STAT 216 Coursepack

Summer 2022
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: Carnegie, N., Hancock, S., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R.* Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/.

Contents

Preface **1**

1 Basics of Data **2**

1.1 Module 1 Reading Guide: Basics of Data 2

1.2 Activity 1: Martian Alphabet 6

2 Study Design **13**

2.1 Module 2 Reading Guide: Sampling, Experimental Design, and Scope of Inference 13

2.2 Activity 2A: American Indian Address 17

2.3 Activity 2B: American Indian Address (continued) 22

2.4 Module 2 Lab: Study Design 27

3 Exploring Categorical and Quantitative Data **34**

3.1 Module 3 Reading Guide: Introduction to R, Categorical Variables, and a Single Quantitative Variable 34

3.2 Activity 3A: Graphing Categorical Variables 42

3.3 Activity 3B: IMDb Movie Reviews — Displaying Quantitative Variables 48

3.4 Module 3 Lab: IPEDs 55

4 Exploring Multivariable Data **62**

4.1 Module 4 Reading Guide: Two Quantitative Variables and Multivariable Concepts 62

4.2 Activity 4A: Movie Profits — Linear Regression 71

4.3 Activity 4B: Movie Profits — Correlation and Coefficient of Determination 75

4.4 Module 4 Lab: Penguins 81

5 Exam 1 Review **84**

6 Inference for a Single Categorical Variable: Simulation-based Methods **90**

6.1 Module 6 Reading Guide: Categorical Inference 90

6.2 Activity 6: Helperer-Hinderer — Simulation-based Hypothesis Test 97

6.3 Module 6 Lab: Helper-Hinderer (continued) 103

7 Inference for a Single Categorical Variable: Theory-based Methods + Errors and Power **108**

7.1 Module 7 Reading Guide: Categorical Inference 108

7.2 Activity 7A: Helper-Hinderer — Simulation-based Confidence Interval 118

7.3 Activity 7B: Handedness of Male Boxers — Theory-based Methods 124

7.4 Module 7 Lab: Errors and Power 130

8 Inference for Two Categorical Variables: Simulation-based Methods **135**

8.1 Module 8 Reading Guide: Hypothesis Testing for a Difference in Proportions 135

8.2 Activity 8A: The Good Samaritan — Simulation-based Hypothesis Test 142

8.3 Activity 8B: The Good Samaritan (continued) — Simulation-based Confidence Interval 148

8.4 Module 8 Lab: Fatal Injuries in the Iliad 154

9 Inference for Two Categorical Variables: Theory-based Methods **158**

9.1 Module 9 Reading Guide: Hypothesis Testing for a Difference in Proportions 158

9.2 Activity 9A: Winter Sports Helmet Use and Head Injuries — Theory-based Hypothesis Test . . . 162

9.3 Week 9B: Winter Sports Helmet Use and Head Injuries — Theory-based Confidence Interval . . 169

9.4 Module 9 Lab: Diabetes 174

10 Exam 2 Review **177**

11 Inference for a Quantitative Response with Paired Samples **182**

11.1 Module 11 Reading Guide: Inference for a Single Mean or Paired Mean Difference 182

11.2 Activity 11A: COVID-19 and Air Pollution 192

11.3 Activity 11B: Color Interference 199

11.4 Module 11 Lab: Swearing 206

12 Inference for a Quantitative Response with Independent Samples **211**

12.1 Module 12 Reading Guide: Inference for a Difference in Two Means 211

12.2 Activity 12: Weather Patterns and Record Snowfall 218

12.3 Module 12 Lab: The Triple Crown 225

13 Inference for Two Quantitative Variables **230**

13.1 Module 13 Reading Guide: Inference for Slope and Correlation 230

13.2 Activity 13A: Diving Penguins 237

13.3 Activity 13B: Golf Driving Distance 243

13.4 Module 13 Lab: COVID Immunization and Infection Rates 250

14 Probability and Relative Risk **255**

14.1 Module 14 Reading Guide: Special Topics 255

14.2 Activity 14A: What's the probability? 259

14.3 Activity 14B: Titanic Survivors — Relative Risk 263

14.4 Module 14 Lab: Efficacy of the COVID Vaccination 268

15 Semester Review **270**

15.1 Final Exam Review 270

15.2 Golden Ticket to Descriptive and Inferential Statistical Methods 277

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University, 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/. Videos assigned in the course calendar and 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, the coursepack includes reading guides to aid in taking notes while you complete the required readings and videos. Bring this workbook with you to class each class period, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

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

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

- *Prior to class meeting*:
  - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
  - Watch assigned videos on that week's content, pausing to take notes and answer video quiz questions.
  - Read through the introduction to the day's in-class activity
  - Read through the week's homework assignment and note any questions you may have on the content.
- *During class meeting*:
  - Work through the 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.
  - Finish watching any remaining assigned videos or readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Basics of Data

1.1 Module 1 Reading Guide: Basics of Data

Sections 1.1 (Case study) and 1.2 (Data basics)

Videos

- Stat 216 Course_Tour
- Instructor bio
- 1.2.1_1.2.2
- 1.2.3_1.2.4_1.2.5

Vocabulary

Data:

Summary statistic:

Case/Observational unit:

Variable:

Quantitative variable:

Discrete variables:

Examples of discrete variables using the `County` data:

Continuous variables:

Examples of continuous variables using the `County` data:

Example of a number which is NOT a numerical variable:

Categorical variable:

Ordinal variable:

Example of an ordinal variable using the `County` data:

Nominal variable:

Examples of nominal variables using the `County` data:

Note: Ordinal and nominal variables will be treated the same in this course. We recommend taking more statistics courses in the future to learn better methods of analysis for ordinal variables.

Data frame:

Scatterplot:

Each point represents:

Positive association:

Negative association:

Association or Dependent variables:

Independent variables:

Explanatory variable:

Response variable:

Observational study:

Experiment:

Placebo:

Notes

Big Idea: Variability is inevitable! We would not expect to get *exactly* 50 heads in 100 coin flips. The statistical question then is whether any differences found in data are due to random variability, or if something else is going on.

The larger the difference, the **less we believe the difference was due to chance.**

In a data frame, rows correspond to _____

and columns correspond to _____.

How many types of variables are discussed? Explain the differences between them and give an example of each.

True or False: A pair of variables can be both associated AND independent.
True or False: Given a pair of variables, one will always be the explanatory variable and one the response variable.
True or False: If a study does have an explanatory and a response variable, that means changes in the explanatory variable must **cause** changes in the response variable.

True or False: Observational studies can show a naturally occurring association between variables.

Example (Section 1.1 — Case study: Using stents to prevent strokes)

1. What is the principle question the researchers hope to answer? (We call this the **research question**.)

2. When creating two groups to compare, do the groups have to be the same size (same number of people in each)?

3. What are the cases or observational units in this study?

4. Is there a clear explanatory and response variable? If so, name the variable in each role and determine the type of variable (discrete, continuous, nominal, or ordinal).

5. What is the purpose of the control group?

6. Is this an example of an observational study or an experiment? How do you know?

7. Consider Tables 1.1 and 1.2. Which table is more helpful in answering the research question? Justify your answer.

8. Describe in words what is shown in Figure 1.1. Specifically, compare the proportion of patients who had a stroke between the treatment and control groups after 30 days as well as after 365 days.

9. Given the notion that the larger the difference between the two groups (for a given sample size), the less believable it is that the difference was due to chance, which measurement period (30 days or 365 days) provide stronger evidence that there is an association between stents and strokes, or that the differences are not due to random chance?

10. This study reported finding evidence that stents *increase* the risk of stroke. Does this conclusion apply to all patients and all stents?

