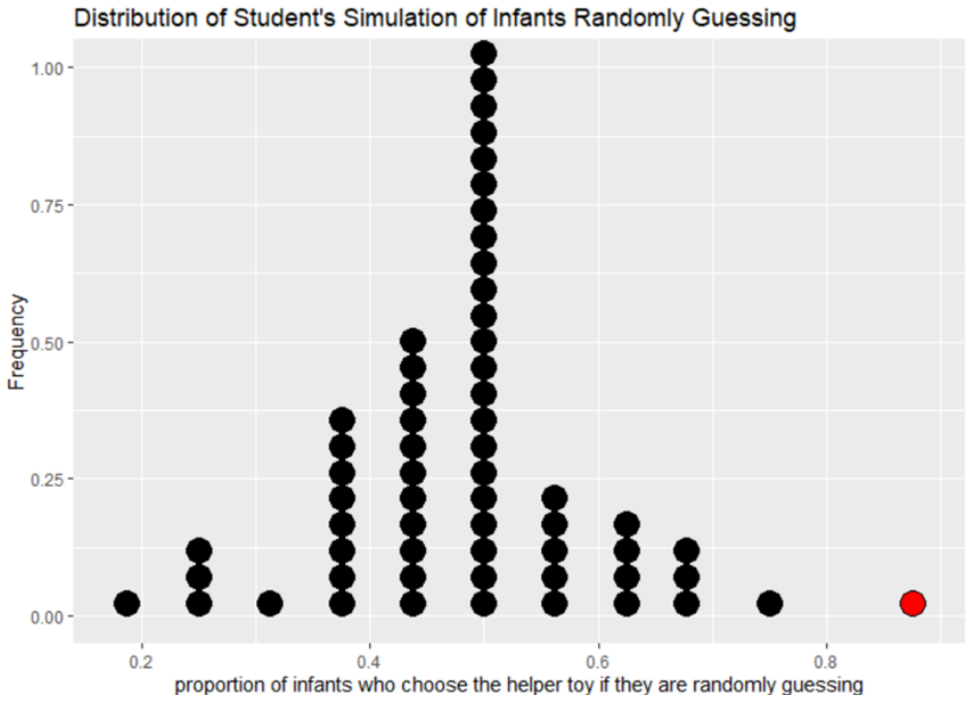
Let's think about how to use a coin to create one simulation of 16 infants under the assumption the null hypothesis is true. Let heads equal infant chose the helper toy and tails equal infant chose the hinderer toy.

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

15. 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 a head.

16. Is the value from question 15 closer to 0.5, the null value, or closer to the sample proportion, 0.875?

The distribution of the proportion of 16 coin flips from a Spring 2023 class is provided below.



17. Circle the observed statistic (value from question 2) 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?

18. Is the observed statistic (0.875) likely to happen or unlikely to happen if the true proportion of infants age 6 to 10 months who choose the helper is 0.5? Explain your answer using the plot.

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

6.2.5 Take-home messages

1. 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.
2. 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.
3. 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.

4. 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.

5. 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**.

6. 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.

7. 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.

8. 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.

6.2.6 Additional notes

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

6.3 Activity 6: Helper-Hinderer (continued)

6.3.1 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

6.3.2 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.

6.3.3 Helper-Hinderer

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

1. Report the sample proportion calculated in the out of class activity.

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

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

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

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

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

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

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

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

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

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

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

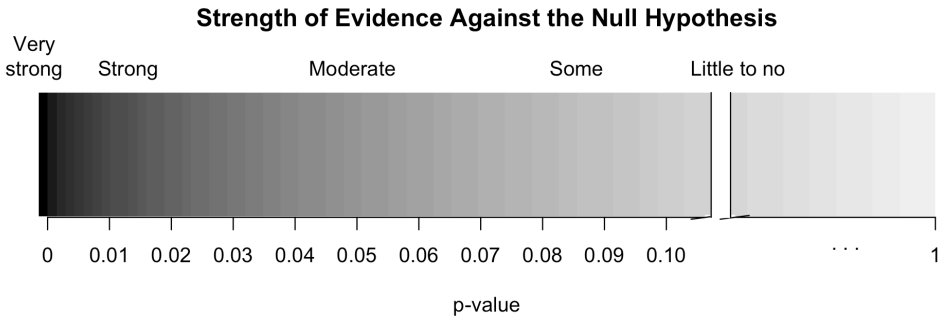
6. Circle the observed statistic (value from question 1) on the distribution you drew in question 4. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

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

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

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

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



Interpret the p-value

The p-value measures the probability that we observe a sample proportion as extreme as what was seen in the data or more extreme (matching the direction of the Ha) IF the null hypothesis is true.

10. What did we assume to create the null distribution?

11. What value did we compare to the null distribution to find the p-value?

12. What direction did we count simulations from the statistic?

13. Fill in the blanks below to interpret the 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) _____.

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?

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

Researchers were interested if infants observe social cues and would be more likely to choose the helper toy over the hinderer toy. In a sample of (sample size) _____infants, (number of successes) _____chose the helper toy. A simulation null distribution with 1000 simulations was created in RStudio. The p-value was found by calculating the proportion of simulations in the null distribution at the sample statistic of 0.875 and greater. This resulted in a p-value of (value of p-value)_____. We would observe a sample proportion of (value of the sample proportion) _____ or (greater, less, more extreme) _____ with a probability of (value of p-value)_____

IF we assume (H_0 in context) _____.
Based on this p-value, there is (very strong/little to no) _____ evidence that the (sample/true)_____ proportion of infants age 6 to 10 months who will choose the helper toy is (greater than, less than, not equal to) _____ 0.5.
In addition, a 95% confidence interval was found for the parameter of interest. We are 95% confident that the (true/sample)_____ proportion of infants age 6 to 10 months who will choose the helper toy is between (lower value)_____ and (upper value)_____. The results of this study can be generalized to (all infants age 6 to 10 months/infants similar to those in this study)_____ as the researchers (did/did not)_____ select a random sample.

6.3.4 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.

6.3.5 Additional notes

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

MODULE 7

Inference for a Single Categorical Variable: Theory-based Methods + Errors and Power

7.1 Lecture Notes Module 7: Inference for One Categorical Variable using Theory-based Methods

Theory-based methods

Central limit theorem

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 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: $N(0, 1)$

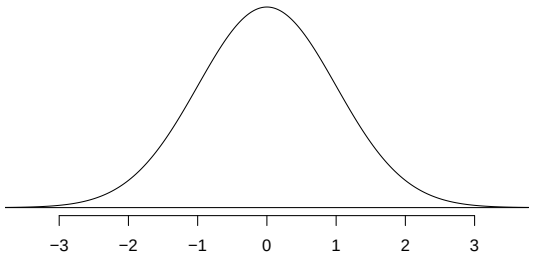


Figure 7.1: A standard normal curve.

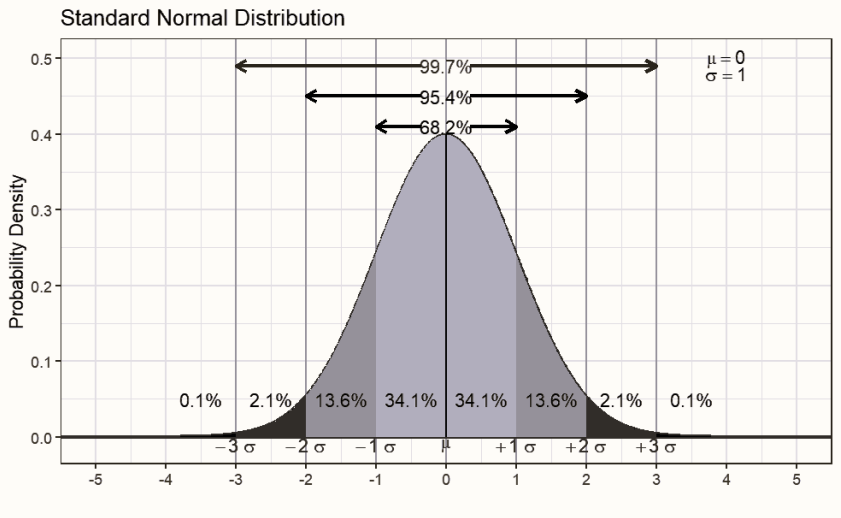
Standardized statistic: Z - score

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

- Measures the _____ of standard _____ the statistic is from the null value

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.

Theory-based methods for a single categorical variable

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:

- 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 sample proportion assuming the null hypothesis is true:

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

Equation for the standardized sample proportion:

- This value measures how many _____ deviations the sample _____ is above/below the _____ value.

Example for Class Discussion: 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?

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

Hypotheses:

In notation:

H_0 :

H_A :

In words:

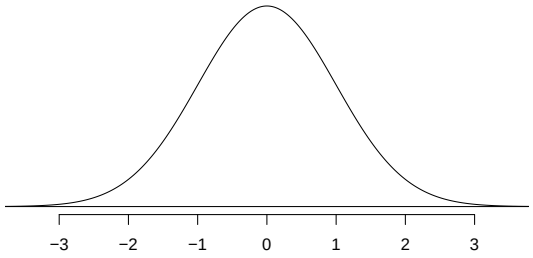
H_0 :

H_A :

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:

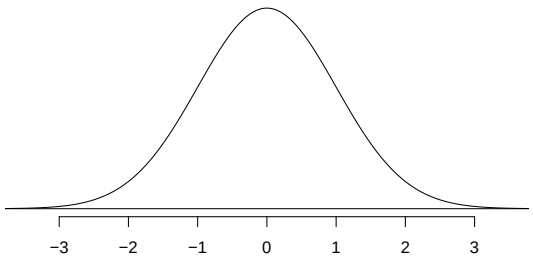
- Can the results of the study be generalized to the target population?

Confidence interval

- Interval of _____ values for the parameter of interest
- $CI =$ statistic $\pm$ margin of error

Theory-based method for a single categorical variable

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



For a 95% confidence interval:

```
qnorm(0.975, lower.tail=TRUE)
#> [1] 1.959964
```

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

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

Example: Estimate the true proportion of Montana residents that have donated blood at least once in their life. Find a 95% confidence interval:

Confidence interval interpretation:

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

Interpreting confidence level

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

- Our goal is to only take _____ sample from the _____ to create a _____ 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.

Errors, power, and practical importance

Type 1 Error: _____ the null hypothesis, when the null is _____.

- Only can have a Type 1 Error when we make the _____ to _____ the null hypothesis.
- The probability of a Type 1 Error is α, the _____ level

Type 2 Error: _____ to reject the null hypothesis, when the null is _____.

- Only can have a Type 2 Error when we make the _____ to _____ to reject the null hypothesis.

Power: probability of _____ the null hypothesis, when the null is _____.

Increasing power:

- Increase _____ _____
- Increase _____ _____
- Use a _____ alternative vs. a _____ alternative
- Increase the _____ size, the _____ between the believed true value and the null value

Confirmation bias: looking for _____ that supports our ideas

- Always should write H_A based on the _____ _____ prior to _____ collection!

Recall from the blood donation study, that we concluded there was very strong evidence that the true proportion of Montana residents who are eligible to donate blood differs from 0.10.

Since, we made the decision to _____ the null hypothesis, we have the possibility of a _____ error.

- What is the probability of this error?
- Write the error in context of the problem.

For each of the following changes to the blood donation study, determine whether the power of the test would increase or decrease.

- If we decreased the sample size from 200 to 100, power would _____.
- If we decreased the significance level from 0.05 to 0.01, power would _____.
- If we changed the research question to only asking if the probability a Montana resident eligible to donate blood actually does so is greater than 0.10, power would _____.
  - This is an example of _____ _____.

Practical importance

- A result can be _____ significant but not _____ important.
- Statistically significant: p-value $< \alpha$
  - Depends on the _____, the _____ _____, and the selected _____ level.
- Practically important: the _____ seen in the data is meaningful and has _____ applications.
  - Depends on the _____ and subjective opinion.

Example: An Austrian study of heights of 507,125 military recruits reported that men born in spring were statistically significantly taller than men born in the fall (p-value < 0.0001). A confidence interval for the true difference in mean height between men born in spring and men born in fall was (0.598, 0.602) cm.

Is there statistical significance?

Is there practical importance?

7.2 Out-of-Class Activity Module 7: Handedness of Male Boxers

7.2.1 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.

7.2.2 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 week 6 covered simulation-based methods for hypothesis tests involving a single categorical variable. This activity covers theory-based methods for testing a single categorical variable.

7.2.3 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.

```
 # 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
```

Review of summary statistics

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

2. Write out the null hypothesis in words.

3. Write out the alternative hypothesis in notation.

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

5. Verify that the independence condition is satisfied.

6. 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:

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

.

The standard error of the sample proportion measures the variability of possible sample proportions from the actual proportion. In other words, how far each possible sample proportion is from the actual proportion on average. 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.

7. 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.

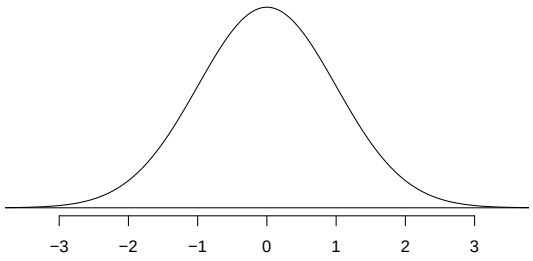


Figure 7.2: Standard Normal Curve

8. 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.*

9. Interpret the standardized sample proportion from question 8 in context of the problem.

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

Notes:

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

```r
pnorm(4.769, # 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] 9.257133e-07
```

10. Report the p-value obtained from the R output.

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

7.2.4 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.

7.2.5 Additional notes

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

7.3 Activity 7: Handedness of Male Boxers — Theory CI

7.3.1 Learning objectives

- Calculate a theory-based confidence interval for a single proportion.
- Check the appropriate conditions to find a theory-based confidence interval.
- Interpret a confidence interval for a single proportion.
- Use the normal distribution to find the multiplier needed for a confidence interval

7.3.2 Terminology review

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

- Parameter of interest
- Multiplier
- Normal distribution

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

7.3.3 Handedness of Male Boxers

In the out-of-class activity we found very strong evidence that the true proportion of male boxers that are left-handed is greater than 0.1. In this activity we will use the same data set to find the theory-based 95% confidence interval.

Remember from the last activity: 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. 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.

Recall that to use theory-based methods we must check the conditions to approximate the sampling distribution with the normal distribution. From the previous activity, we saw that independence was satisfied as the researchers took a random sample and that the sample had more than 10 successes and 10 failures.

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.

1. 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 4 and 5 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 4. This will give the value of the multiplier for a 95% confidence interval.

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

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

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

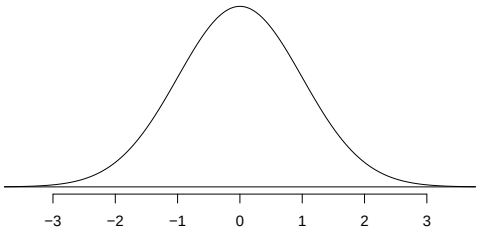


Figure 7.3: Standard Normal Curve

4. 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:

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

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

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

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

7. Is the null value, 0.1, contained in the 95% confidence interval? Explain, based on the p-value from the last activity, why you expected this to be true.

Simulation Methods

In the out-of-class activity, we found that the success-failure condition was met to use theory-based methods. Here we will use simulation methods to find a 95% confidence interval for the parameter of interest.

Use the `one_proportion_bootstrap_CI()` function in R 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` in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations.
- Make sure to run the library(catstats) function before running the one_proportion_bootstrap_CI function.
- Highlight and run lines 9–13

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

8. Report the simulation 95% confidence interval. Is this confidence interval similar to the confidence interval calculated in question 5? Explain why this makes sense.

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"?