11. This study reported finding evidence that stents *increase* the risk of stroke. This conclusion implies a causal link between stents and an increased risk of stroke. Is that conclusion valid? Justify your answer.

1.2 Activity 1: Martian Alphabet

1.2.1 Learning outcomes

- Describe the statistical investigation process.
- Identify observational units, variables, and variable types in a statistical study.

1.2.2 Terminology review

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

- Observational units or cases
- Variables: categorical or quantitative
- Proportions
- Graphs: frequency bar plot and relative frequency bar plot
- Distribution

For more on these concepts, read Sections 1.2 and 2.1 in the textbook.

1.2.3 General information labs

For each module 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. Today's activity is Lab 0 in Gradescope for practice submitting.

1.2.4 Can you read "Martian?"

How well can humans distinguish one "Martian" letter from another? In today's activity, we'll find out. When shown the two Martian letters, Kiki and Bumba, write down whether you think Bumba is on the left or on the right.

1. Were you correct or incorrect in identifying Bumba?

Steps of the statistical investigation process

Step 1: The first step of any statistical investigation is to *ask a research question.* In this study the research question is: Can we as a class read Martian? (We will refine this later on!).

Step 2: To answer any research question, we must *design a study and collect data.* For our question, the study consists of each student being presented with two Martian letters and asking which was Bumba. 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.

2. What are the observational units in this study?

3. How many students are in class today? This is the **sample size**.

A **variable** is information collected or measured on each observational unit or case. Each column in a data set will represent a different variable. Today we are only measuring one variable on each observational unit.

4. **Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?** *Hint*: Your answer to question 1 is the outcome for the variable measured on one observational unit.

We will look at two types of variables: **quantitative** and **categorical** (see Figure 1.1).

Quantitative variables are numerical measurements that can be discrete (whole, non-negative numbers) or continuous (any value within an interval). The number of pets one owns would be a discrete variable as you can not have a partial pet. GPA would be a continuous variable ranging from 0 to 4.0.

The outcome of a categorical variable is a group or category such as eye color, state of residency, or whether or not a student lives on campus. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered nominal variables. All categorical variables will be treated as nominal for analysis in this course.

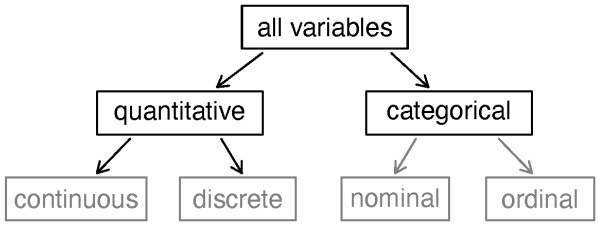


Figure 1.1: Types of variables.

5. Is the variable identified in question 4 categorical or quantitative?

Step 3: Once we have collected data, the next step is to *summarize and visualize the data.*

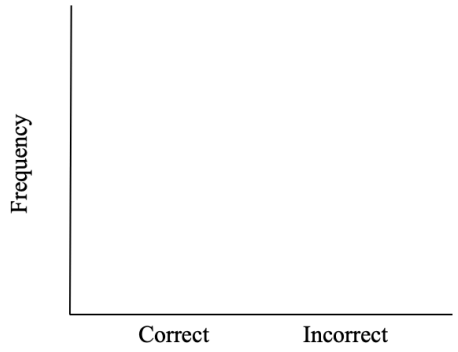
6. How many people in your class were correct in identifying Bumba? Using the class size from question 3, calculate the proportion of students who correctly identified Bumba.

$$\text{proportion} = \frac{\text{number of students who correctly identified Bumba}}{\text{total number of students}}$$

The proportion in question 6 is called a **summary statistic**—a single value that summarizes the data set. It is important to note that a variable is different than a summary statistic. A *variable* is measured on a *single observational unit* while a summary statistic is calculated from a group of observational units. For example, the variable "whether or not a student lives on campus" can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who live on campus, the summary statistic. Look back and make sure you wrote the variable in question 4 as a variable, NOT a summary statistic.

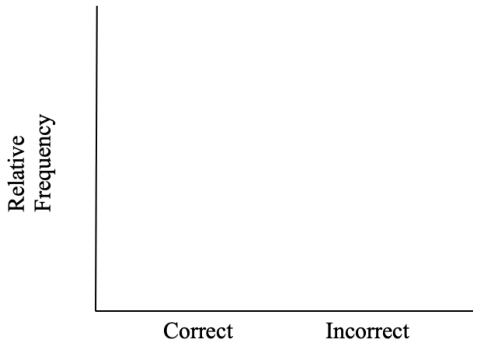
Looking at the data set and the summary statistic is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels or outcomes, correct or incorrect, we will create two bars—one for each level.

7. Plot the observed class data using a frequency bar plot. Be sure to add a scale to the y-axis.



We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot. Be sure to add a scale to the y-axis.



Step 4: The next step is to *use statistical analysis methods to draw inferences from the data.* To answer the research question, we will simulate what *could* have happened in our class given random chance, repeat many times to understand the expected *variability* between different "randomly guessing" classes, then compare our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) the class's result would occur if students were all merely guessing, allowing us to determine if the data provides evidence that we as a class can in fact read Martian.

9. If humans really don't know Martian and are just guessing which is Bumba, what are the chances of getting it right?

   How could we use a coin to simulate each student "just guessing" which Martian letter is Bumba?

   How could we use coins to simulate the entire class "just guessing" which Martian letter is Bumba?

   How many people in your class would you expect to choose Bumba correctly just by chance? Explain your reasoning.

10. Each student will flip a coin one time to simulate your "guess" under the assumption that we can't read Martian. Let Heads = correct, Tails = incorrect. What was the result of your one simulation?

    What was the result from your class's simulation? What proportion of students "guessed" correctly in the simulation?

11. If students really don't know Martian and are just guessing which is Bumba, which seems more unusual: the result from your class's **simulation** or the observed proportion of students in your class that were correct (this is your summary statistic from question 6)? Explain your reasoning.

12. While your observed class data is likely far different from the simulated "just-guessing" class, comparing our class data to a single simulation does not provide enough information. The differences seen could just be due to the randomness of that set of coin flips! Let's simulate another class. Each student should flip their coin again. What was the result from your class's second simulation? What proportion of students "guessed" correctly in the second simulation? Create a plot to compare the two simulated results with the observed class result.

13. **We still only have a couple of simulations to compare our class data to. It would be much better to be able to see how our class compared to hundreds or thousands of "just-guessing" classes. Since we don't want to flip coins all class period, your instructor will use a computer simulation to get 1000 trials. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 trials (repetitions).**

    Probability of correct guesses: _____

    Sample size: _____

    Number of repetitions: _____

14. Sketch the distribution displayed by your instructor here. Label each axis appropriately.

What does one dot on the plot above represent in context of the problem?

15. Is your class particularly good or bad at Martian? Use the plot in question 14 to explain your answer.

16. Is it *possible* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

17. **Is it *likely* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.**

Step 5: The next step in the statistical investigation process is to *communicate the results and answer the research question.*

18. Does this activity provide strong evidence that students were not just guessing at random? If so, what do you think is going on here? Can we as a class read Martian?[1]

Introduction to R

In Stat 216 we will use the statistical package R to analyze data through the IDE (integrated development environment) RStudio. Though it is possible to download R and RStudio on your own computer, we will use this program through the MSU RStudio server: https://rstudio.math.montana.edu/.

Read through the preliminaries chapter in the textbook and watch the video "Starting with R" before completing the following questions.