9. In the last part of the activity we found a 95% confidence interval for the parameter of interest. 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.*

10. 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 9 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 9?

- 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 9?

- How many intervals out of 100 contain π, the true value chosen in question 9? *Hint*: This is given to the left of the graph of green and red intervals.

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

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

7.3.4 Take-home messages

1. In theory-based methods, we add and subtract a margin of error to the sample statistic. The margin of error is calculated using a multiplier that corresponds to the level of confidence times the variability (standard error) of the statistic.
2. The confidence interval calculated using theory-based methods should be similar to the confidence interval found using simulation methods provided the success-failure condition is met.
3. 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.

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 Module 7 Lab: 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 ACL recovery

It is widely reported that the median recovery time for athletes who undergo surgery to repair a torn anterior cruciate ligament (ACL) is 8 months, indicating that 50% of athletes return to their sport within 8 months after an ACL surgery. Suppose a local physical therapy company hopes to advertise that their rehabilitation program can increase this percentage.

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

2. **Use proper notation to write the null and alternative hypothesis the company would need to test in order to check their advertisement 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 lab will be spent investigating how different factors influence the power of a test, after which you will complete a power analysis for this physical therapy company.

- Navigate to https://istats.shinyapps.io/power/. *Please note that this applet uses p_0 to represent the null value rather than π_0.*
- Use the scale under "Null Hypothesis value p_0" to change the value to your null value from question 2.
- Change the "Alternative Hypothesis" to the direction you wrote in question 2.
- Leave all boxes un-checked. Do not change the scales under "True value of p_0", "Sample size n", or "Type I Error α"

The red distribution you see is the scaled-Normal distribution representing the null distribution for this hypothesis test, if the sample size was 50 and the significance level was 0.05. This means the red distribution is showing the probability of each possible sample proportion of athletes who returned to their sport within 8 months ($\widehat{p}$) if we assume the null hypothesis is true.

3. Based off this distribution and your alternative hypothesis, give one possible sample proportion 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 *minimum* sample proportion which would lead to reject the null hypothesis. What is this value?

5. Notice that there are some sample proportions under the red line (when the null hypothesis is true) which would lead us to reject the null hypothesis. Give the range of sample proportions 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 maximum 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 (what is the probability of making a Type 1 Error)?**

Let's say this physical therapist company believes their program can get 70% of athletes back to their sport within 8 months of an ACL surgery. In the applet, set the scale under "True value of p" to 0.7.

7. Where is the blue distribution centered?

The blue distribution that appears represents what the company believes, that 0.7 (not 0.5) is the true proportion of its clients who return to their sport within 8 months of ACL surgery. This blue distribution represents the idea that the **null hypothesis is false**.

8. Consider the definition of power provided earlier in this lab. 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 what power and Type 2 Error are conditional on.*

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 p" to see the effect on Power.

True value of p	0.60	0.65	0.70	0.75	0.80
Power					

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

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

13. Using the same significance level, set the "True value of p" to 0.7 and change the sample size to see the 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 p", set the sample size to 50 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 physical therapy company. The company believes 70% of their patients will return to their sport within 8 months of ACL surgery. 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 athletes the company will need to collect data from in order to meet these goals? Use the applet to answer this question, then download your image created and upload the file to Gradescope.**

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

- Not being able to advertise their ACL recovery program is better than average when their program really is better.
- Advertising their ACL recovery program is better even though it is not.

MODULE 8

Inference for Two Categorical Variables: Simulation-based Methods

8.1 Lecture Notes Module 8: Inference for Two Categorical Variables using Simulation-based Methods

Two categorical variables

- In this week, we will study inference for a _____ explanatory variable and a _____ response.
- The summary measure for two categorical variables is the _____ in _____.

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

$\pi_1 - \pi_2$:

Notation for the Sample Statistics

- Sample proportion for group 1:
- Sample proportion for group 2:
- Sample difference in proportions:
- Sample size for group 1:
- Sample size for group 2:

Example for class discussion: 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:

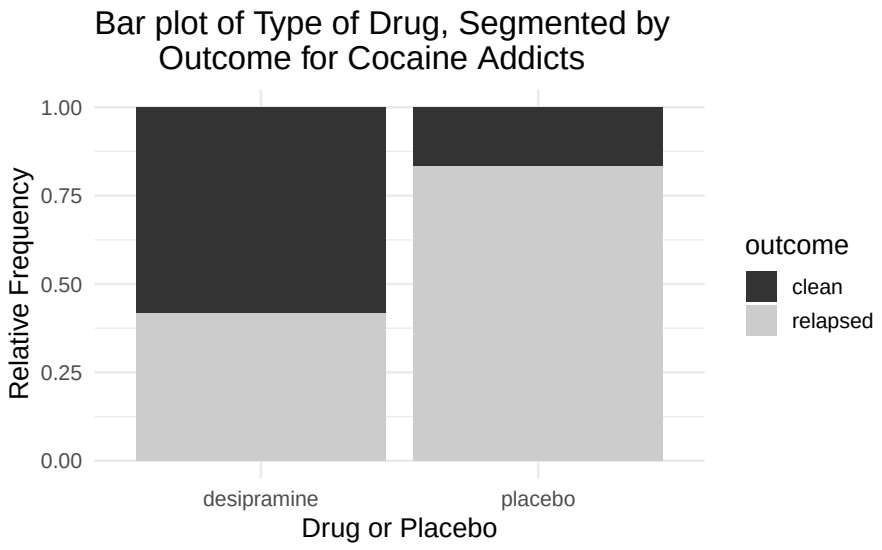
Parameter of interest:

```
(cocaine.table<-table(cocaine$outcome, cocaine$drug))
#>
#>            desipramine placebo
#>   clean             14       4
#>   relapsed          10      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()
```



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 alternative hypothesis.

Write the null and alternative hypotheses for the cocaine study:

In words:

H_0 :

H_A :

In notation:

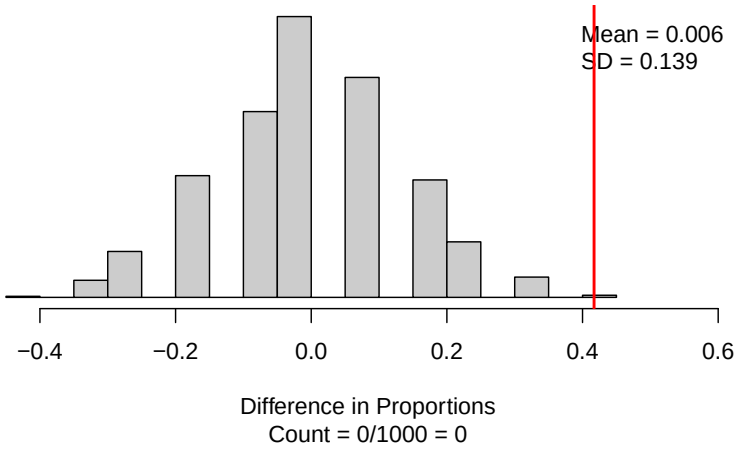
H_0 :

H_A :

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 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\hat{p}_1 - \hat{p}_2$

```r
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 = 1000, # Always use a minimum of 1000 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")
```



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

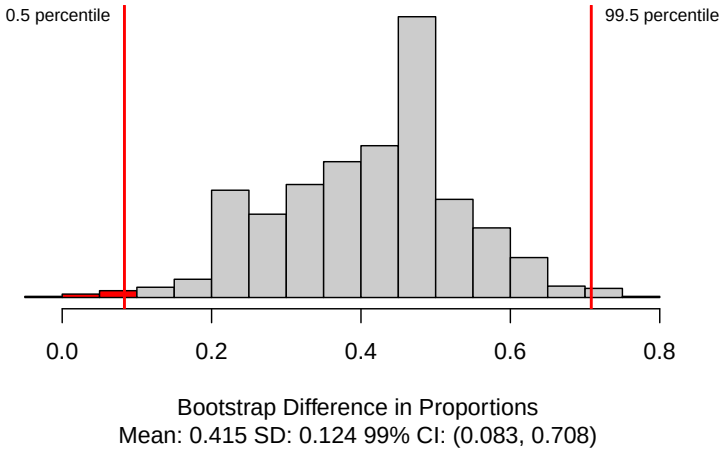
To estimate the difference in true proportion we will create a confidence interval.

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 1000 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 = 1000, # Always use a minimum of 1000 repetitions
        confidence_level = 0.99) # Enter the level of confidence as a decimal
```



Confidence interval interpretation:

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

Does the confidence interval agree with the p-value?

8.2 Out-of-Class Module Week 8: The Good Samaritan — Intro

8.2.1 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

8.2.2 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.

8.2.3 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:

```
# 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
```

Vocabulary review

1. What is the name of the explanatory variable as it is written in the R output? What are its categories?

2. What is the response variable in the R output? What are its categories?

3. Fill in the blanks with one answer from each set of parentheses: This is an
   _____ (experiment/observational study) because
   _____ (hurry or no hurry/help or no help) _____ (was/was not)
   randomly _____ (assigned/selected).

4. Put an X in the box that represents the appropriate scope of inference for this study.

	Study Type	
Selection of Cases	Randomized Experiment	Observational Study
Random Sample (no sampling bias)		
Non Random Sample (sampling bias)		

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.

5. Write the two parameters of interest in context of the study.

   π_{hurry} —

   $\pi_{\text{no hurry}}$ —

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

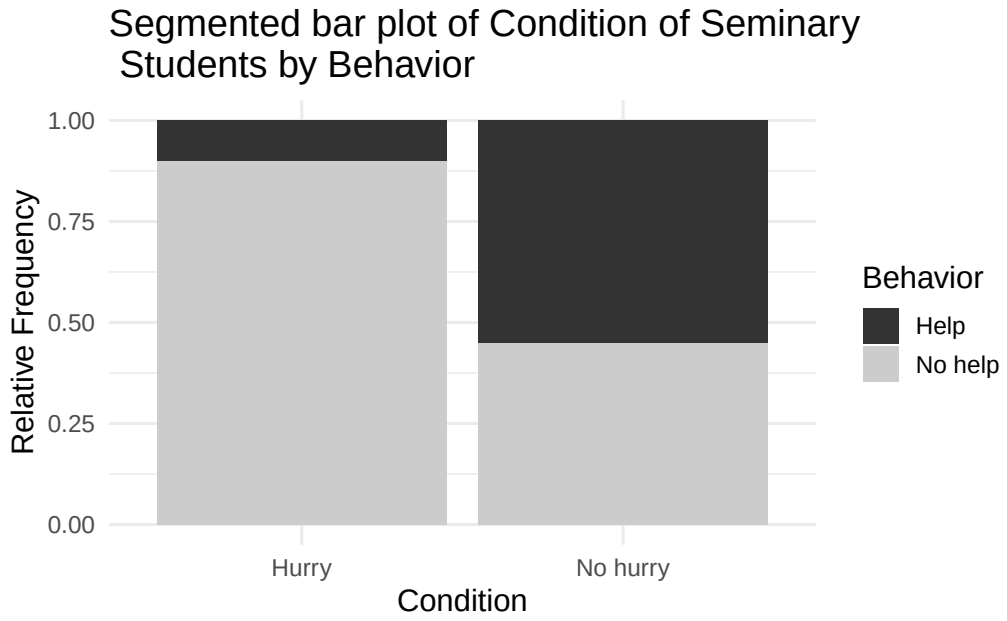
6. Write the null hypothesis out in words using your answers to question 5.

7. 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

```
good %>%
  ggplot(aes(x = Condition, fill = Behavior))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Condition of Seminary \n Students by Behavior", #Title your plot
       y = "Relative Frequency", #y-axis label
       x = "Condition") + #x-axis label
  scale_fill_grey()
```



8. Using the provided segmented bar plot, is there an association between whether a Seminary student helps the actor and condition assigned?

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

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

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

12. What is the notation used for the value calculated in question 11?

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.

13. 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?

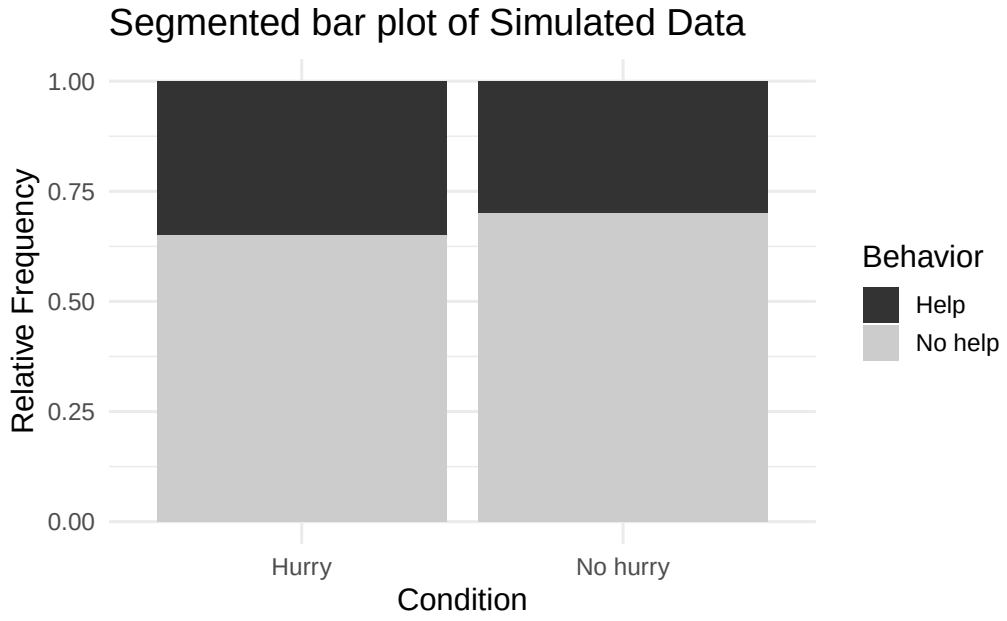
14. Next, we would mix the cards together and shuffle into two piles.

    How many cards would be in each pile?

    What would each pile represent?

15. 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?

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



16. Compare the segmented bar plot for the simulated data to the previous segmented bar plot of the original data. Explain how the segmented bar plot of the simulated data reflects the null hypothesis.

8.2.4 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.

8.2.5 Additional notes

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

8.3 Activity 8: The Good Samaritan (continued) — Simulation-based Hypothesis Test & Confidence Interval

8.3.1 Learning outcomes

- Identify the parameter of interest for a difference in proportions.
- 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.
- Create and interpret a simulation-based confidence interval for a difference in proportions.

8.3.2 Terminology review

In today's activity, we will use simulation methods to estimate the difference in two proportions. Some terms covered in this activity are:

- Hypothesis test
- P-value
- Parameter of interest
- Bootstrapping
- Confidence interval
- Types of errors

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

8.3.3 The Good Samaritan

In the out of class activity, we began a test of significance to see if people in a hurry are less likely to help those in need. Today we will use RStudio to continue to assess this research question.

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

1. Simulate one sample assuming the null hypothesis is true using the cards provided by your instructor. Write down the value of the simulated statistic. How does the value of your group's simulated statistic compare to the other groups at your table? Are the simulated values closer to the null value of zero than the actual calculated difference in proportions?

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`.

2. 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 16 in place of the `xx`'s to produce the null distribution with 1000 simulations; highlight and run lines 1–16.