The RStudio workflow operates best by the use of "Projects." You should create a separate project for each activity or assignment in this course that requires the use of R. To get started with this activity, follow these steps:

- Log onto the RStudio server using your NetID and password: https://rstudio.math.montana.edu/.
  - Please note: Your netID password expires every 6 months. It is HIGHLY recommended that you reset your netID password BEFORE attempting to login to the Rstudio server. You can reset your netID password in the MSU password portal (https://pwreset.montana.edu/react/).
- In the top right corner, you will see a dropdown menu next to "Project" that currently says "(None)." Click on this menu and choose "New Project." (Alternatively, you can click the "File" menu in the top left and select "New Project.")
  - A "New Project Wizard" window should pop up: click "New Directory," then click "New Project."
  - Give your project directory a name (e.g., Activity1). *Do not use spaces or other characters in the name.*
  - Click "Browse" and choose a location where you would like to save your project (you can create a new folder if desired). Note that this location is on your server account, not on your computer.
  - Leave all other boxes unchecked, and click "Create Project." (Now, if you click on the home icon in the top right, you will see your RStudio account, and the project should be listed under "Projects.")
- Download the Martian Alphabet R script file from D2L.
- Click "Upload" in the "Files" tab in the bottom right window of RStudio. Click "Choose File," and navigate to the folder where the Martian Alphabet R script file is saved. Then click "Open"; then click "Ok."
- You should see the uploaded file appear in the list of files. Click on the filename to open the file.

[1] Reference for "Martian alphabet" is a TED talk given by Vilayanur Ramachandran in 2007. The synesthesia part begins at roughly 17:30 minutes: `http://www.ted.com/talks/vilayanur_ramachandran_on_your_mind`.

In the Martian Alphabet `R` script file, highlight the lines of code that starts with `library` and click "Run." This will load the **package** (or library) `catstats` needed for this activity; each package is a collection of `R` functions. We review a few of these packages here.

- Throughout the semester we will use the package `tidyverse` to allow us to use chaining (see Section 1.7 in the textbook for more on this symbol `%>%`.) Contained in `tidyverse` is the package `ggplot2`, used to create graphs in RStudio.
- The package `mosaic` contains the `favstats()` function to find summary statistics for quantitative variables.
- We will use the package `catstats`, starting in Chapter 5 (and in this activity), to create simulations for statistical inference.

These packages are already installed in the RStudio server, but you need to use the `library()` function to call the package into your R environment. We will only use the package `catstats` for this activity.

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.

In the Martian Alphabet `R` script file for the `one_proportion_test()` function arguments, enter your class size (Q3 from the in-class activity) for `sample_size` and the number of students who were correct in identifying Bumba (Q6 from the in-class activity) for `as_extreme_as` argument. Highlight lines 3 – 8 and click run.

Is the distribution created from this code similar to what you saw in class in Q14?

1.2.5 Take-home messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses" when asked to identify Bumba) to a distribution of many simulated results under an assumption like "blind guessing."
2. Blind guessing between two outcomes will be correct only about half the time. We can simulate data using a computer program to fit the assumption of blind guessing.
3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

1.2.6 Additional notes

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

MODULE 2

Study Design

2.1 Module 2 Reading Guide: Sampling, Experimental Design, and Scope of Inference

Section 1.3 (Sampling principles and strategies)

Videos

- 1.3

Vocabulary

(Target) Population:

Sample:

Anecdotal evidence:

Bias:

Selection bias:

Non-response bias:

Response bias:

Convenience sample:

Simple Random Sample:

Non-response rate:

Representative:

Notes

Ideally, how should we sample cases from our target population? Using what sampling method?

Notes on types of sampling bias

- Someone must first be *chosen* to be in a study and refuse to participate in order to have **non-response bias**.
- There must be a valid reason for someone to lie or be untruthful to justify saying **response bias** is present. Yes, anyone could lie at any time to any question. Response bias is when those lies are predictable and systematic based on outside influences.

True or False: Convenience sampling tends to result in non-response bias.

True or False: Volunteer sampling tends to result in response bias.

True or False: Random sampling helps to resolve selection bias, but has no impact on non-response or response bias.

Sections 1.4 (Observational studies), 1.5 (Experiments), and 1.6 (Scope of inference)

Videos

- 1.4to1.6

Reminders from Section 1.2

Explanatory variable: The variable researchers think *may be* affecting the other variable. What the researchers control/assign in an experiment. If comparing groups, the explanatory variable puts the observational units into groups.

Response variable: The variable researchers think *may be* influenced by the other variable. This variable is always observed, never controlled or assigned.

Vocabulary

Observational study:

Observational data:

Prospective study:

Retrospective study:

Confounding variable:

Experiment:

Randomized experiment:

Blocking:

Treatment group:

Control group:

Placebo:

Placebo effect:

Blinding:

Scope of inference:

Generalizability:

Causation:

Notes

What are the four principles of a well-designed randomized experiment?

Fill in the appropriate scope of inference for each study design.

	Study Type	
Selection of Cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)		
Non-random sample (or other sampling bias)		

True or False: Observational studies can show an association between two variables, but cannot determine a causal relationship.

True or False: In order for an experiment to be valid, a placebo must be used.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be generalized to the entire target population.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be inferred as a causal relationship between the explanatory and response variables.

2.2 Activity 2A: American Indian Address

2.2.1 Learning outcomes

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

2.2.2 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
- Types of samping bias
- Generalization

To review these concepts, see Section 1.3 in the textbook.

Types of sampling bias.

In today's activity, we will look at sampling and types of bias (selection, non-response, or response).

In these next questions, identify the target population, the sample selected, the variable, and the type of bias present.

1. 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.

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

2. A television station is interested 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.

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

3. 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?"

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

4. The Bozeman school district is 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.

2.2.3 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

5. 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.

6. 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		

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

8. Report your mean word length to your instructor. Your instructor will guide the class in creating a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive x-axis label.

9. Based on the plot of sample mean word lengths in question 8, what is your best guess for the average word length of the population of all 359 words in the speech?

10. 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 our plot above?

11. If your samples were truly representative, what proportion of sample means would you expect to be below 3.95?

12. What proportion of students' computed sample means were lower than the true mean of 3.95 letters?

13. Based on your answers to questions 11 and 12, would you say the sampling method used by the class is biased or unbiased? Justify your answer.

14. If the sampling method is biased, what type of bias is present? What is the direction of the bias, i.e., does the method tend to overestimate or underestimate the population mean word length?

15. 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?

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 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 2B: 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 1.3 in the textbook.

Random selection

Today we will return to the American Indian Address introduced in Activity 2A. First, refresh your memory where the activity finished.

1. Explain how you determined the sampling method used by the class to select words from the American Indian address resulted in selection bias. What did each student need to do? What did you plot from each student? What did you compare to that plot and how did you use the plot to determine bias was present in the sampling method?

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.

2. 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?"

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

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

4. Report your mean word length to your instructor. Your instructor will guide the class in creating a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive x-axis label.

5. Where does the value 3.95, the true mean word length, fall in the distribution created in question 4?

6. How does the plot generated in question 4 compare to the plot generated in question 8 from Activity 2A?

   Which features are similar?

   Which features differ?

   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 Sample(s)."

7. The plot labeled "Statistic" 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.

8. What is the center value of the distribution created in question 7?

9. 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."

10. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from question 7.

Effect of sample size

We will now consider the impact of sample size.

11. First, consider if each student had selected 20 words, instead of 10, by eye. Do you think this would make the plot from question 8 in Activity 2A centered on 3.95 (the true mean word length)? Explain your answer.

12. 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 Sample(s)."

The plot labeled "Statistic" 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.

13. Compare the distribution created in question 12 to the one created in question 7.

    Which features are similar?

    Which features differ?

14. Compare the spreads of the plots in question 12 and in question 7. You should see that in one plot all sample means are closer to the population mean than in the other. Which plot shows this?

15. 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 of sample estimates? Explain your answer

16. 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 our true parameter.
3. Random selection eliminates selection bias. Random selection will not eliminate response or non-response bias however.
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 1.2 through 1.6 in the textbook.

2.4.3 General information labs

Remember that for each module 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 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 explanatory variable, the response variable, and the study design (observational study or experiment).

1. 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 last fall. 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.

   Explanatory variable:

   Response variable:

   Study design:

2. **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.**

   Explanatory variable:

   Response variable:

   Study design:

2.4.5 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.

3. 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:

Males: Paul, Antonio, Davieon, Chao, Aryan, Jabari, Tong, Andres, John, Liu, Lucas, Rashidi, Shiwoo, Jihoon, Alejandro, Daniel

Non-males: An, Nailah, Jasmine, Ka Nong, Keyaina, Mary, Adah, Sassandra

4. Is this a simple random sample or a convenience sample? How do you know?

5. Based on the sampling method, to what population should the results of this study be generalized?

6. One way to separate into two groups would be give all the males the placebo and all the non-males the new drug. Would this be a reasonable strategy? Explain your answer.

7. Could the scientists fix the problem with the strategy presented in question 6 by creating equal sized groups by putting 4 males and 8 non-males into the drug group and the remaining 12 males in the placebo group? Explain your answer.

8. A third strategy would be to **block** on sex. In this type of study, the scientists would assign 4 non-males and 8 males to each group.

Using this strategy, how many males are in each group?

What is the sample size of each group?

Is the proportion of males the same in the drug and placebo groups?

9. **Assume the scientists used the strategy in question 8, but they put the four tallest non-males and eight tallest males into the placebo group and the remaining subjects into the control 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 9 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.

10. 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 7.
- 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)?

11. 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 under "Simulation" and click Randomize again. Did you get the same difference in proportion of males between the placebo and drug groups?

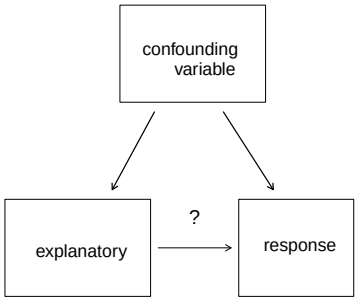
12. Change Repetitions under "Simulation" to 998 (for 1000 total). Sketch the plot of the distribution of difference in proportions from each of the 1000 random assignments here. Be sure to include a descriptive x-axis label.

13. **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 12 to justify your answers.**

14. 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.

15. Suppose there is a genetic component to how well permanent atrial fibrillation responds to medication. The scientists do not know about this gene ahead of time, but if you select Reveal gene? under "Choose variables" then change the drop-down menu under Group 2 from "height" to "gene," we can see how random assignment impacts the distribution of this gene between the two groups. Explain what happens to the gene variable, in the long run, if random assignment is used to create the two groups. Use the dot plot to justify your answer.

The diagram below summarizes these ideas about confounding variables and random assignment. When a confounding variable is present (such as sex, height, or a gene), 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.



16. **What is the purpose of random assignment of the subjects in a study to the explanatory variable groups?**

17. 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.

18. **Both the sampling method (which we covered earlier this week) 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 the table below to determine the scope of inference of this trial study described in question 17.**

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

19. Use the table to determine the scope of inference for the study in question 1.

20. Use the table to determine the scope of inference for the study in question 2.

2.4.6 Take-home messages

1. The study design 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 Module 3 Reading Guide: Introduction to R, Categorical Variables, and a Single Quantitative Variable

Section 1.7 (Data in R)

Videos

- Starting_with_R

Notes

R is case sensitive, meaning it reads `data` differently from `Data`. If you get an error message, check that your capitalization is correct.

R does not like spaces or special characters This means the column and row headers in the data set should not have spaces, periods, commas, etc. Instead of titling the variable `column header`, use `column_header` or `ColumnHeader`.

Tidy data: Data frames with

1 row per _____,

1 column per _____.

We highly recommend completing Tutorial 1 at the end of Chapter 1 (all four lessons) to give you practice with R/RStudio AND to help reflect on the content of Chapter 1: basics of data, sampling, study design, and scope of inference. These tutorials have some content questions and some places for you to practice using R online with some guidance.

___ indicate spots you need to type in functions, data sets, or variable names.

There are Hint and Solution buttons on the R code box to help you.

We would not expect you to know the coding right now, especially for things like mutations or creating new variables in the data set. But seeing some initial coding for these more difficult functions will only make you more comfortable using the functions needed for this course!

Functions

State what these introductory functions do in R:

`glimpse(data_set_name)`

`head(data_set_name)`

`data_set_name$variable_name`

`%>%`

`<-`

Section 2.1 (Exploring categorical data)

Videos

- 2.1
- MosaicPlots

Vocabulary

Frequency table:

Relative frequency table:

Contingency or two-way table:

Unconditional proportion:

Conditional proportion:

Row proportions:

Column proportions:

Statistic:

Sample proportion:

Notation:

Parameter:

Population proportion:

Notation:

Bar plot:

Segmented bar plot:

Simpson's Paradox:

Notes

In a contingency table, which variable (explanatory or response) generally will make the columns of the table? Which variable will make the rows of the table?

In a segmented bar plot, the bars represent the levels of which variable? The segments represent the levels of which variable?

What type of plot(s) are appropriate to display a single categorical variable?

What type of plot(s) are appropriate to display two categorical variables?

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

True or false: Pie charts are generally highly recommended ways to graphically display categorical data.

True or false: Two categorical variables are associated if the conditional proportions of a particular outcome (typically of the response variable) differ across levels of the other variable (typically the explanatory variable).

True or false: When a segmented bar plot has segments that sum to 1 (or 100%), the segment heights correspond to the proportions conditioned on the **segment**.

Review of Simpson's Paradox

Based on the segmented bar plot in Figure 2.6, which race of defendant was more likely to have the death penalty invoked?

Based on the segmented bar plot in Figure 2.7 and Table 2.9, which race of defendant was more likely to have the death penalty invoked when the victim was Caucasian?

Based on the segmented bar plot in Figure 2.7 and Table 2.9, which race of defendant was more likely to have the death penalty invoked when the victim was African American?

The direction of the relationship between the _____ and _____ variables is **reversed** when accounting for a _____ variable.

Section 2.3 (Exploring quantitative data)

Videos

- 2.3

Type of Plots

Scatterplot:

Dot plot:

Histogram:

Density plot:

Box plot:

Vocabulary

Four characteristics of a scatterplot:

Form:

Strength:

Direction:

Unusual observations or outliers:

Data density:

Tail:

Skew:

Symmetric:

Modality:

Distribution (of a variable):

Four characteristics of the distribution of one quantitative variable:

Center:

Variability:

Shape:

Outliers:

Point estimate:

Deviation:

Five number summary:

X^{th} percentile:

Interquartile range (IQR):

Robust statistics:

Notes

What type of plot(s) are appropriate for displaying one quantitative variable?

What type of plot(s) are appropriate for displaying two quantitative variables?

What type of plot(s) are appropriate for displaying one quantitative variable and one categorical variable?

What are the two ways to measure the 'center' of a distribution? Which one is considered robust to skew/outliers?

What are the three ways to measure the 'variability' of a distribution? Which one is considered robust to skew/outliers?

How are variance and standard deviation related?

Fill in the following table with the appropriate notation.