```
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 = 1000, # Always use a minimum of 1000 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")
```

3. Sketch the null distribution created here.

4. What value is the null distribution centered around? Explain why this makes sense.

5. What is the value of the p-value? *Remember*: This is the value given at the bottom of the null distribution.

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

7. How much evidence does the p-value provide against the null hypothesis? *Hint*: Refer to the guidelines given in Week 6.

8. Do you expect the null value to be in a 99% confidence interval? Explain your answer.

Use statistical analysis methods to draw inferences from the data

In this part of the activity, we will estimate the difference in true proportion of people who will help others for those in the hurry condition and those not in the hurry condition by finding a confidence interval.

9. Write the parameter of interest in context of the study. Use proper notation.

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 `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 confidence level as a decimal.

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

10. 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? `"Hurry"` or `"No hurry"`):
- Response value numerator (What is the outcome for the response variable that is considered a success? `"Help"` or `"No help"`):
- Number of repetitions:
- Confidence level (entered as a decimal):

Using the R script file for this activity, enter your answers for question 7 in place of the `xx`'s to produce the bootstrap distribution with 1000 simulations; highlight and run lines 16–21.

```
two_proportion_bootstrap_CI(formula = Behavior ~ Condition,
        data=good, # 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 = 1000, # Always use a minimum of 1000 repetitions
        confidence_level = xx) # Enter the level of confidence as a decimal
```

11. Where is the bootstrap distribution centered? Explain why.

12. Report the bootstrap 99% confidence interval.

13. What percentile of the bootstrap distribution does the upper value of the confidence interval represent?

14. Interpret the 99% confidence interval in context of the problem.

15. Write a conclusion to the test.

8.3.4 Take-home messages

1. We use the same guidelines for the strength of evidence as we did in Activity 6A.
2. To create one simulated sample on the bootstrap distribution for a difference in sample proportions, label $n_1 + n_2$ cards with the outcomes for the original responses. 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 the proportion of successes.
3. If the null value is not contained in a 99% confidence interval, then there is evidence against the null hypothesis and the p-value is less than the significance level of 0.01.

8.3.5 Additional notes

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

8.4 Module 8 Lab: Poisonous Mushrooms

8.4.1 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.

8.4.2 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 Week 8 lab. Upload and import the csv file, `mushrooms_edibility`.
- Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 8.
- Highlight and run lines 1–9 to get the counts for each combination of categories.

```
mushrooms <- datasetname # 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 = 1000, # Always use a minimum of 1000 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 = 1000, # Always use a minimum of 1000 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 9

Inference for Two Categorical Variables: Theory-based Methods

9.1 Lecture Notes Module 9: Theoretical Inference for Two Categorical Variables

Hypothesis testing using theory-based methods

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:
- 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 _____ is from the _____ value, on average.

Equation for the standardized difference in sample proportions:

- This value measures how many _____ deviations the sample difference in _____ is above/below the _____ value.

Example for class discussion: In the week 3 Lab, 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?

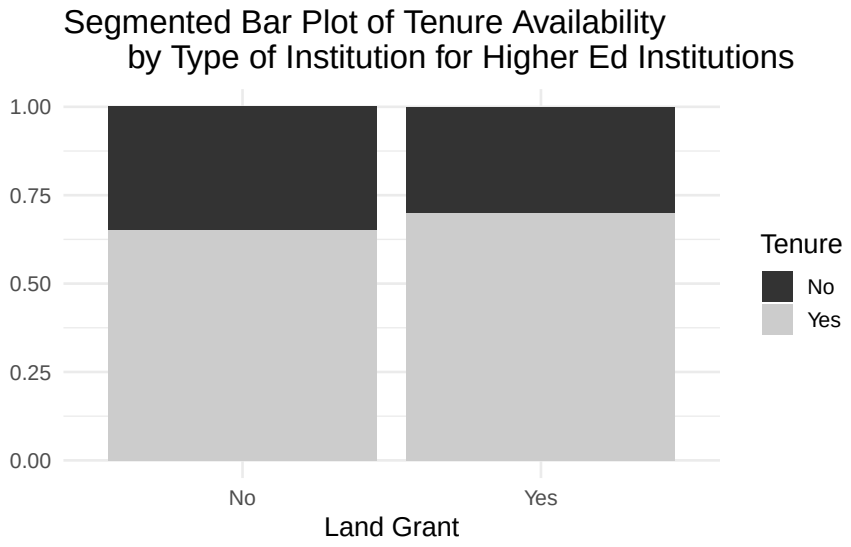
Are the conditions met to analyze the university data using theory-based methods?

```
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
```



What is the explanatory variable?

What is the response variable?

Write the parameter of interest:

Hypotheses:

In notation:

H_0 :

H_A :

In words:

H_0 :

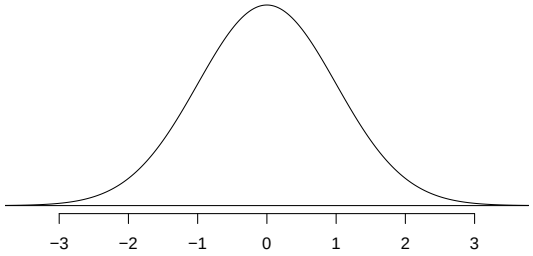
H_A :

Report the summary statistic:

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

- Estimate the _____ in true _____
- CI = statistic $\pm$ 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

```
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 Out-of-Class Activity Module 9: Winter Sports Helmet Use and Head Injuries — Theory-based Confidence Interval

9.2.1 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.

9.2.2 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.

9.2.3 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.

1. Explain why a theory-based confidence interval for the Good Samaritan study from last week would NOT be similar to the bootstrap interval created.

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

2. Write the parameter of interest, in words, for this study, in context of the problem.

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.

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.

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \text{ where}$$
$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

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.

To calculate the standard error for a difference in proportions to create a 90% confidence interval we substitute in the two sample proportions and the sample size for each group into the equation above.

$$n_1 = 752, n_2 = 2810, \hat{p}_h = \frac{96}{752} = 0.128, \hat{p}_n = \frac{480}{2810} = 0.171$$

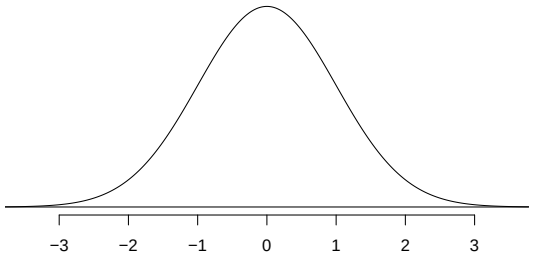
$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.128 \times (1 - 0.128)}{752} + \frac{0.171 \times (1 - 0.171)}{2810}} = 0.014$$

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
```

4. 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.

$$ME = z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

5. Using the multiplier of $z^* = 1.645$ and the calculated standard error, calculate the margin of error for a 90% confidence interval.

6. Calculate the 90% confidence interval for the parameter of interest.

7. Interpret the confidence interval found in question 6 in context of the problem.

8. Interpret the level of confidence in context of the problem. What does it mean to be 90% confident in the confidence interval?

9. What decision (reject or fail to reject the null hypothesis) would you make based on your confidence interval? Explain your answer.

9.2.4 Effect of sample size

Suppose in another sample of skiers and snowboards involved in accidents we saw these results:

	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.

$$\hat{p}_h = \frac{135}{1056} = 0.127, \;\; \hat{p}_n = \frac{674}{3944} = 0.171$$

10. Calculate the standard error for the difference in sample proportions for this new sample.

11. Calculate the margin of error for a 90% confidence interval using a multiplier of $z^* = 1.645$ for this new sample. Is the margin of error larger or smaller than the margin of error for the original study?

12. Calculate the 90% confidence interval for this new study using the margin of error from question 10.

13. Is the confidence interval calculated in question 12 with the larger sample size wider or narrower than the confidence interval in question 6? Why?

9.2.5 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*. 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.
3. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and thus a narrower confidence interval.

9.2.6 Additional notes

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

9.3 Activity Module 9: Winter Sports Helmet Use and Head Injuries — Theory-based Hypothesis Test

9.3.1 Learning outcomes

- Given a research question involving two categorical 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 difference in proportions.
- Calculate the Z test statistic for a difference in proportions.
- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a difference in proportions.

9.3.2 Terminology review

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

- Conditional proportion
- Z test
- z^* multiplier
- Null hypothesis
- Alternative hypothesis
- Test statistic
- Standard normal distribution
- Independence and success-failure conditions
- Relative risk

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

9.3.3 Helmet use and head injuries

For this activity we will again use the Helmet Use ann Head Injury data set. In the out-of-class activity we found that the null value of zero was not contained in the 90% confidence interval. In today's activity, we will calculate the standardized difference in sample proportion to find the p-value of the test.

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. Is there evidence that safety helmet use is associated with a reduced risk of head injury for skiers and snowboarders?

For this study the observational units are skiers and snowboarders involved in accidents. A success will be considered a head injury in this context and we are comparing the groups helmet use (group 1) and no helmet use (group 2). Use helmet use - no helmet use as the order of subtraction.

- Highlight and runs lines 1–6 in the provided Rscript file to create the summary data table.

```
injury <- read.csv("https://math.montana.edu/courses/s216/data/HeadInjuries.csv")
injury %>% group_by(Helmet) %>% count(Outcome)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Helmet [2]
#>   Helmet Outcome            n
#>   <chr>  <chr>          <int>
#> 1 No     Head Injury      480
#> 2 No     No Head Injury  2330
#> 3 Yes    Head Injury       96
#> 4 Yes    No Head Injury   656
```

1. Fill in the following two-way table using the R output.

	Helmet Use		
Head Injury	Yes	No	Total
Head Injury			
No Head Injury			
Total			

2. Write the null and alternative hypotheses in notation.

H_0:

H_A:

3. Calculate the summary statistic (difference in proportions) for this study. Use appropriate notation with clear subscripts.

4. Interpret the difference in sample proportions in context of the study.

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 week 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.

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.

5. Is the independence condition met? Explain your answer.

6. Is the success-failure condition met for each group? Explain in context of the study.

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.

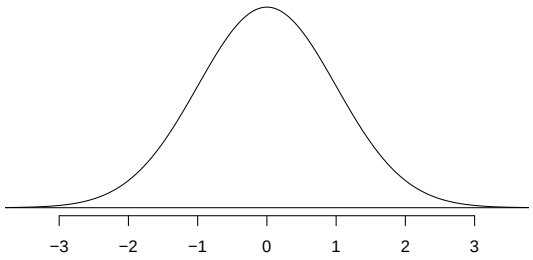
$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{576}{3562} = 0.162$$

We use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$.

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{752} + \frac{1}{2810}\right)} = 0.015$$

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 statistic on the standard normal distribution above and shade the area to find the p-value.



We will use the `pnorm()` function in R to find the p-value.

- Use the provided R script file and enter the value of the standardized statistic found in question 7 at `xx` in line 11
- Highlight and run lines 11–13.

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

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

11. Write a conclusion to the research question based on the p-value found.

12. What is the scope of inference for this study?

Impacts on the p-value

Suppose that we want to show that there is a **difference** in true proportion of head injuries for those that wear helmets and those that do not.

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

14. How would this impact the p-value?

Suppose in a larger sample of skiers and snowboarders involved in accidents we saw the following results.

	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.

$$\hat{p}_h = \frac{135}{1056} = 0.127, \hat{p}_n = \frac{674}{3944} = 0.171$$

15. The standard error for the difference in proportions for this new sample is 0.013 ($SE(\hat{p}_h - \hat{p}_n) = 0.013$). Calculate the standardized statistic for this new sample.

Use Rstudio to find the p-value for this new sample.

- Enter the value of the standardized statistic found in question 15 for xx in line 18.
- 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 greater than the standardized statistic
```

16. How does the increase in sample size affect the p-value?

17. Suppose another sample of 3562 skiers and snowboarders was taken. In this new sample a difference in proportions of head injuries was found to be -0.009, ($\hat{p}_h - \hat{p}_n = -0.009$) with a standard error for the difference in proportions of 0.015, ($SE(\hat{p}_h - \hat{p}_n) = 0.015$). Calculate the standardized statistic for this new sample.

Use Rstudio to find the p-value for this new sample.

- Enter the value of the standardized statistic found in question 17 for xx in line 25.
- Highlight and run lines 25–27.

```
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 greater than the standardized statistic
```

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

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

a) Switching to a two-sided test.

b) Using a larger sample size.

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

9.3.4 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. The standardized statistic when the response variable is categorical is a Z-score and is compared to the standard normal distribution to find the p-value. To find the standardized statistic, we take the value of the statistic minus the null value, divided by the null standard error of the statistic. The standardized statistic measures the number of standard errors the statistic is from the null value.
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.

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 Module 9 Lab: Diabetes

9.4.1 Learning outcomes

- Given a research question involving two categorical 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 difference in proportions.
- Describe and perform a simulation-based hypothesis test for a difference in proportions.
- Calculate the Z test statistic for a difference in proportions.
- Find, interpret, and evaluate the p-value for a hypothesis test 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 for Week 9 lab. Upload and import the csv file, `diabetes`.
- Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 7.
- Highlight and run lines 1–8 to get the counts for each combination of categories.

```
glycemic <- datasetname
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. Write the parameter of interest for the research question.

5. Using the research question, write the alternative hypothesis in notation.

6. **Calculate the summary statistic (difference in proportions). Use appropriate notation.**

- Fill in the missing values/names in the R script file in the two-proportion_test function to create the null distribution and find the simulation p-value for the test.

```
two_proportion_test(formula = outcome~treatment, # response ~ explanatory
         data= glycemic, # Name of data set
         first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
         number_repetitions = 1000, # Always use a minimum of 1000 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")
```

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

8. **Will the theory-based p-value be similar to the simulation p-value? Explain your answer.**

9. **Calculate the number of standard errors the sample difference in proportion is from the null value of zero.**

10. **Will a 95% simulation confidence interval contain the null value of zero? Explain your answer.**

11. Calculate the standard error for a difference in proportions to create a 95% confidence interval.

12. Use the multiplier of $z^* = 1.96$ and the standard error found in question 11 to calculate a 95% confidence interval for the parameter of interest.

13. 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 p-value interpretation and scope of inference to Gradescope.

Paragraph (continued):

MODULE 10

Probability and Relative Risk

10.1 Lecture Notes Module 10: Probability and Relative Risk

Probability

- 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

Finding probabilities from a table

	A	A^c	Total
B	A and B	A^c and B	Total B
B^c	A and B^c	A^c and B^c	Total B^c
Total	Total A	Total A^c	TOTAL

Calculating unconditional probabilities:

$P(A) =$

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

Calculating conditional probabilities:

$P(A|B) =$

$P(B|A) =$

Example for class discussion: 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?

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 for class discussion: 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:

Relative Risk

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

Relative Risk:

- 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.

Example for class discussion: 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.

- Interpret the value of relative risk 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 :

10.2 Out-of-Class Activity Module 10: Titanic Survivors – Relative Risk

10.2.1 Learning outcomes

- Interpret the value of relative risk in terms of a percent increase or decrease.
- Evaluate the association between two categorical variables using relative risk.

10.2.2 Terminology review

In today's activity, we will look another summary. Some terms covered in this activity are:

- Conditional proportion
- Relative risk

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

10.2.3 Titanic Survivors

A complete data set exists listing all those aboard HMS Titanic and includes related facts about each person including age, how much they paid for their ticket, which boat they survived in (if they survived), and their job if they were crew members. Stories, biographies and pictures can be found on the site: www.encyclopedia-titanica.org/. Did all passengers aboard the Titanic have the same chance of survival? Was the risk of death higher among 3rd class passengers compared to 1st class passengers?

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

```
# Read data set in
survive <- read.csv("https://math.montana.edu/courses/s216/data/Titanic.csv")
survive <- survive %>%
  filter(Class_Dept == "1st Class Passenger" | Class_Dept == "3rd Class Passenger")
survive %>% group_by(Class_Dept) %>% count(Survived)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Class_Dept [2]
#>   Class_Dept          Survived     n
#>   <chr>               <chr>    <int>
#> 1 1st Class Passenger Alive      166
#> 2 1st Class Passenger Dead       108
#> 3 3rd Class Passenger Alive      147
#> 4 3rd Class Passenger Dead       509
```

Data Exploration

1. Fill in the data from the R output to complete the two-way table.

	Class		
Outcome	1st Class Passenger	3rd Class Passenger	Total
Dead			
Alive			
Total			

2. Calculate the conditional proportion of 1st class passengers that died.

3. Calculate the conditional proportion of 3rd class passengers that died.

4. Calculate the difference in conditional proportions of death for 3rd and 1st class passengers. Use 3rd − 1st as the order of subtraction.