Summary Measure	Parameter	Statistic
Mean		
Variance		
Standard deviation		

How are outliers denoted on a box plot? How can you mathematically determine if a data set has outliers?

Section 2.4 (R: Exploratory data analysis) and Section 2.5 (Chapter 2 review)

Section 2.4 presents four tutorials on analyzing quantitative data in R. We recommend you complete all four.

Notes

Statistics summarize _____ .
Parameters summarize _____.

Fill in the following table with the appropriate notation for each summary measure.

Summary measure	Statistic	Parameter
Sample size		
Proportion (used to summarize one categorical variable)		
Mean (used to summarize one quantitative variable)		
Correlation (used to summarize two quantitative variables)		
Regression line slope (used to summarize two quantitative variables)		

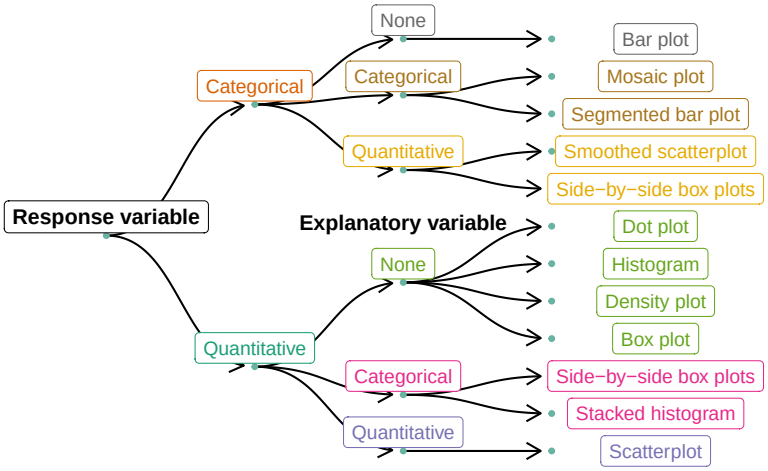
Look at the table of vocabulary terms. If there are any you do not know, be sure to review the appropriate section of your text.

Data visualization summary

Fill in the following table to help associate type of plot for each of several scenarios.

	Appropriate plot(s)
One categorical variable (categorical response, no explanatory)	
One quantitative variable (quantitative response, no explanatory)	
Two categorical variables (categorical response, categorical explanatory)	
One of each (quantitative response, categorical explanatory)	
Two quantitative variables (quantitative response, quantitative explanatory)	

Decision tree for determining an appropriate plot given a number of variables and their types from Chapter 2 review:



3.2 Activity 3A: Graphing 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 Sections 2.1 and 2.2 in the textbook.

3.2.3 Graphing categorical variables

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 2 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. These lines of code read in the data set and name the data set `myopia`. Highlight and run lines 1–6 in the R script file to load the needed packages and the data.

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

Displaying a single categorical variable

If we wanted to know how many children in our data set were in each level of myopia, we would create a frequency bar plot of the variable `Sight`. Enter the variable name, `Sight`, for `variable` into the `ggplot` code at line 10 in the R script file. Highlight and run lines 9–15 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).

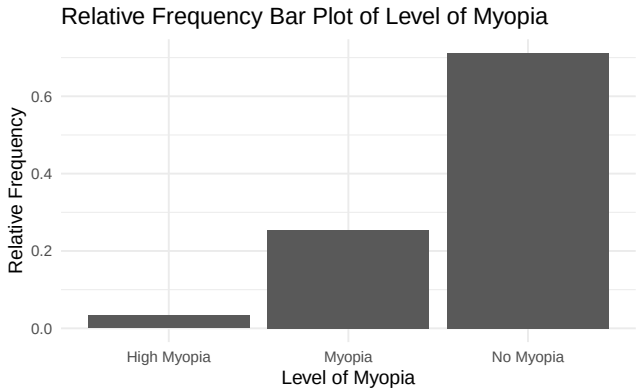
```
myopia %>% # Data set piped into...
ggplot(aes(y = variable)) +   # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Frequency",   # Label the x axis
       y = "Level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

2. Sketch the bar chart created below. Be sure to label the axes.

3. 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, divide the count in each level of myopia by the sample size. These are sample proportions. 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 = ..prop.., group = 1)) +  # Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



4. 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? To examine the differences in level of myopia for the level of light, we would create a segmented bar plot of `Light` segmented by `Sight`. To create the segmented bar plot enter the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in the `R` script file in line 27. Highlight and run lines 26–33.

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

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

6. From the segmented bar plot, estimate the proportion of no myopia for those that used a nightlight.

7. Which level of light has the highest proportion of `No Myopia`?

We could also plot the data using a mosaic plot. Fill in the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in line 38 in the `R` script file. Highlight and run lines 36–43.

```
myopia %>% # Data set piped into...
  ggplot() +   # This specifies the variables
  geom_mosaic(aes(x=product(explanatory), fill = response)) +  # 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()  # Make figure black and white
```

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

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

Fill in the name of the explanatory variable and the response variable in line 46 in the `R` script file, highlight and run line 46 to get the counts for each combination of levels of variables.

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

10. 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				

11. Calculate the proportion of children with high myopia. Use appropriate notation.

12. Calculate the proportion of children that slept with full light that have high myopia. Use appropriate notation.

13. Calculate the proportion of children that slept with no light that have high myopia. Use appropriate notation.

14. Calculate the difference in proportion of children with high myopia for those that slept with full light minus those who slept with no light. Give the appropriate notation. Label group 1 as full light and group 2 as no light.

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.

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 3B: 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 Section 2.3 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. Here we have the summary statistics for the variable `imdb_score`. The summary statistics give the two measures of center and two measures of spread for IMDb score. Highlight and run lines 1 – 8 in the provided `R` script file to load the data set. Check that the summary statistics match that printed 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. Give the values for the two measures of center (mean and median).

2. Calculate the interquartile range (IQR = Q3 - Q1).

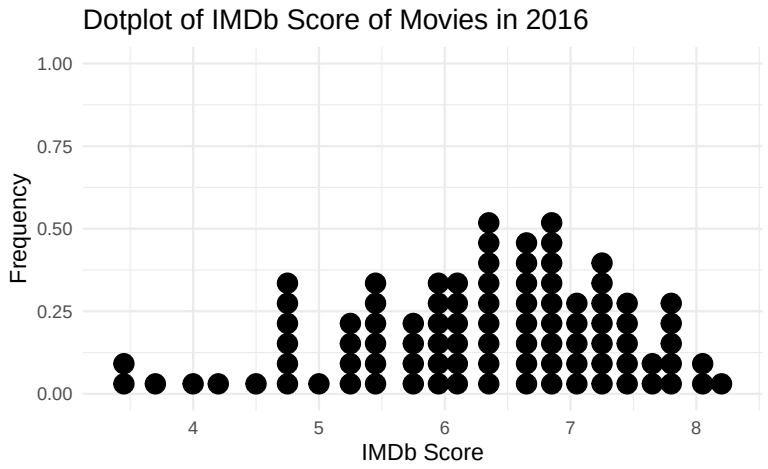
3. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

4. What are the three types of plots used to plot a single quantitative variable?

A dotplot will plot a dot for each value in the data set. The following code will create a dotplot of IMDb scores. Notice that we put in the variable name `imdb_score` for `x =` in the ggplot function.

```
movies %>% # Data set piped into...
ggplot(aes(x = imdb_score)) +   # Name variable to plot
  geom_dotplot() +  # Create dotplot
  labs(title = "Dotplot of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```



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

To create a histogram of the IMDb scores, enter the variable name, `imdb_score` in the provided `R` script file for `variable` at line 20, highlight and run lines 19–24. Visually, this shows us the range of IMDb scores for Movies released in 2016.