5. Interpret the difference in proportions in context of the problem.

Relative Risk

Another summary statistic that can be calculated for two categorical variables is the relative risk. The relative risk is calculated as the ratio of the conditional proportions:

$$\text{relative risk} = \frac{\widehat{p}_1}{\widehat{p}_2}.$$

6. Calculate the relative risk of death for 3rd class passengers compared to 1st class passengers.

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

8. Calculate the increase or decrease in risk of death for 3rd class passengers compared to 1st class passengers.

9. Interpret the increase or decrease in risk of death in context of the problem.

10. Calculate the percent increase or percent decrease in death.

11. Interpret the value of relative risk as a percent increase or percent decrease in death.

12. Based on the summary statistic, was the risk of death higher among 3rd class passengers compared to 1st class passengers? By what percent?

10.2.4 Risk in the News

13. Find a recent news article discussing 'risk'. Summarize the article below by answering the following questions.

- What is the article discussing the risk of? (This is the a *success* for the study.)

- What two groups are being compared? (These are the two levels of the *explanatory* variable.)

- What is the percent increase/decrease in risk reported? What is the relative risk comparing the two groups?

- Does the news report appear to indicate that the reported difference in the groups is statistically significant? Do you agree with the report? If so, explain why. If not, what further information would you need to assess statistical significance?

- Does the news report appear to indicate a causal relationship exists based on the reported relative risk? Do you agree with the report? Justify your answer.

10.2.5 Take-home messages

1. Relative risk calculates the ratio of the proportion of successes in group 1 compared to the proportion of successes in group 2.
2. Relative risk evaluates the percent increase or percent decrease in the response variable attributed to the explanatory variable. To find the percent increase or percent decrease we calculate the following percent change $= (RR - 1) \times 100\%$. If relative risk is less than 1 there is a percent decrease. If relative risk is greater than 1 there is a percent increase.

10.2.6 Additional notes

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

10.3 Activity 10: What's the probability?

10.3.1 Learning outcomes

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

10.3.2 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.

10.3.3 Probability

1. In a large general education class, 60% of students are science majors and 40% are liberal arts majors. Twenty percent of the science majors are seniors, while 30% of the liberal arts majors are seniors. Given the following two-way table answer the following questions.

   Let A = the event the student is a senior, and B = the event the student is a science major.

	Senior (A)	Not a Senior (A^c)	Total
Science Major (B)	12000	48000	60000
Not a Science Major (B^c)	12000	28000	40000
Total	24000	76000	100000

   a. What is the probability that a randomly selected senior is a science major? Use appropriate probability notation.

   b. What is the probability that a randomly selected student is both a senior and a science major. Use appropriate probability notation.

   c. What is the probability that a randomly selected student is not a senior given they are a liberal arts major. Use appropriate probability notation.

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

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

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

   $0.86 =$

   $0.92 =$

   $0.37 =$

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

	A	A^c	Total
B			
B^c			
Total			100,000

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

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

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

3. In a computer store, 30% of the computers in stock are laptops and 70% are desktops. Five percent of the laptops are on sale, while 10% of the desktops are on sale.

   Let L = the event the computer is a laptop, and S = the event the computer is on sale.

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

   $0.30 =$

   $0.70 =$

   $0.05 =$

   $0.10 =$

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

	L	L^c	Total
S			
S^c			
Total			100,000

c. Calculate the probability that a randomly selected computer will be a desktop, given that the computer is on sale. What is the notation used for this probability?

d. Find $P(S^C|L^C)$. What does this probability represent in context of the problem?

e. What is the probability a randomly selected computer is both a laptop and on sale? Give the appropriate probability notation.

10.3.4 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.

10.3.5 Additional notes

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

MODULE 11

Group Exam 2 Review

Use the provided data set from the Islands (Bulmer, n.d.) (Exam2ReviewData.csv) and the appropriate Exam 2 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.

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
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. Use the appropriate Exam 2 Review R script file and analyze the following research question: "Is there evidence that those with a higher education level are less likely to smoke?"

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?

MODULE 12

Inference for a Quantitative Response with Paired Samples

12.1 Lecture Notes Module 12: Inference for Paired Data

Single categorical, single quantitative variables

- In this week, 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 12.1: 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?

For a paired experiment, we look at the difference between responses for each unit (pair), rather than just the average difference between treatment groups

- The summary measure for paired data is the _____.
- Mean difference: the average _____ in the _____ variable outcomes for observational units between _____ variable groups

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

μ_d :

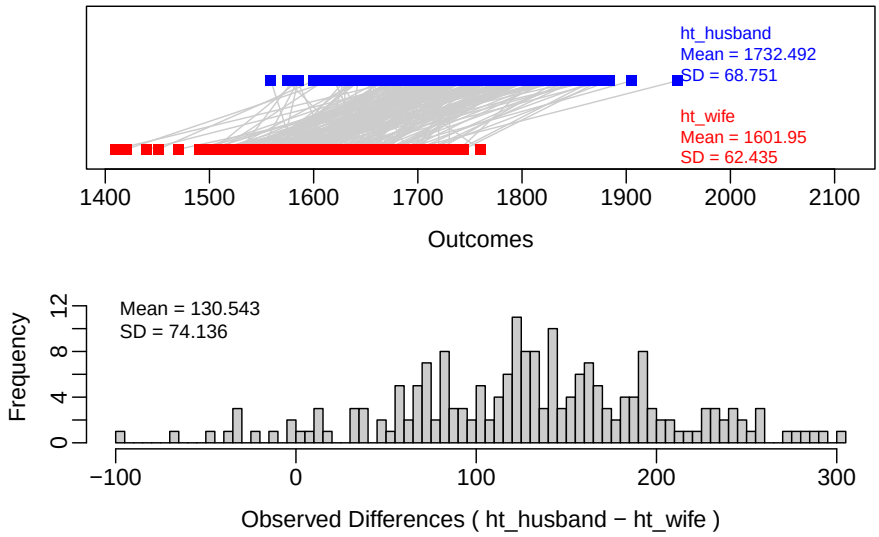
Notation for the Sample Statistics

- Sample mean of the differences:
- Sample standard deviation of the differences:

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.).

Parameter of interest:

```
hw <-read.csv("data/husbands_wives_ht.csv")
paired_observed_plot(hw)
```




```
hw_diff <- hw %>%
  select(ht_husband, ht_wife) %>%
  mutate(ht_diff = ht_husband-ht_wife)
```

```
hw_diff %>%
    summarise(favstats(ht_husband))
#>    min   Q1 median   Q3  max     mean       sd   n missing
#> 1 1559 1691   1725 1774 1949 1732.492 68.75067 199       0
```

```
hw_diff %>%
    summarise(fav_stats(ht_wife))
#>    min   Q1 median   Q3  max    mean     sd   n missing
#> 1 1410 1560   1600 1650 1760 1601.95 62.435 199       0
```

```
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
```

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 1000 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)

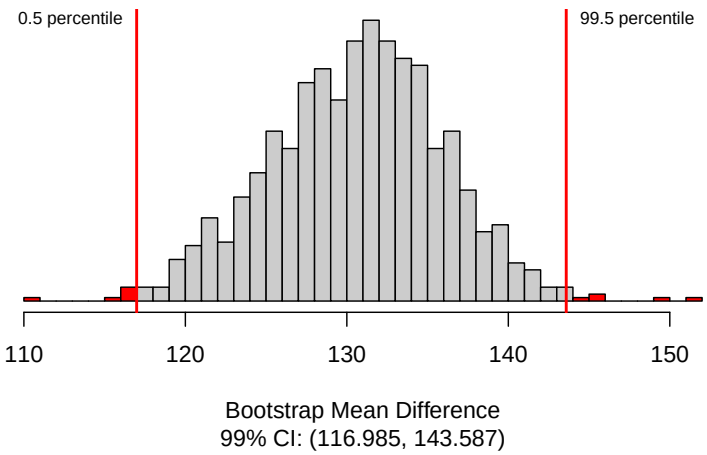
Conditions for inference for paired data:

- Independence:

Is the independence condition met for the height study?

Simulated bootstrap distribution:

```
set.seed(216)
paired_bootstrap_CI(data = hw_diff$ht_diff, # Enter vector of differences
            number_repetitions = 1000, # Number of bootstrap samples for CI
            confidence_level = 0.99,  # Confidence level in decimal form
            which_first = 1)  # Not needed when entering vector of differences
```



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

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.

Conditions for inference using theory-based methods for paired data (categorical explanatory and quantitative response):

- Independence: (same as for simulation); the difference in outcome for one observational unit will not influence another observation.
- Large enough sample size:
  - Normality: The data should be approximately normal or the sample size should be large.

    $n < 30$:

    $30 \leq n < 100$:

    $n \geq 100$:

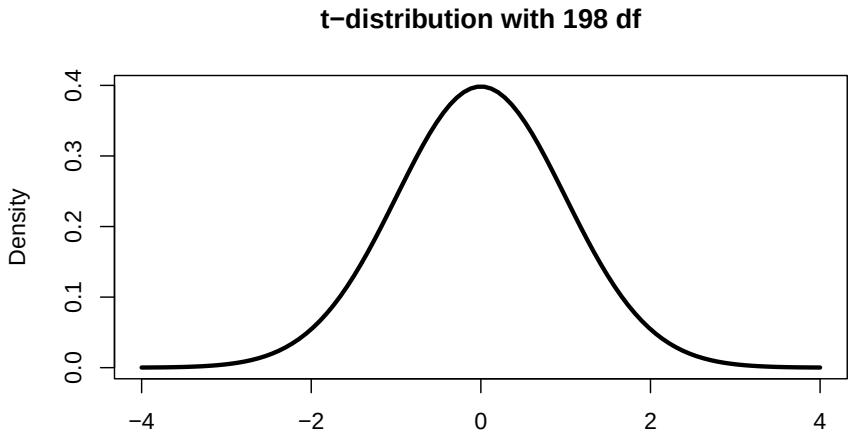
Theory-based Confidence Interval

statistic $\pm$ margin of error

Equation for the standard error for the sample mean difference:

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.



To find the t^* multiplier for a 99% confidence interval:

```
qt(0.995, df=198, lower.tail = TRUE)
#> [1] 2.600887
```

Calculate the margin of error:

Calculate the theory-based confidence interval.

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 alternative hypothesis.

Write the null and alternative for the height study:

In words:

H_0 :

H_A :

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 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}_d$

Reminder of summary statistics:

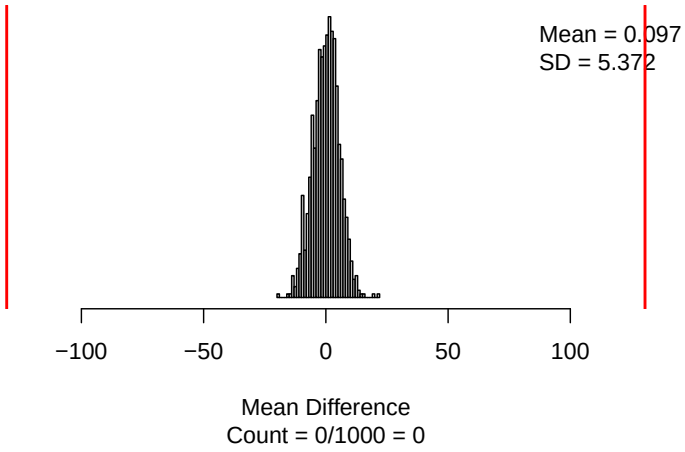
```
hw_diff %>%
    summarise(fav_stats(ht_diff))
#>   min   Q1 median  Q3 max     mean       sd   n missing
#> 1 -96 83.5    131 179 303 130.5427 74.13608 199       0
```

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 = 1000,  # 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

Theory-based method

- 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 standardized sample mean difference:

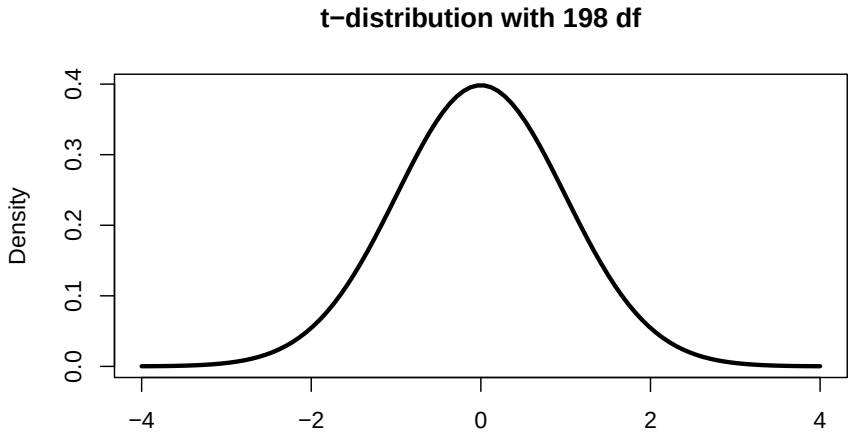
Reminder of summary statistics for height data:

```
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



Interpret the standardized statistic:

What theoretical distribution should we use to find the p-value using the value of the standardized statistic?

To find the p-value:

```
pt(24.84, df = 198, lower.tail=FALSE)*2
#> [1] 9.477617e-63
```

12.2 Out-of-Class Activity Module 12: Color Interference

12.2.1 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.

12.2.2 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.

12.2.3 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.

Ask a research question

2. Write out the null hypothesis in words, in the context of this study.

3. 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.

Conditions for the sampling distribution of $\bar{x}$ to follow an approximate Normal distribution:

- **Independence**: The sample's observations are independent. For paired data, that means each pairwise difference should be independent.
- **Normality**: The data should be approximately normal or the sample size should be large.
  - $n < 30$: 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.

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 Normal distribution. Enter the ***t*-distribution**!

As you can seen from Figure 12.2, 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!)

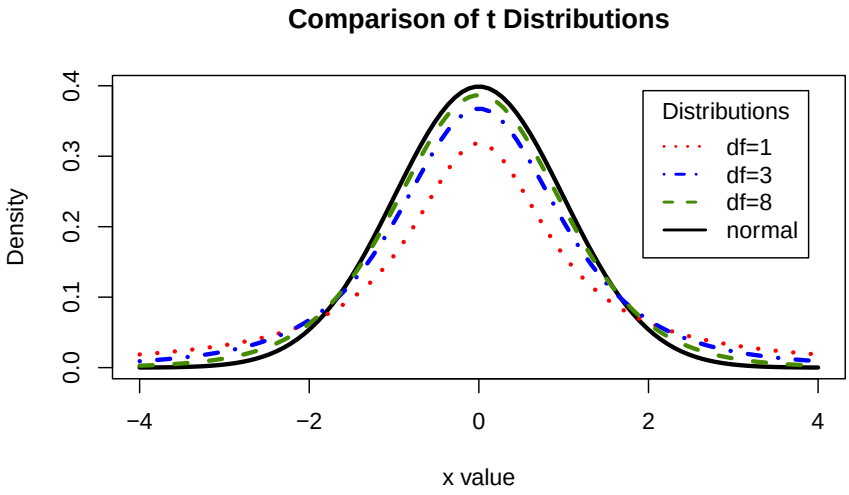


Figure 12.2: Comparison of the standard Normal vs t-distribution with various degrees of freedom

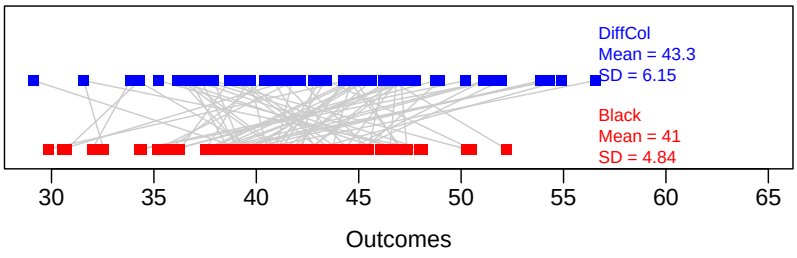
Summarize and visualize the data

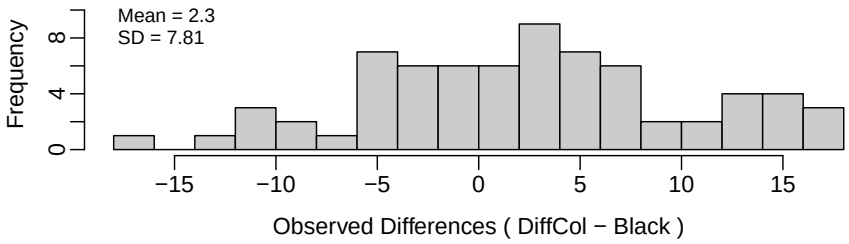
Since the original data from the study are not available, we simulated data to match the means and standard deviations reported in the article. We will use these simulated data in the analysis below.

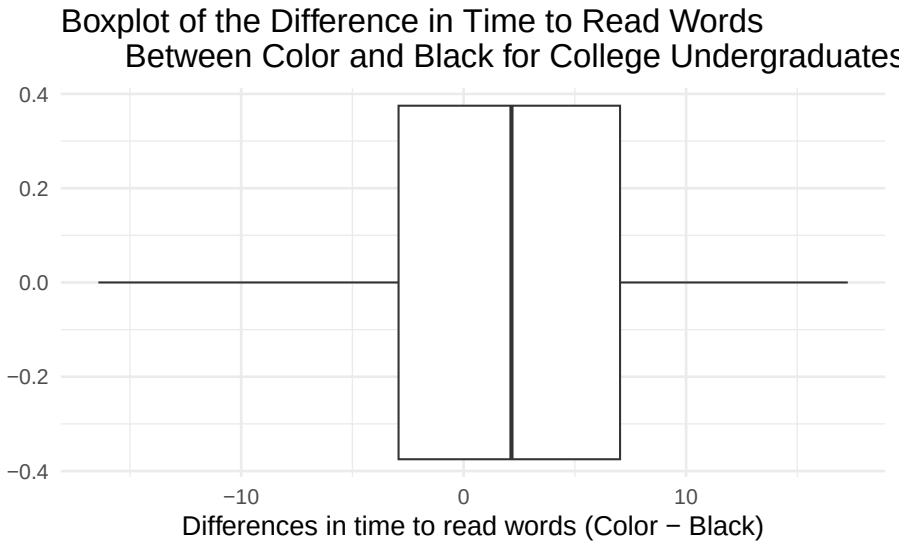
The following code plots each subject's time to read the colored words (above) and time to read the black words (below) connected by a grey line, a histogram of the differences in time to read words between the two conditions, and a boxplot displaying the pairwise differences in time (color − black).

```
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.

```
color_diff %>%
  summarise(favstats(differences))
#>      min     Q1 median     Q3   max mean       sd  n missing
#> 1 -16.42 -2.925   2.15 7.0325 17.27  2.3 7.810196 70       0
```

Check theoretical conditions

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

5. 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:

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

6. Calculate the standard error of the sample mean difference.

7. How many standard errors is the observed mean difference from the null mean difference?

To find the p-value we enter the value for the standardized statistic (T = 2.464) into the pt function in R. If you did not get his answer for question 7, double check your work. 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.

```
pt(2.464, df=69, lower.tail=FALSE)
#> [1] 0.008117801
```

8. Explain why we found the area above the T-score using `lower.tail = FALSE` in the code above.

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

On the next page 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()`. The code below will return the 95th percentile of the t distribution with `df` $= n_d - 1 = 70 - 1 = 69$.

```
qt(0.95, df = 69, lower.tail=TRUE)
#> [1] 1.667239
```

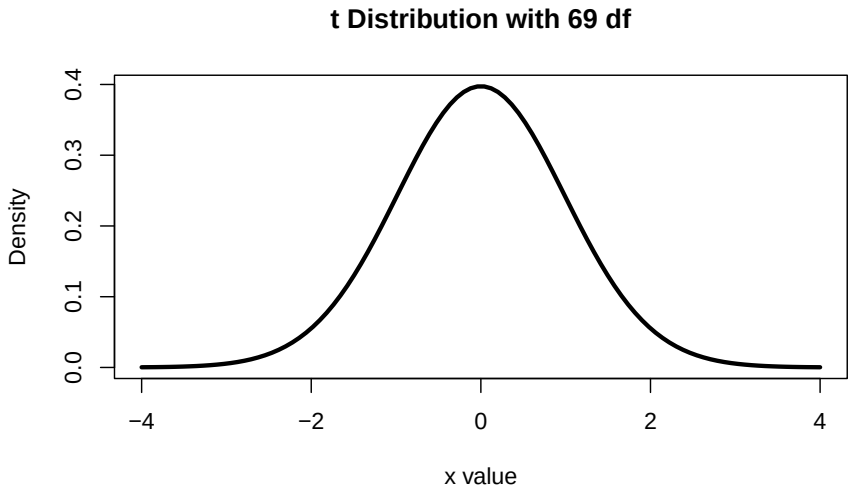


Figure 12.3: t-distribution with 69 degrees of freedom

10. In Figure 12.3, you see a t-distribution with 69 degrees of freedom. Label $t^\star$ and $-t^\star$ on that distribution. Write on the plot the percent of the t_{69}-distribution that is below $-t^\star$, between $-t^\star$ and $t^\star$, and above $t^\star$. Then use your plot to determine the confidence level associated with the $t^\star$ value obtained.

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. Do the results of the CI agree with the p-value? Explain your answer.

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

12.2.4 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.

12.2.5 Additional notes

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

12.3 Activity 12: COVID-19 and Air Pollution

12.3.1 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.

12.3.2 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.

12.3.3 COVID-19 and air pollution

In June 2020, the social distancing efforts and stay-at-home directives to help combat the spread of COVID-19 appeared to help 'flatten the curve' across the United States, albeit at a high cost to many individuals and businesses. The impact of these measures, though, goes far beyond the infection and death rates from the disease. You may have seen images comparing air quality in large international cities like Rome, Milan, Wuhan, and New Delhi such as the one pictured in Figure 12.4, which seem to indicate, perhaps unsurprisingly, that fewer people driving and factories being shut down have reduced air pollutants.

Have high population-density US cities seen the same improved air quality conditions? To study this question, data were gathered from the US Environmental Protection Agency (EPA) AirData website which records the ozone (O3) and fine particulate matter (PM2.5) values for cities across the US (US Environmental Protection Agency, n.d.). These measures are used to calculate an air quality index (AQI) score for each city each day of the year. Thirty-three of the most densely populated US cities were selected and the AQI score recorded for April 20, 2020 as well as the five-year median AQI score for April 20th (2015–2019). Note that higher AQI scores indicate worse air quality. A box plot of the differences in AQI scores for the 33 cities and a table of summary statistics are shown on the next page. Use Current - 5-year median as the order of subtraction.

Figure 12.4: The India Gate in New Delhi, India. Source: Reuters/Anushree Fadnavis/Adnan Abidi

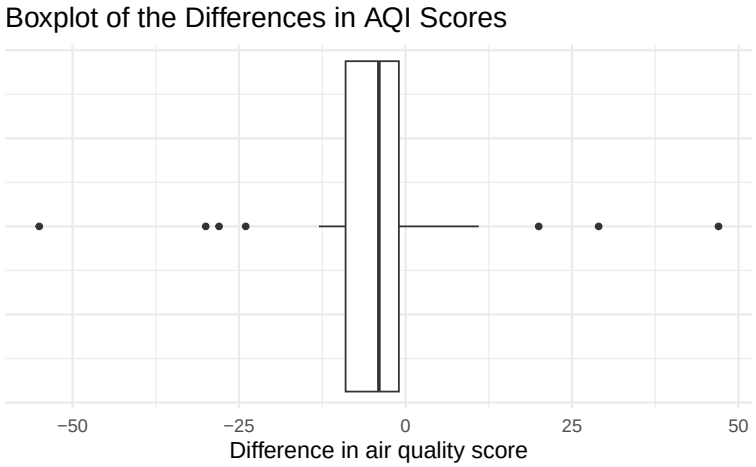


Table 12.1: Summary statistics for current AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

	Mean	Standard deviation	Sample size
Current	$\bar{x}_1 = 47.394$	$s_1 = 14.107$	$n_1 = 33$
5 Year Median	$\bar{x}_2 = 51.545$	$s_2 = 17.447$	$n_2 = 33$
Differences	$\bar{x}_d = -4.152$	$s_d = 17.096$	$n_d = 33$

Vocabulary review.

1. Identify the variables in this study. What role (explanatory or response) do each have?

2. Are the differences in AQI scores independent for each case (US city)? Explain.

3. Why is this treated as a paired study design and not two independent samples?

Ask a research question

4. Write the null hypothesis in words.

5. What is the research question?

6. Write the alternative hypothesis in notation.

Summarize and visualize the data

7. Report the summary statistic of interest (mean difference) for the data.

8. What notation is used for the value in question 7?

Use statistical inferential methods to draw inferences from the data

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.

9. 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 9 for `xx` to simulate the null distribution and enter the summary statistic from question 7 for `yy` to find the p-value.
- Highlight and run lines 1–24.

```
paired_test(data = Air$Difference,    # 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 = 1000,  # Number of simulated samples for null distribution
        which_first = 1)  # Not needed when using calculated differences
```

10. Sketch the null distribution created using the R output here.

11. Explain why the null distribution is centered at zero.

12. What proportion of samples are at or less than the observed sample mean difference in AQI scores for current scores minus 5 year median scores? What is the statistical term for this proportion?

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

14. How much evidence does this provide for improved air quality in US cities?

15. If evidence was found for improved air quality in US cities, could we conclude that the stay-at-home directives *caused* the improvement in air quality? 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.

16. Write out the parameter of interest in context of the study.

17. Using the provided R script file, fill in the missing value at `xx` to find a 99% bootstrap confidence interval; highlight and run lines 29–32. Report the confidence interval in interval notation.

```
paired_bootstrap_CI(data = Air$Difference, # Enter vector of differences
                    number_repetitions = 1000, # Number of bootstrap samples for CI
                    confidence_level = xx,  # Confidence level in decimal form
                    which_first = 1)  # Not needed when entering vector of differences
```

Communicate the results and answer the research question

18. Interpret the 99% confidence interval in the context of the problem.

19. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

12.3.4 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.

[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 Week 12.

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.

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 Module 12 Lab: Swearing

12.4.1 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.

12.4.2 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 Week 12 lab.
- Upload and import the csv file, `pain_tolerance`.
- Enter the name of the data set (see the environment tab) for datasetname 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 <- datasetname
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 `group_1` and the outcome for group 2 (`Neutral`) for `group_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 = group_1 - group_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 1000 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 1000 simulations. Highlight and run lines 23–29.

```
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 34–37. **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 = 1000, # 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)?

MODULE 13

Inference for a Quantitative Response with Independent Samples

13.1 Lecture Notes Module 13: Inference for Independent Samples

Single categorical, single quantitative variable with independent samples

- In this week, 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 _____ 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.

- The summary measure for two independent groups is the _____ in _____.
- Difference in means: the difference in average _____ variable outcome for observational units between _____ variable groups

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

$\mu_1 - \mu_2$:

Notation for the Sample Statistics

- Sample mean for group 1:
- Sample mean for group 2:
- Sample difference in means:
- Sample standard deviation for group 1:
- Sample standard deviation for group 2:
- Sample size for group 1:
- Sample size for group 2:

Example for class discussion: 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?

Why should we treat this as two independent groups rather than paired data?

Parameter of interest:

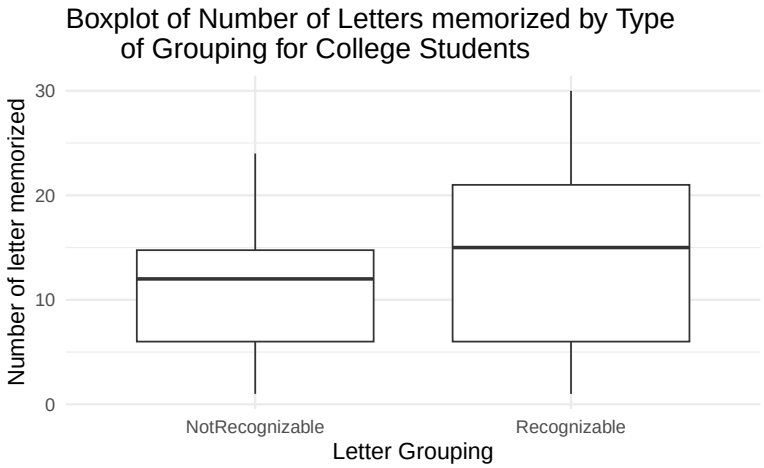
```
letters<-read.csv("data/letters.csv")
letters %>%
    reframe(favstats(Memorized~Grouped))
#>           Grouped min Q1 median    Q3 max     mean       sd  n missing
#> 1 NotRecognizable   1  6     12 14.75  24 11.15385 6.576883 26       0
#> 2    Recognizable   1  6     15 21.00  30 14.32000 8.518216 25       0
```

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 letter memorized", #y-axis label
       x = "Letter Grouping") #x-axis label
```



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 words:

H_0 :

H_A :

In notation:

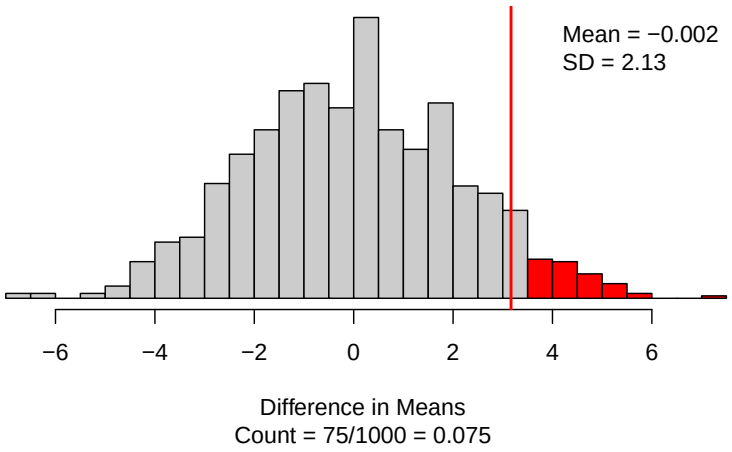
H_0 :

H_A :