Notice that the **bin width** 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 = "Histogram of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

6. Sketch the histogram created here.

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

8. Which five summary statistics are used in creating a box plot? *Hint*: Together they are called the **five-number summary** of the variable.

9. Using the code below we see that the three smallest IMDb scores in the data set are 3.4, 3.5, and 3.7 and the three largest IMDb scores are 8.0, 8.1, and 8.2:

```
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_min(imdb_score, n = 3)  # Show 3 smallest values
```

```
#>   imdb_score
#> 1        3.4
#> 2        3.5
#> 3        3.7
```

```
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_max(imdb_score, n = 3)  # Show 3 largest values
```

```
#>   imdb_score
#> 1        8.2
#> 2        8.1
#> 3        8.0
```

Using the summary statistics given in the `R` output before question 1, and the smallest and largest values of the variable to check for outliers, sketch a box plot of IMDb Score. Be sure to label the axes.

10. Compare the three graphs of IMDb scores created above.

Which graph is best used to show the shape of the distribution?

Which graph is best used to show the outliers of the distribution?

Summary statistics for a single categorical and single quantitative Variable

Is there an association between content rating and budget for movies 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 37–39 in the provided `R` script file and check that the summary statistics match those provided in the coursepack.

```
movies %>% # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  summarise(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
```

11. Which content rating has the largest IQR?

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

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

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

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 44, highlight and run code lines 42–48. 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",  # Title
       x = "Content Rating",    # x-axis label
       y = "Budget (in Millions)")  # y-axis label
```

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

16. 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 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 third quartile (Q3) for the PG-13 rating? Interpret this value in context.

17. Which variable is the explanatory variable? Response variable?

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 $Q_1 - 1.5 * IQR$ and greater than $Q_3 + 1.5 * IQR$ 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 data.
- 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).
- Use `R` to create graphs of variables.

3.4.2 The Integrated Postsecondary Education Data System (IPEDS)

Download and open the provided `R` script file for the week 3 lab to answer the following questions. **Remember that 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
Control	• Public • Private
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 2018-19. It includes in-out-of-state tuition and fees, books and supplies, on campus room and board, and other on 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 2018-19. It includes in-state tuition and fees, books and supplies, on campus room and board, and other on 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.

Summarizing a single quantitative variable

1. What are the observational units for this study?

2. Identify in the chart above which variables are categorical (C) and which variables are quantitative (Q).

Upload the data set `IPEDS_Data_2018` to the `R` Studio server. Click on Import Dataset in the Environment tab in the upper right hand corner. Choose `From Text(base)` and select the correct csv file. Be sure that `Yes` is selected next to `Heading` in the pop-up screen. Click `Import`.

Enter the name of the data set (see the environment tab) for `datasetname` in the `R` script file in line 6. We will look at the retention rates for the 4-year institutions. Enter the variable name `Retention` for `variable` in line 12. Highlight and run lines 1 – 12. Note that the two lines of code (lines 8 and 10) are filtering to remove the 2-year institutions so we are only assessing Public 4-year and Private 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
```

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

4. Calculate the interquartile range for this study.

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

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

Next we will create both a histogram and a boxplot of the variable `Retention`. Enter the name of the variable in both line 16 and line 23 for `variable` in the `R` script file. **Give each plot a descriptive title.** Highlight and run lines 15 – 27 to give the histogram and boxplot. **Export and upload both plots to Gradescope for your group.** To export the graphs: in the bottom right corner in the Plots tab, click on `Export`, then 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...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 5) +  # Create histogram with specified binwidth
  labs(title = "Title", # Title for plot
       x = "Rentention Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Title", # Title for plot
       x = "Retention Rates", # Label for x axis
       y = "Frequency") # Label for y axis
```

7. What is the shape of the distribution of retention rates?

8. Identify any outliers in the data set.

Robust Statistics

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

9. Report the two measures of center for retention given in the R output.

10. Report the two measures of spread for retention given in 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 30–38.

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

```
IPEDS %>% # Data set piped into...
  ggplot(aes(x = Retention_Inc)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of Adjusted Revenue of Movies in 2016", # Title for plot
       x = "Revenue (in Millions)", # Label for x axis
       y = "Frequency") # Label for y axis
```

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

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

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

14. 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 `Control` in the data set.

15. **Which variable will we treat as the explanatory variable? Response variable?**

Enter the name of the explanatory variable and the name of the response variable in lines 42 and 45 of the R script file. Highlight and run lines 41 – 49 to find the summary statistics and create side by side boxplots of the data.

```
IPEDS %>%  # Data set piped into...
  summarise(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 = "Side by side box plot of retention rates by control",  # Title
       x = "Control",    # x-axis label
       y = "Retention Rates")  # y-axis label
```

16. **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 outliers?

17. 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.

18. Does there appear to be an association between retention rates and type of university? Explain your answer.

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 53. Highlight and run lines 52 – 58 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", # Title
       x = "Sector", # x-axis label
       y = "") + # remove y-axis label
  scale_fill_grey()
```

19. **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 Module 4 Reading Guide: Two Quantitative Variables and Multivariable Concepts

Section 3.1 (Fitting a line, residuals, and correlation)

Videos

- Chapter3

Reminders from Section 2.3

Scatterplot: displays two quantitative variables; one dot = two measurements (x, y) on one observational unit.

Four characteristics of a scatterplot:

- *Form*: pattern of the dots plotted. Is the trend generally linear (you can fit a straight line to the data) or non-linear?
- *Strength*: how closely do the points follow a trend? Very closely (strong)? No pattern (weak)?
- *Direction*: as the x values increase, do the y-values tend to increase (positive) or decrease (negative)?
- Unusual observations or *outliers*: points that do not fit the overall pattern of the data.

Vocabulary

Residual:

Formula:

Residual plot:

Correlation:

Notes

General equation of a linear model for a *population*: $y = \beta_0 + \beta_1 x + \epsilon$, where

x represents

y represents

β_0 represents

β_1 represents

ϵ represents

General equation of a linear regression model from *sample* data: $\hat{y} = b_0 + b_1 x$, where

x represents

$\hat{y}$ represents

b_0 represents

b_1 represents

Fill in the following table with the appropriate notation for each summary measure.

Summary Measure	Parameter	Statistic
Correlation		
Slope		
y-intercept		

Fill in the blanks below to define some of the properties of correlation:

The value of correlation must be between _____. (Includes the endpoints of the interval)

The sign of correlation gives the _____ of the linear relationship.

The magnitude of correlation gives the _____ of the linear relationship.

True or false: A scatterplot that shows random scatter would be considered non-linear.

True or false: If the correlation between two quantitative variables is equal to zero, then the two variables are not associated.

True or false: To calculate a predicted y-value from a given x-value, just look at the scatterplot and estimate the y-value.

True or false: A positive residual indicates the data point is above the regression line.

Example: Brushtail possums

1. What are the observational units?

2. Look at the scatterplot in Figure 3.5.

   a) What is the explanatory variable? The response variable? What type is each?

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

3. Write the equation of the regression line, in context (do not use x and y, use variable names instead).

4. Calculate the predicted head length for a possum with a 76.0 cm total length.

5. One of the possums in the data set has a total length of 76.0 cm and a head length of 85.1 mm. Calculate the residual for this possum. Does this possum lie above or below the regression line?

Section 3.2 (Least squares regression)

You may skip the special topic Sections 3.2.3.1 and 3.2.6.

Videos

- Chapter3

Vocabulary

Least squares criterion:

Least squares line:

`lm()` R function: `name_of_model <- lm(response ~ explanatory, data = data_set_name)`

slope:

y-intercept:

Extrapolation:

Coefficient of determination:

s_y^2 (or SST) represents

s_{RES}^2 (or SSE) represents

Notes

Two methods for determining the best line:

1.

2.

Notation for the coefficient of determination:

Formulas for calculating the coefficient of determination:

True or false: A correlation between two quantitative variables implies a causal relationship exists between the variables.

True or false: The slope of the line tells us how much to expect the y variable to increase or decrease when the x variable increases by 1 unit.

True or false: The coefficient of determination is just the square of the correlation.

Example: Elmhurst College

1. What are the observational units?

2. Look at the scatterplot in Figure 3.13.

   a) What is the explanatory variable? The response variable?

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

3. Write the equation of the regression line, in context (do not use x and y, use variable names instead).

4. Interpret the slope of the line, in the context of the problem. Remember that both family income and gift aid from the university are measured in \$1000s.

5. Interpret the y-intercept of the line, in the context of the problem. Remember that both family income and gift aid from the university are measured in \$1000s.

6. Is your interpretation in question 5 an example of extrapolation?

7. Give and interpret, in context, the value of the coefficient of determination.

Section 3.3 (Outliers in linear regression)

Videos

- Chapter3

Vocabulary

Outlier:

Leverage:

Influential:

Notes

Investigate, but do not remove, outliers. Unless you find there was an actual error in the data collection, ignoring outliers can make models poor predictors!

True or false: All high leverage outliers are influential.

True or false: An outlier is considered high leverage if it is extreme in its x-value.

Section 3.4 (R: Correlation and regression) and Section 3.5 (Chapter 3 review)

Videos

- Chapter3

Section 3.4 presents five tutorials on analyzing two quantitative variables in `R`. We recommend you complete all five.

Notes

Statistics summarize:

Parameters summarize:

What are the two ways to calculate the coefficient of determination?

What is the formula for calculating a residual?

Determine whether each of the following statements about the correlation coefficient are true or false:

1. The correlation coefficient must be a positive number.
2. Stronger linear relationships are indicated by correlation coefficients far from 0.
3. The correlation coefficient is a robust statistic.

4. When two variables are highly correlated, that indicates a causal relationship exists between the variables.
5. The sign of the correlation coefficient will be the same as the sign of the regression line slope, though the values are typically different.

Fill in the blanks to correctly interpret:

- Slope:

  For every _____, we expect _____ to increase (if slope is _____) or decrease (if slope is _____) by the absolute value of the _____.

- y-intercept:

  If _____, we predict the _____ to equal _____.

Look at the table of vocabulary terms. If there are any you do not know, be sure to review the appropriate section of your text.

Section 4.1 (Gapminder world)

Videos

- Chapter4

Reminder from Section 3.1

Use color and a legend to add a third variable to a scatterplot. E.g., Color the dots to represent different levels of a categorical variable or use shading of the dots to represent different values of a quantitative variable.

Vocabulary

Interaction:

Aesthetic:

Notes

If the response and one predictor are quantitative and the other predictor is categorical, we fit a regression line for each level of the categorical predictor.

- Parallel slopes would indicate that that the two predictors _____ in explaining the response.
- Non-parallel slopes would indicate that the two predictors _____ in explaining the response.

True or false: Scatterplots can only display two variables at a time.

Section 4.2 (Simpson's Paradox, revisited)

Videos

- Chapter4

Reminder from Section 2.1

Simpson's Paradox: when the relationship between the explanatory and response variable is reversed when looking at the relationship within different levels of a confounding variable.

Notes

True or false: Simpson's Paradox can only occur when the explanatory, response, and confounding variables are all categorical.

Example: SAT scores

1. What are the observational units?

2. Look at the scatterplot in Figure 4.5.

   a) What is the explanatory variable? The response variable?

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

3. What would need to be done to the study design in order to eliminate the confounding variable: percent of eligible students taking the SAT?

4. What features of the scatterplots in Figure 4.6 demonstrate that the percent of eligible students taking the SAT is a confounding variable?

5. How does Figure 4.7 demonstrate Simpson's Paradox?

Section 4.4 (Chapter 4 review)

Section 4.3 discusses multiple regression and presents five tutorials on analyzing multiple variables in `R`. This section is a special topic, meaning you are not required to read or complete these tutorials.

Videos

- Chapter4

Notes

To determine if the relationship between two quantitative variables differs across levels of a categorical variable, you should compare

Simpson's Paradox:

4.2 Activity 4A: Movie Profits — Linear Regression

4.2.1 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.

4.2.2 Terminology review

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

- Scatterplot
- Least-squares line of regression
- Slope and y-intercept
- Residuals

To review these concepts, see Chapter 3 in the textbook.

4.2.3 Movies released in 2016

We will revisit the data set used last week collected on Movies released in 2016 ("IMDb Movies Extensive Dataset" 2016). Here is a reminder of the variables collected on these movies.

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

Vocabulary review

1. What type of plot should be used to display the relationship between `budget_mil` and `revenue_mil`?

2. What three summary statistics could be used to describe the relationship between two quantitative variables?

We will look at the relationship between budget and revenue for movies released in 2016. Enter the explanatory variable name, `budget_mil`, for `explanatory` and the response variable name, `revenue_mil`, for `response` at line 7 in the `R` script file to create the scatterplot. (Note: both variables are measured in "millions of dollars" ($MM).) Highlight and run lines 1–12.

```
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 = "Revenue vs. Budget") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

3. Sketch the scatterplot created from the code.

4. Assess the four features of the scatterplot that describe this relationship. Describe each feature using a complete sentence!

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

5. 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.

```
# 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
```

6. 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.

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

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

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

10. The prediction in question 9 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 line (model).

$$\begin{aligned} &\text{Data} = \text{Model} + \text{Residual} \\ \implies &\text{Residual} = \text{Data} - \text{Model} \\ &e_i = y_i - \hat{y}_i \end{aligned}$$

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

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

4.2.4 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**.

4.2.5 Additional notes

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

4.3 Activity 4B: Movie Profits — Correlation and Coefficient of Determination

4.3.1 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.

4.3.2 Terminology review

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

- Correlation (r or R)
- Coefficient of determination (r-squared or R^2)

To review these concepts, see Chapter 3 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.

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

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

Correlation

Correlation measures the strength and the direction of the linear relationship between two quantitative variables. The closer the value of correlation to $+1$ or -1, the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables. The following output shows a correlation matrix between several pairs of quantitative variables. Highlight and run lines 1–12 to produce the same table as below.

```
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. Using the output above, which two variables have the *strongest* correlation? What is the value of this correlation?

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

3. Based on the value of correlation found in question 2, what would the sign of the slope be? Positive or negative? Explain.

4. Does your answer to question 3 match the direction you choose in question 4 in Activity 4A?

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

Coefficient of determination (squared correlation)

Another summary measure used to explain the linear relationship between two quantitative variables is the coefficient of determination (r^2). The coefficient of determination, r^2, can also be used to describe the strength of the linear relationship between two quantitative variables. The value of r^2 (a value between 0 and 1) represents the **proportion of variation in the response that is explained by the least squares line with the explanatory variable**. There are two ways to calculate the coefficient of determination:

Square the correlation coefficient: $R^2 = (R)^2$

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

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

7. The variance of the response variable, revenue in \$MM, is about $s_{revenue}^2 = 8024.261$ \$MM^2 and the variability in the residuals is about $s_{RES}^2 = 4244.832$ \$MM^2. Use these values to calculate the coefficient of determination. Verify that your answers to 6 and 7 are the same.