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 1000 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 = 1000,  # Number of simulations
              as_extreme_as = 3.166,  # Observed statistic
              direction = "greater")  # Direction of alternative: "greater", "less", or "two-sided"
```



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

Theory-based method

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:

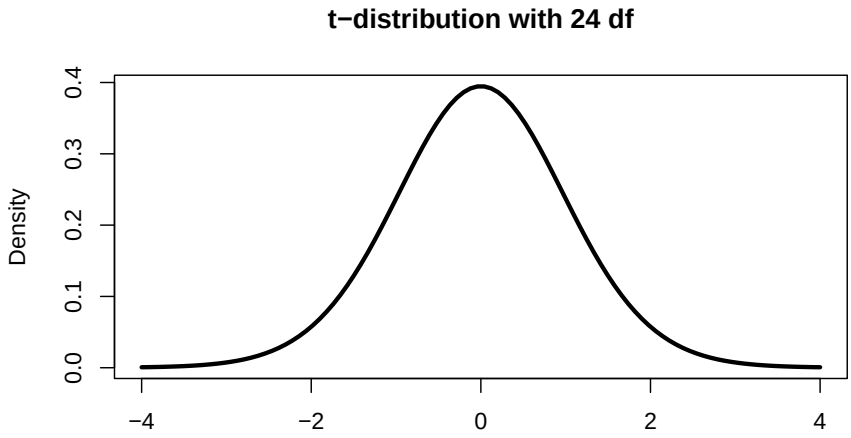
Equation for the standardized difference in sample mean:

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

Calculate the standardized difference in memorized letters.

- First calculate the $SE(\bar{x}_1 - \bar{x}_2)$

- Then calculate the T-score



Interpret the standardized statistic:

To find the theory-based p-value:

```
pt(1.482, df=24, lower.tail=FALSE)*2
#> [1] 0.1513523
```

Confidence interval

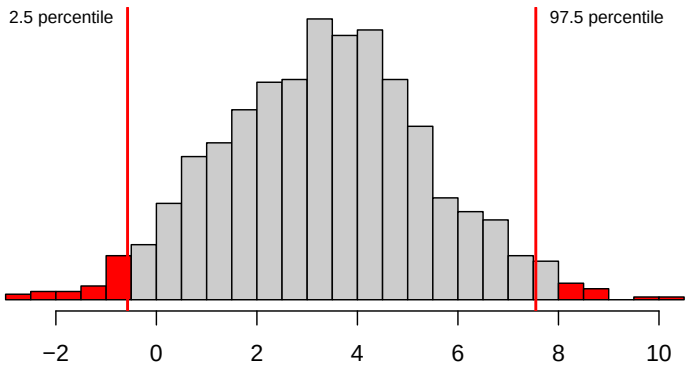
To estimate the difference in true mean we will create a confidence interval.

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

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 = 1000,  # Number of simulations
                      confidence_level = 0.95)
```



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

Using the letters data, calculate the 99% confidence interval.

```
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
```

- Need the t^* multiplier for a 99% confidence interval from a t-distribution with _____ df.

```
qt(0.995, df=24, lower.tail = TRUE)
#> [1] 2.79694
```

- We will use the same value for the $SE(\bar{x}_1 - \bar{x}_2)$ as calculated for the standardized statistic.

13.2 Out-of-Class Activity Module 13: Does behavior impact performance?

13.2.1 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 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.

13.2.2 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

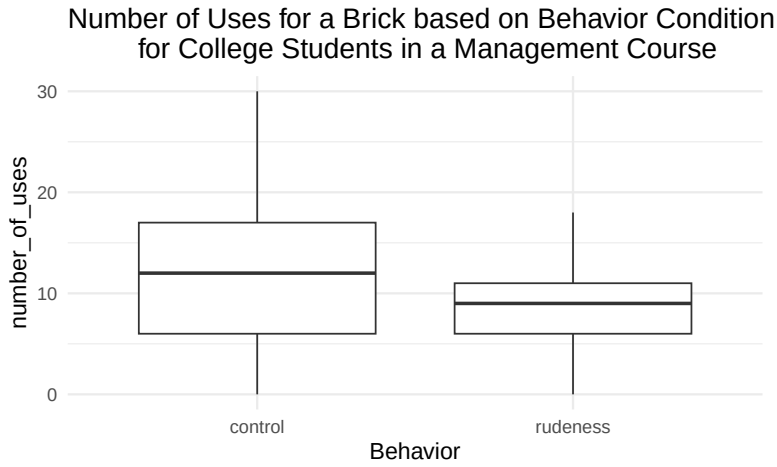
To review these concepts, see Chapter 19 in the textbook.

13.2.3 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.

```
# Read in data set
rude <- read.csv("https://math.montana.edu/courses/s216/data/rude.csv")
```

```
# 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
```

Quantitative variables review

1. The two variables assessed in this study are behavior and number of uses for a brick. Identify the role for each variable (explanatory or response).

2. Which group (control or rudeness) has the highest center in the distributions of number of uses for a brick? Explain which measure of center you are using.

3. Using the side-by-side box plots, which group has the largest spread in number of uses for a brick? How did you make that choice?

4. Is this an experiment or an observational study? Justify your answer.

5. Is this a paired data set or two independent groups? Explain your reasoning.

Ask a research question

6. Write out the parameter of interest in context of the study. Use proper notation and be sure to define your subscripts.

7. Write out the null hypothesis in words.

8. Write the alternative hypothesis in notation.

Summarize and visualize the data

9. Calculate the summary statistic of interest (difference in means). What is the appropriate notation for this statistic?

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.

10. How many cards would we start with?

11. What would we write on each card?

12. Next, we would mix the cards together and shuffle into two piles. How many cards will go into each pile? What should we label the piles?

13. What value would be calculated from the cards and plotted on the null distribution? *Hint*: What statistic are we calculating from the data?

14. Would you expect your simulated statistic to be closer to the null value of zero than the difference in means calculated from the sample? Explain why this makes sense.

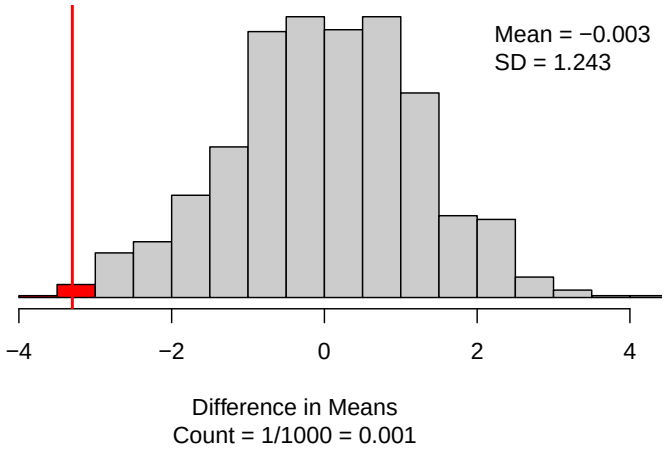
We will use the `two_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of differences in sample means and compute a p-value.

15. When using the `two_mean_test()` function, we need to enter the name of the response variable, `number_of_uses`, and the name of the explanatory variable, `condition`, for the formula. The name of the data set as shown above is `rude`. What values should be entered for each of the following to create 1000 simulated samples?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"control"` or `"rudeness"`):
- Number of repetitions:
- As extreme as:
- Direction (`"greater"`, `"less"`, or `"two-sided"`):

The code below will simulate a null distribution and compute the p-value. Check that your answers from question 15 match what is entered below.

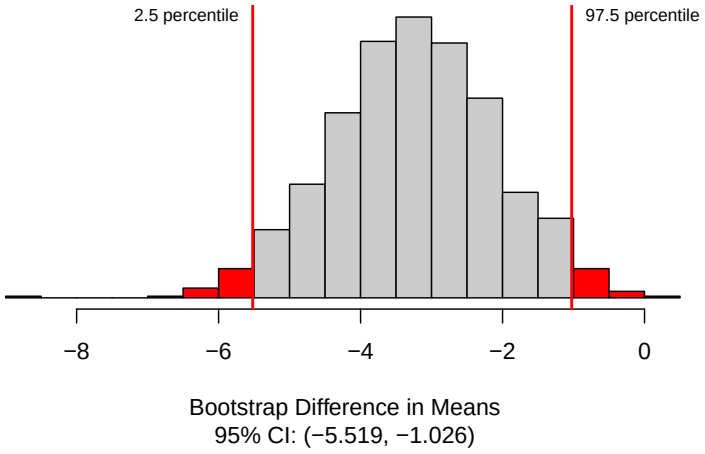
```
set.seed(216)
two_mean_test(number_of_uses~condition, #Enter the names of the variables
              data = rude,  # Enter the name of the dataset
              first_in_subtraction = "rudeness", # First outcome in order of subtraction
              number_repetitions = 1000,  # Number of simulations
              as_extreme_as = -3.3,  # Observed statistic
              direction = "less")  # Direction of alternative: "greater", "less", or "two-sided"
```



16. Report the p-value. Based off of this p-value, write a conclusion to the hypothesis test.

Confidence interval We will use the `two_mean_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample means and calculate a 95% confidence interval.

```
set.seed(216)
two_mean_bootstrap_CI(number_of_uses ~ condition, #Enter the name of the variables
                      data = rude,  # Enter the name of the data set
                      first_in_subtraction = "rudeness", # First value in order of subtraction
                      number_repetitions = 1000,  # Number of simulations
                      confidence_level = 0.95)
```



17. Report the 95% confidence interval.

18. Interpret the interval reported in question 17.

13.2.4 Take-home messages

1. This activity differs from the activities in Week 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.

13.2.5 Additional notes

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

13.3 Activity 13: The Triple Crown

13.3.1 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.

13.3.2 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.

13.3.3 The triple crown

The Triple Crown of "Thru" hiking consists of hiking the Appalachian Trail, the Pacific Crest Trail (PCT), and the Continental Divide Trail (CDT). Each year halfwayanywhere.com conducts a survey to better understand the people who hike these trails. One variable which is queried in the survey is the pre-hike "base weight" of a hiker's pack which is the total weight of gear without food, water, and worn gear. The 131 hikers surveyed who completed the CDT had a mean base weight of 15.266 lbs (sd = 5.128 lbs). The 484 hikers surveyed who completed the PCT had a mean base weight of 17.837 lbs (sd = 7.823 lbs). Is there a difference in average base weight for PCT hikers and CDT hikers? Use order of subtraction CDT - PCT.

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
- **Normality**: Each sample should be approximately normal or have a large sample size. For *each* sample:
  - $n < 30$: 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 provided R script file.
- Upload and import the csv file, `Trail_Weight`.
- Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 10.
- Write a title for the boxplots in line 15.
- Highlight and run lines 1–16 to load the data and create plots of the data.

```
hikes <- datasetname
hikes %>%  # Data set piped into...
  ggplot(aes(y = Baseweight, x = Trail))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "xx",  # Title
       x = "Trail",    # x-axis label
       y = "Baseweight(lbs)")  # 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.

```
hikes %>%
  reframe(favstats(response~explanatory))
```

6. **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:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

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 28. Report the p-value calculated.

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

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 (as a decimal) for `xx` and degrees of freedom for `yy` into the `qt()` function at line 34 to find the appropriate t^* multiplier

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

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?

13.3.4 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.

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 Module 13 Lab: Dinosaurs

13.4.1 Learning outcomes

- Identify whether a study is a paired design or independent groups
- 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 hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a hypothesis test for a difference in means.
- Find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.

13.4.2 Type of 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.

13.4.3 Dinosaurs

The backbone of heavy, two-legged, carnivorous dinosaurs, such as the *T. rex*, is subject to stress. Intriguingly, these dinosaurs have protrusions (rugosity) at the top and sides of their spinal vertebrae, potentially for extra support. These protrusions do not seem to be present in smaller carnivorous dinosaurs. MSU paleontologists hypothesize that the presence of the protrusions is associated with the size of the two-legged carnivorous dinosaurs, potentially allowing them to grow big (Wilson 2016). To test this hypothesis, the researchers collected multiple scientific papers describing the fossil bones of 57 two-legged carnivorous dinosaur species. Then, they checked for the presence or absence of these rugose protrusions from photographs published in the papers and collected measurements of the length in centimeters of the femur (or thigh) bone. Femur length is a proxy for dinosaur size. Is there evidence that the presence of the protusions result in larger dinosaurs? Use present − absent as the order of subtraction.

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

5. 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, `dinosaur`.
- Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 10 and the name of the explanatory and response variable in line 12.
- Highlight and run lines 1–16 to load the data and create a plot of the data.

```
dinos <- datasetname
dinos %>%
  ggplot(aes(y = response, x = explanatory)) +
  geom_boxplot() +
  labs(title = "Side-by-side Box Plots of Femur Length by Rugosity
       for Carnivorous Dinosaur",
       x = "Rugose structures on the backbone",
       y = "Femur length (cm)")
```

6. Based on the plots, does there appear to be some evidence in favor of the alternative hypothesis? How do you know?

- Enter the name of the explanatory variable for `explanatory` and the response variable for `response` in line 23.
- Run lines 22–23 to find the summary statistics.

```
dinos %>%
    summarize(favstats(response~explanatory))
```

7. Calculate the summary statistic for the research question. Use proper notation.

8. **How far, on average, is each difference in mean femur length from the difference in true mean femur length? 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 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 there is no difference in true mean femur length for dinosaurs with protrusions and dinosaurs without protrusions.

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.

10. Simulate a null distribution and compute the p-value.

- Using the R script file for this lab, enter the correct values in place of the `xx`'s to produce the null distribution with 1000 simulations.
- Highlight and run lines 28–33.

```
two_mean_test(response~explanatory, #Enter the names of the variables
              data = dinos,  # Enter the name of the dataset
              first_in_subtraction = "xx", # First outcome in order of subtraction
              number_repetitions = 1000,  # Number of simulations
              as_extreme_as = xx,  # Observed statistic
              direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

Sketch the null distribution created using the `two_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 `two_mean_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample differences in 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 `two_mean_bootstrap_CI()` function to create the 98% confidence interval.
- Highlight and run lines 38–42. **Upload a copy of the bootstrap distribution created to Gradescope for your group.**

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = dinos,  # Enter the name of the data set
                      first_in_subtraction = "xx", # First value in order of subtraction
                      number_repetitions = 1000,  # Number of simulations
                      confidence_level = xx)
```

Report the 98% confidence interval in interval notation.

14. Write a paragraph summarizing the results of this study as if you were describing the results to your roommate. **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)?

Paragraph:

MODULE 14

Inference for Two Quantitative Variables

14.1 Lecture Notes Module 14: Inference for Two Quantitative Variables

Summary measures and plots for two quantitative variables.

Scatterplot:

- Form: linear or non-linear?
- Direction: positive or negative?
- Strength: how clear is the pattern between the two variables?
- Outliers: points that are far from the pattern or bulk of the data
  - Influential points: outliers that are extreme in the x- variable.

The summary measures for two quantitative variables are:

- _____, interpreted as the on average change in the response variable for a one unit increase in the explanatory variable.
- _____, which measures the strength and direction of the linear relationship between two quantitative variables.
- _____, interpreted as the percent of variability in the response variable that is explained by the relationship with the explanatory variable.
- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem)

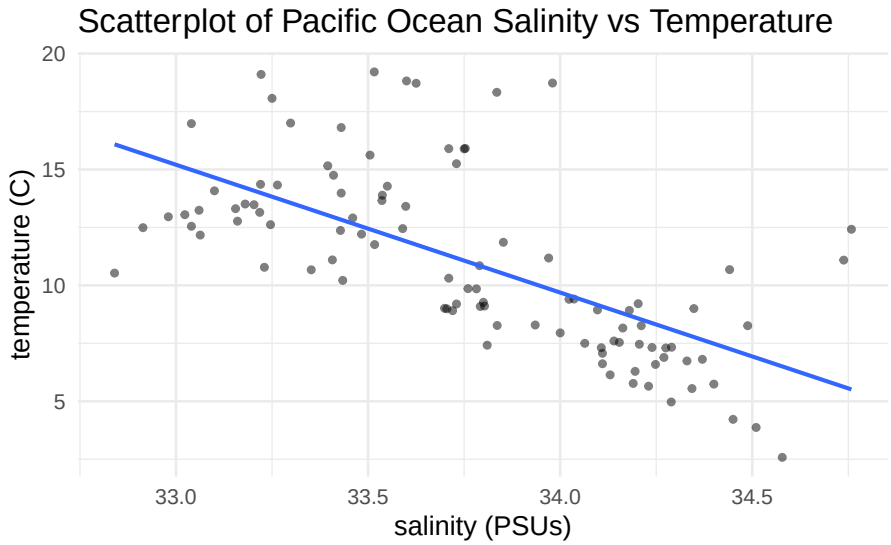
Notation:

- Population slope:
- Population correlation:
- Sample slope:
- Sample correlation:

Example for class discussion: 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?

```
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
```



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:

Hypothesis Testing

Conditions:

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

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 words:

H_0 :

H_A :

In notation:

H_0 :

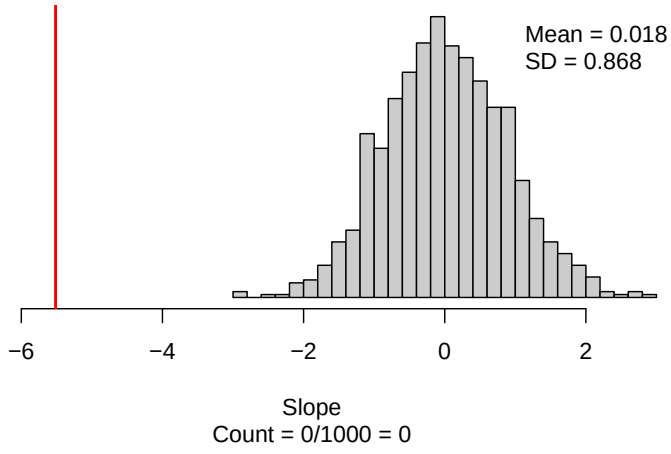
H_A :

Simulation-based method

- 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 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as b_1 or r

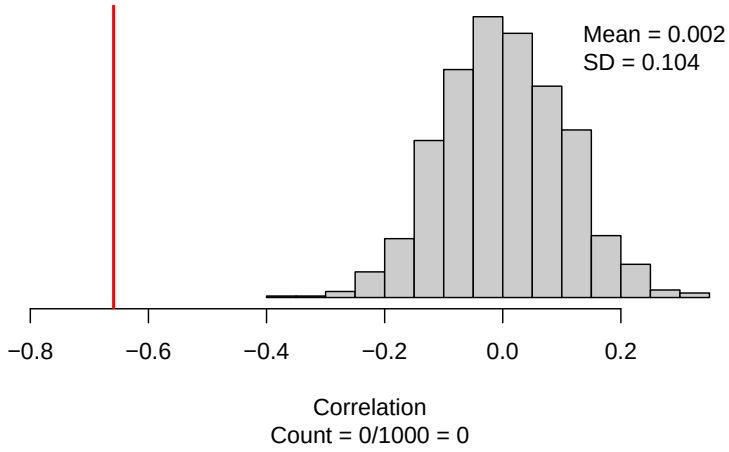
To test slope:

```
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 = 1000) # Number of simulated samples for null distribution
```



To test correlation:

```
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 = 1000) # Number of simulated samples for null distribution
```



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

Theory-based method

Conditions:

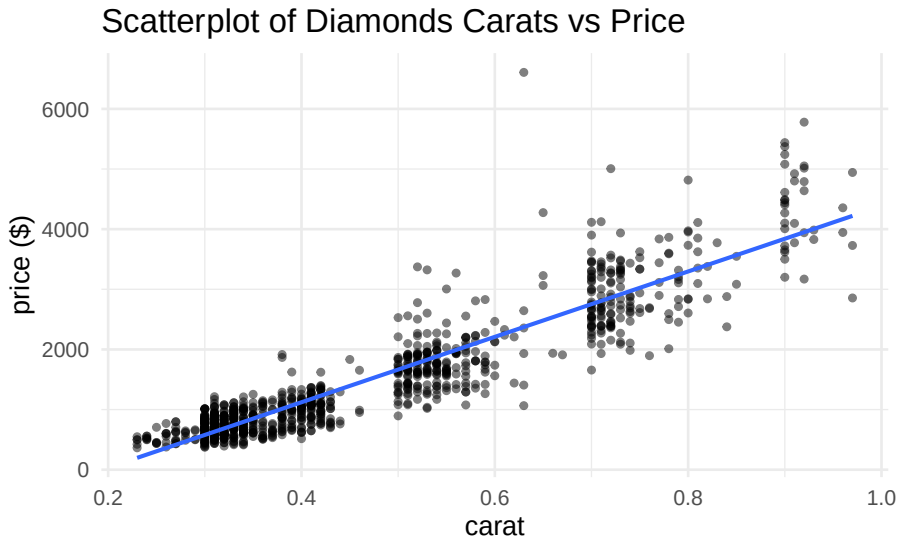
- Linearity (for both simulation-based and theory-based methods): the data should follow a linear trend.
  - Check this assumption by examining the _____ of the two variables, and _____. The pattern in the residual plot should display a horizontal line.
- Independence (for both simulation-based and theory-based 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:

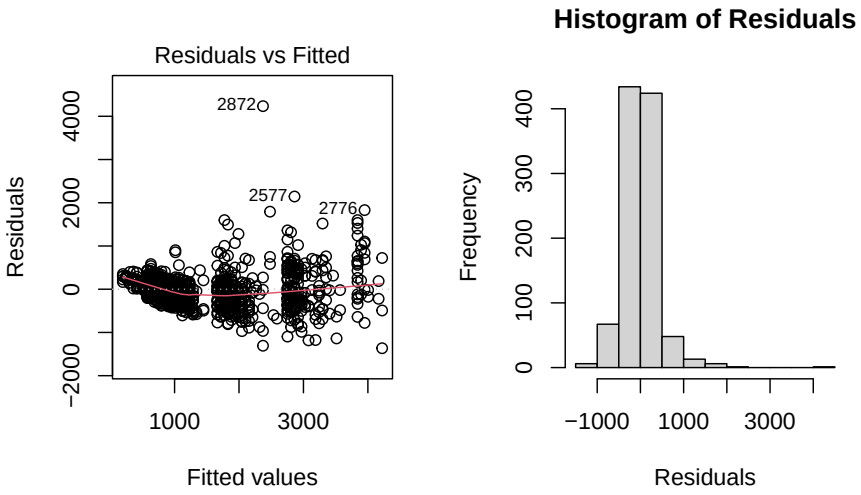
It is a generally accepted fact that the more carats a diamond has, the more expensive that diamond will be. The question is, how much more expensive? Data on thousands of diamonds were collected for this data set. We will only look at one type of cut ("Ideal") and diamonds less than 1 carat. Does the 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
```



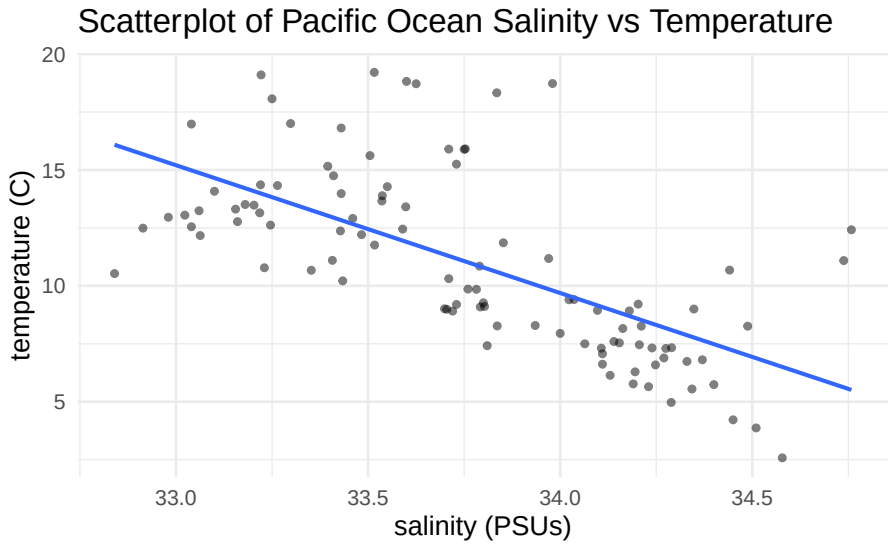
Diagnostic plots:



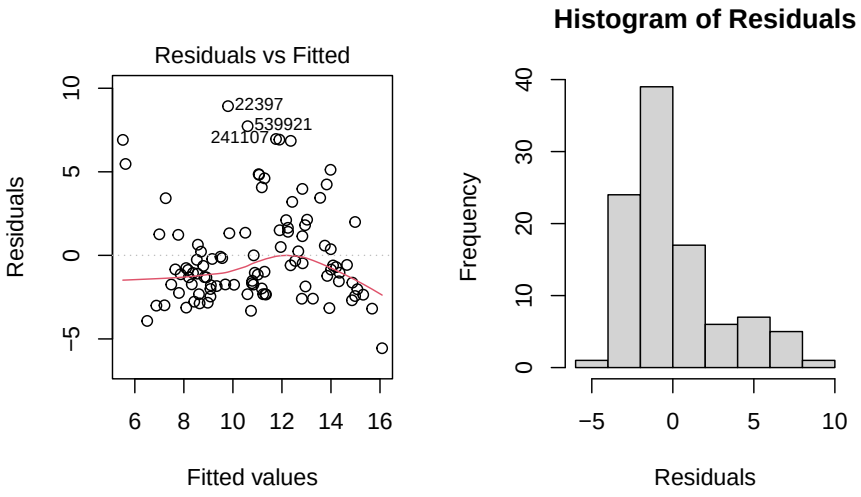
Check the conditions for the ocean data:

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.

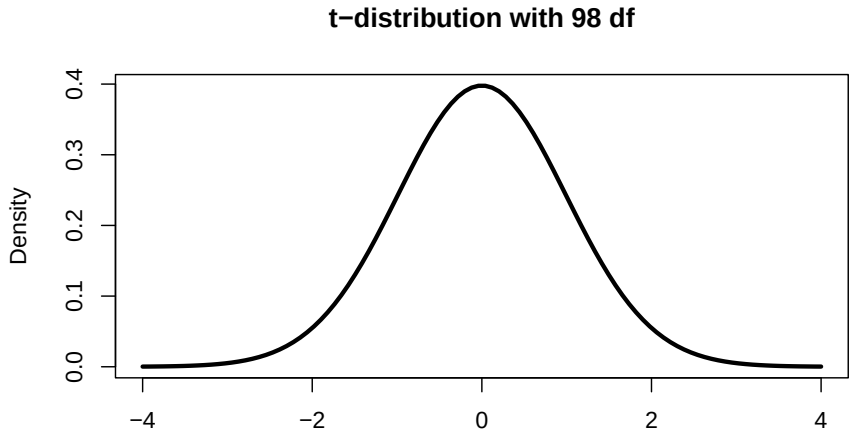
Theory-based test:

- 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)
summary(lm.water)$coefficients
#>               Estimate Std. Error   t value     Pr(>|t|)
#> (Intercept) 197.156160 21.4778118  9.179527 7.304666e-15
#> Salnty       -5.513691  0.6359673 -8.669770 9.257446e-14
```



Interpret the standardized statistic:

To find the theory-based p-value:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
summary(lm.water)$coefficients
#>               Estimate Std. Error   t value     Pr(>|t|)
#> (Intercept) 197.156160 21.4778118  9.179527 7.304666e-15
#> Salnty       -5.513691  0.6359673 -8.669770 9.257446e-14
```

or

```
pt(-8.670, df = 98, lower.tail=TRUE)
#> [1] 4.623445e-14
```

Confidence interval

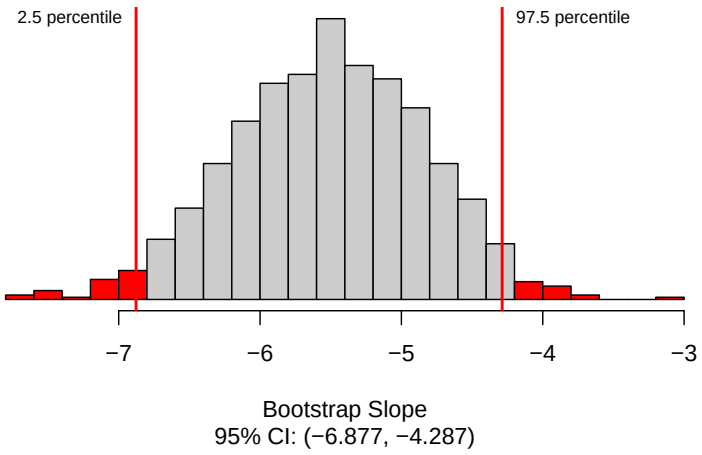
To estimate the true slope (or true correlation) we will create a confidence interval.

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 1000 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 = 1000) # Number of simulated samples for bootstrap distribution
```

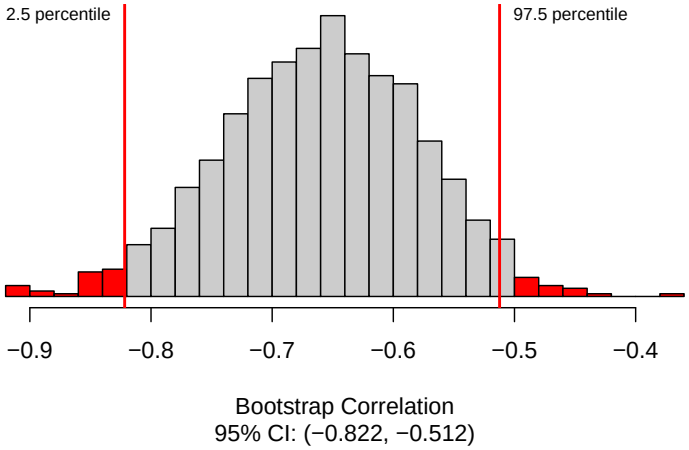


Confidence interval interpretation:

- 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.

```
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 = 1000) # Number of simulated samples for bootstrap distribution
```



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

```
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
```

14.2 Out-of-Class Activity Module 14: Prediction of Crocodilian Body Size

14.2.1 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.

14.2.2 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.

14.2.3 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?

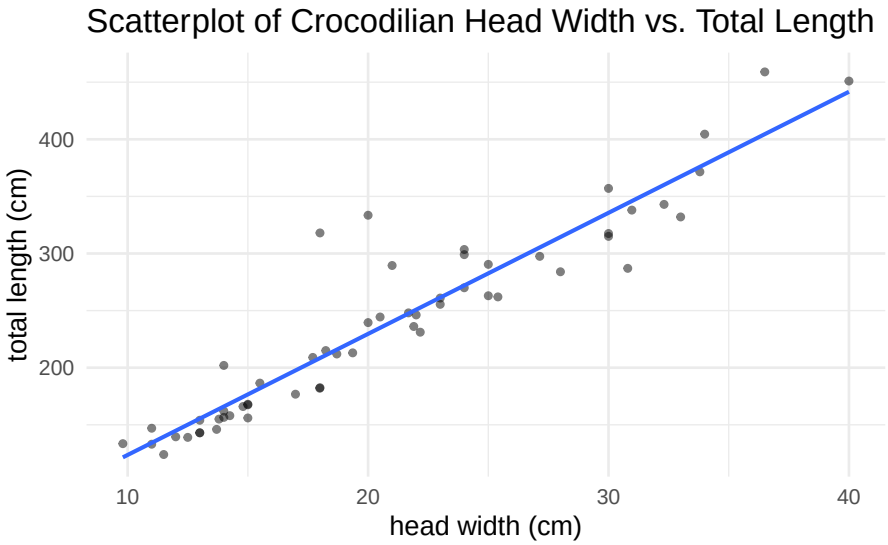
```
# Read in data set
croc <- read.csv("https://math.montana.edu/courses/s216/data/Crocodylian_headwidth.csv")
croc <- croc %>%
    na.omit()
```

Vocabulary review

1. Explain why regression methods are appropriate to use to address the researchers' question. Make sure you clearly define the variables of interest in your explanation and their roles.

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.

```
croc %>% # Pipe data set into...
ggplot(aes(x = HW_cm, y = TL_cm))+  # 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
```



2. Describe the features of the plot above, addressing all four characteristics of a scatterplot.

If you indicated there are potential outliers, which points are they?

Ask a research question

3. Write out the null hypothesis in words to test **slope**.

4. Using the research question, write the alternative hypothesis in notation using **slope** as the summary measure.

Summarize and visualize the data

The linear model output for the data is given below.

```
lm.croc <- lm(TL_cm~HW_cm, data=croc) #lm(response~explanatory)
round(summary(lm.croc)$coefficients, 5)
#>             Estimate Std. Error  t value Pr(>|t|)
#> (Intercept) 17.61250   11.36269  1.55003  0.12687
#> HW_cm       10.59983    0.51294 20.66494  0.00000
```

The value of correlation is given below.

```
cor(croc$HW_cm, croc$TL_cm)
#> [1] 0.9412234
```

5. Using the output from the evaluated R code above, write the equation of the regression line in the context of the problem using appropriate statistical notation.

6. Interpret the estimated slope in context of the problem.

7. Report the value of correlation between head width and total body length.

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.

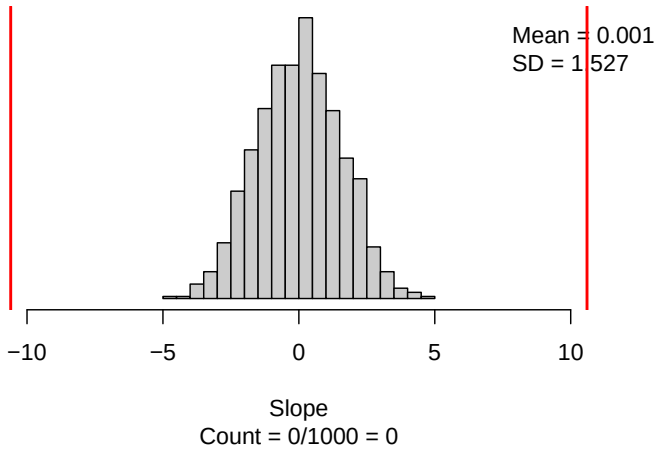
The response variable name is `TL_cm` and the explanatory variable name is `HW_cm` for these data.

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:

Check that your answers to question 8 reflect what is shown below in the R code to produce the null distribution for slope.

```
set.seed(216)
regression_test(TL_cm~HW_cm, # response ~ explanatory
                data = croc, # Name of data set
                direction = "two-sided", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "slope", # "slope" or "correlation"
                as_extreme_as = 10.600, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

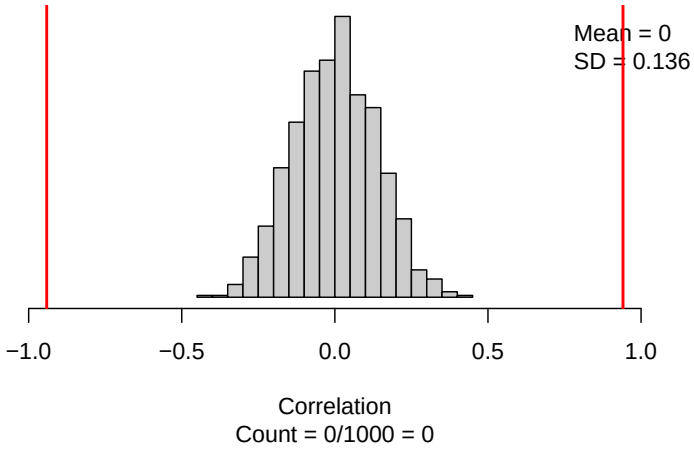


9. Report the p-value from the R output.

10. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which two inputs in question 8 would need to be changed to test for correlation? What inputs should you use instead?

Check that your answers to question 10 reflect what is shown below in the R code to produce the null distribution for correlation.

```
set.seed(216)
regression_test(TL_cm~HW_cm, # response ~ explanatory
                data = croc, # Name of data set
                direction = "two-sided", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "correlation", # "slope" or "correlation"
                as_extreme_as = 0.941, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

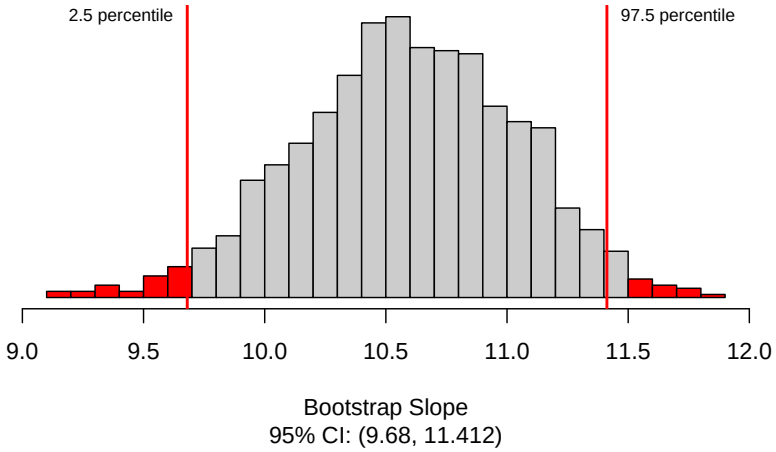


11. The p-values from the test of slope and the test of correlation should be similar. Explain why the two p-values should match. *Hint: think about the relationship between slope and correlation!*

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. The following code gives the 95% confidence interval for the true slope.

```
set.seed(216)
regression_bootstrap_CI(TL_cm~HW_cm, # response ~ explanatory
   data = croc, # Name of data set
   confidence_level = 0.95, # Confidence level as decimal
   summary_measure = "slope", # Slope or correlation
   number_repetitions = 1000) # Number of simulated samples for bootstrap distribution
```



12. Report the bootstrap 95% confidence interval in interval notation.

13. Interpret the interval in question 12 in context of the problem. *Hint: use the interpretation of slope in your confidence interval interpretation.*

Communicate the results and answer the research question

14. Based on the p-value, write a conclusion in context of the problem.

15. Does the conclusion based on the p-value agree with the results of the 95% confidence interval? What does each tell you about the null hypothesis?

14.2.4 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.

14.2.5 Additional notes

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

14.3 Activity 14: Golf Driving Distance

14.3.1 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.

14.3.2 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.

14.3.3 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?"

```
# Read in data set
golf <- read.csv("https://math.montana.edu/courses/s216/data/golf.csv")
```

Plot review.

- Use the provided R script file to create a scatterplot to examine the relationship between the driving distance and percent accuracy by filling in the variable names (`Driving_Distance` and `Percent_Accuracy`) for `explanatory` and 'response" in line 9.
- Highlight and run lines 1–17.

```
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
```

1. Sketch the plot created below. Based on your plot, does it appear that there is a relationship between driving distance and percent accuracy? Note: `Driving Distance` should be on the x-axis.

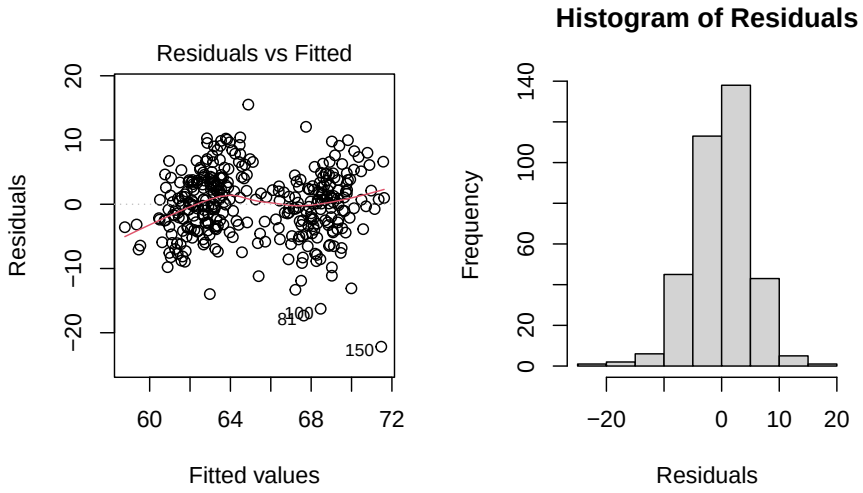
Conditions for the least squares line

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[1].

[1] A better plot for checking the normality assumption is called a *normal quantile-quantile plot* (or QQ-plot). However, this type of plot will be covered in a future course

The scatterplot generated in question 1 and the residual plots shown below will be used to assess these conditions for approximating the data with the t-distribution.



2. 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

3. Write out the null hypothesis in words to test the slope.

4. Using the research question, write the alternative hypothesis in notation to test the slope.

Summarize and visualize the data

- Using the provided R script file, enter the response variable name, `Percent_Accuracy`, into the `lm()` (linear model) function for `response` and the explanatory variable name, `Driving_Distance`, for `explanatory` in line 30 to get the linear model output.
- Highlight and run lines 30–31.

```
lm.golf <- lm(response~explanatory, data=golf) # lm(response~explanatory)
round(summary(lm.golf)$coefficients, 5)
```

5. Using the output from the evaluated R code above, write the equation of the regression line in the context of the problem using appropriate statistical notation.

6. Interpret the estimated slope in context of the problem.

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.

7. What are the values of b_1 and $SE(b_1)$? Where in the linear model R output can you find these values?

8. Calculate the standardized statistic for slope. Identify where this calculated value is in the linear model R output.

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

10. 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.

11. 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. Recall for a 95% confidence interval, we use the 97.5% percentile (95% of the distribution is in the middle, leaving 2.5% in each tail). The sample size for this study is 354 so we will use the degrees of freedom 352 $(n-2)$.

```
qt(0.975, 352, lower.tail = TRUE) # 95% t* multiplier
```

```
#> [1] 1.966726
```

12. Calculate the 95% confidence interval for the true slope.

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

Communicate the results and answer the research question

14. 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.

- Highlight and run lines 38–46 to produce the multivariable plot.

```
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_viridis_d(end=0.8)
```

15. Does the association between driving distance and percent accuracy change depending on which league the golfer is a part of? Explain your answer.

16. Explain the association between league and each of the other two variables.

14.3.4 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.

14.3.5 Additional notes

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

14.4 Module 14 Lab: Big Mac Index

14.4.1 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.

14.4.2 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 and import the csv file, `big_mac_adjusted_index_S22.csv`.
- Enter the name of the data set (see the environment tab) 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 <- 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.

```r
# Fit linear model: y ~ x
bigmacLM <- lm(response~explanatory, data=mac)
summary(bigmacLM)$coefficients # 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 28–34 to create the scatterplot to check for linearity.

```r
#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.

```r
#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 country 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 1000 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 = 1000) # 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.

```
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 = 1000) # 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 15

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 1 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
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. Use the appropriate Final Exam Review R script file to analyze the following research question, "Does drinking a caffeinated drink increase blood glucose levels, 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?

2. Use the appropriate Final Exam Review R script file to analyze the following research question: "Do Islanders who listen to classical music take less time 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?

3. Use the appropriate Final Exam Review R script file to analyze the following research question: "Can height be used to predict the balance time for Islanders?"

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	Paired Differences	Two Quantitative Variables	Quant. Response and Categ. Explanatory (independent samples)
Type of plot	Bar plot	Segmented bar plot, Mosaic plot	Dotplot, histogram, boxplot	Scatterplot	Side-by-sided boxplots, Stacked dotplots or histograms
Summary measure	Proportion	Difference in proportions	Mean Difference	Slope or correlation	Difference in means
Parameter notation	π	$\pi_1 - \pi_2$	μ_d	β_1 or ρ	$\mu_1 - \mu_2$
Statistic notation	$\hat{p}$	$\widehat{p_1} - \widehat{p_2}$	$\bar{x}_d$	b_1 or r	$\bar{x}_1 - \bar{x}_2$
Null hypothesis	$H_0: \pi = \pi_0$	$H_0: \pi_1 - \pi_2 = 0$	$H_0: \mu_d = 0$	$H_0: \beta_1 = 0$ or $H_0: \rho = 0$	$H_0: \mu_1 - \mu_2 = 0$
Conditions for simulation methods	Independent cases;	Independence (within and between groups);	Independent cases;	Independent cases; Linear form;	Independence (within and between groups);
Simulation test (how to generate a null distn) p-value = proportion of null simulations at or beyond (H_A direction) the observed statistic	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 1000's of 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 1000's of times. Centered at 0.	Shift the original data by adding $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Plot sample mean difference. Repeat 1000's of times. Centered at 0.	Hold the x values constant; shuffle new y's to x's. Find the regression line for shuffled data; plot the slope or the correlation for the shuffled data. Repeat 1000's of 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 1000's of times. Centered at 0.
Bootstrap CI (how to generate a boot. distn) X% CI: $(\frac{1-X}{2}\%tile,$ $\left(X + \frac{1-X}{2}\right)\%tile)$	Label n cards with the original responses. Randomly draw with replacement n times. Plot the resampled proportion of successes. Repeat 1000's of times. Centered at $\hat{p}$.	Label $n_1 + n_2$ cards with the original responses. 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 1000's of times. Centered at $\widehat{p_1} - \widehat{p_2}$	Label n cards with the original responses. Randomly draw with replacement n times. Plot the resampled mean difference. Repeat 1000's of times. Centered at $\bar{x}_d$.	Label n cards with the original (response, explanatory) values. Randomly draw with replacement n times. Plot the resampled slope or correlation. Repeat 1000's of times. Centered at b_1 or r.	Label $n_1 + n_2$ cards with the original responses. Randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Plot the resampled difference in means. Repeat 1000's of times. Centered at $\bar{x}_1 - \bar{x}_2$.
Theory-based distribution	Standard Normal	Standard Normal	t- distribution with $n - 1$ df	t- distribution with $n - 2$ df	t- distribution with min of n_1-1 or n_2-1 df
Conditions for theory-based hypothesis tests and confidence intervals	Independent cases; Number of successes and number of failures in the 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 $30 \le n < 100$ with no extreme outliers OR $n \ge 100$	Linear form; Independent cases; Nearly normal residuals; Variability around the regression line is roughly constant.	Independent cases (within and between groups); In each sample, $n < 30$ with no clear outliers OR $30 \le n < 100$ with no extreme outliers OR $n \ge 100$
Theory-based standardized statistic (test statistic)	$z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})}$ $SE_0(\hat{p}) = \sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}$	$z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{SE_0(\widehat{p_1} - \widehat{p_2})}$ $SE_0(\widehat{p_1} - \widehat{p_2}) = \sqrt{\widehat{p_{pool}} \times (1 - \widehat{p_{pool}}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	$t = \frac{\bar{x}_d - 0}{SE(\bar{x}_d)}$ $SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$	$t = \frac{b_1 - 0}{SE(b_1)}$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.	$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Theory-based confidence interval	$\hat{p} \pm z^* \times SE(\hat{p})$ $SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$	$\widehat{p_1} - \widehat{p_2} \pm z^* \times SE(\widehat{p_1} - \widehat{p_2})$ $SE(\widehat{p_1} - \widehat{p_2}) = \sqrt{\frac{\widehat{p_1} \times (1 - \widehat{p_1})}{n_1} + \frac{\widehat{p_2} \times (1 - \widehat{p_2})}{n_2}}$	$\bar{x}_d \pm t^* \times SE(\bar{x}_d)$ $SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$	$b_1 \pm t^* \times SE(b_1)$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.	$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

References

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

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.

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.

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).

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.

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

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

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

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

"Physician's Health Study." n.d. https://phs.bwh.harvard.edu/.

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

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

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

"Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status." 2021. CDC. https://covid.

cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination.

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

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

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

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

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.

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).