In the next part of the activity we will explore what the coefficient of determination measures. Go to the website www.rossmanchance.com/ISIapplets.html and click on Corr/Regresssion under Quantitative Response. Click `Clear` below the box containing the sample data. Download and open the csv file "Movie2016" from D2L. Copy the two columns containing `budget_mil` and `revenue_mil` including the headers and paste into the sample data box. Click 'Use Data'.

8. Click on `Show Moveable Line`. Write down the equation of the line given. Why is the slope zero for this line?

9. Click on `Show Squared Residuals`. Write down the value for SSE. Since this is the sum of squared errors (SSE) for the horizontal line we call this the total sum of squares (SST).

10. Click on `Show Regression Line`. Write down the equation of the line given. Does this match the least squares line found in Activity 4A question 6?

11. Click on `Show Squared Residuals`. Write down the value for SSE.

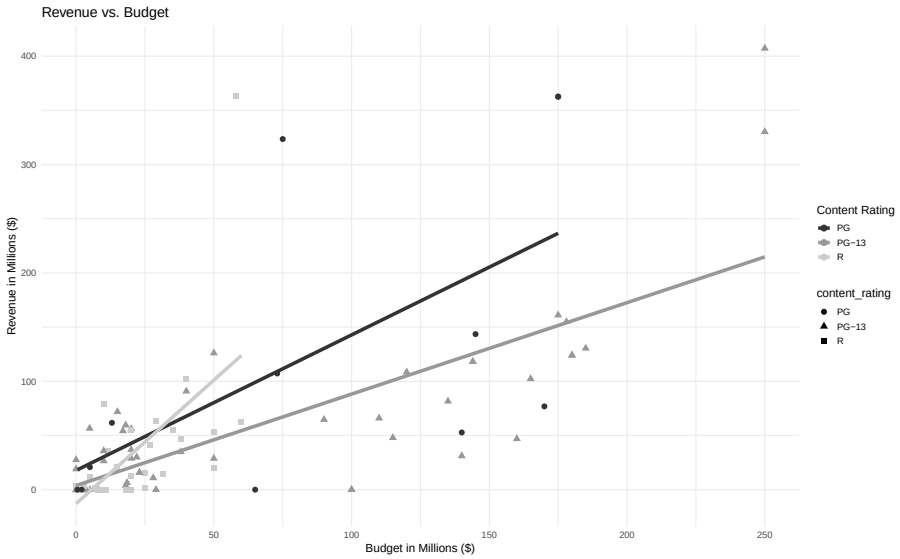
12. Calculate the value for R^2 using the values found for SST and SSE.

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

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 = 3) +  # 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 = "Revenue vs. Budget") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE, lwd = 2) + # Add regression lines
  scale_color_grey() # Make black and white
```



14. Identify the three variables plotted in this graph.

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

4.3.4 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 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.

4.3.5 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.

4.4.2 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 flipper length and body mass. The researchers were interested in the relationship between flipper length and body mass and wondered if flipper length could be used to accurately predict the body mass of these three penguin species.

Upload and import the `Antarctica_Penguins` csv file and 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 4.

First we will create a scatterplot of the flipper length and body mass. Notice that we are using flipper length to predict body mass. This makes flipper length the explanatory variable. **Make sure to give your plot a descriptive title.** Highlight and run lines 1–13 in the `R` script file. **Upload a copy of your scatterplot to Gradescope.**

```
penguins <- datasetname #Creates the object penguins
penguins %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "flipper length (mm)",  # Label x-axis
       y = "body mass (g)",  # Label y-axis
       title = "Title") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

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

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

Highlight and run lines 16–20 to get the correlation matrix 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, which two variables have the *strongest* correlation? What is the value of this correlation?

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.**

Enter the variable `body_mass_g` for `response` and the variable name `flipper_length_mm` for `explanatory` in line 23 in the R script file. Highlight and run lines 23–24.

```
# 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 flipper length of 181 mm.**

8. One penguin had a flipper length of 181 mm and a body mass of 3750 g. Find the residual for this penguin.

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

Highlight and run lines 27–34 to get the multivariate plot.

```
penguins %>%
  ggplot(aes(x = flipper_length_mm, y = body_mass_g, color=species))+  # Specify variables
  geom_point(aes(shape = species), size = 3) +  # Add scatterplot of points
  labs(x = "flipper length (mm)",  # Label x-axis
       y = "body mass (g)",  # Label y-axis
       color = "species",
       title = "TITLE") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

10. What three variables are plotted on this plot?

11. **Does adding the variable species affect the relationship between body mass and flipper length? Explain.**

MODULE 5

Exam 1 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the Exam 1 Review `R` script file to answer the following questions. Each adult (>21) islander was selected at random from all the adult islanders. Variables and their descriptions are listed below. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measure before listening to the music and then 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
`Gender`	Gender of Islander (M = male, F = Female)
`Age`	Age of Islander
`Married`	Marital status of Islander
`Smoking_Status`	Whether the Islander is a current smoker
`Children`	Whether the Islander has children
`weight_kg`	Weight measured in kg
`height_cm`	Height measured in cm
`respiratory_rate`	Breaths per minute
`Type_of_Music`	Music type (Classical or Heavy Medal) Islander was randomly assigned to listen to
`Before_PuzzleCube`	Time to complete puzzle cube (minutes) before listening to assigned music
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed (note: missing data depicted by missing)
`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. List all the variables that are categorical.

3. List all the variables that are quantitative.

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

5. Use the Exam 1 Review `R` script file to find the appropriate summary statistic and graphical display of the data to assess the following research question, "Is the proportion of married Islanders greater than 50%?"

   Variable:

   Value of Summary Statistic (with notation):

   Interpretation:

   Type of Graph:

   Sketch of the graph:

   To what group could the results of this study be applied to?

6. Use the Exam 1 Review `R` script file to find the appropriate 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.

   Explanatory Variable:

   Response variable:

   Value of Summary Statistic (with notation):

      Interpretation:

   Type of Graph:

   Sketch of the graph:

Based on the graph, does there appear to be an association between the two variables? Explain your answer.

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

What is the scope of inference for this study?

7. Use the 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.

   Explanatory Variable:

   Response Variable:

   Value of Summary Statistic (with notation):

      Interpretation:

   Type of Graph:

   Sketch of the graph:

Based on the graph, does there appear to be an association between the two variables? Explain your answer.

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

   Shape:

   Center:

   Spread:

   Outliers:

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

What is the scope of inference for this study?

8. Use the 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 breathes per minute?"

   Explanatory Variable:

   Response Variable:

   Value of Summary Statistic (with notation):

   Slope:

   Interpretation:

   Correlation:

   Interpretation:

   Coefficient of Determination:

   Interpretation:

   Type of Graph:

   Sketch of the graph:

Based on the graph, does there appear to be an association between the two variables? Explain your answer.

Compare the two plots using the four characteristics to describe scatterplots.

Form:

Direction:

Strength:

Outliers:

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

What is the scope of inference for this study?

MODULE 6

Inference for a Single Categorical Variable: Simulation-based Methods

6.1 Module 6 Reading Guide: Categorical Inference

Section 5.1 (Foundations of inference: Hypothesis tests)

Please note that Theory-based inference will be covered next week.

Videos

- 5.1

Vocabulary

Statistical inference:

Hypothesis test:

Also called a 'significance test.'

Simulation-based method:

Theory-based method:

Central Limit Theorem:

Sampling distribution:

Standard deviation of a statistic:

Standard error of a statistic:

Null hypothesis (H_0):

Alternative hypothesis (H_A):

P-value:

Point estimate:

Test statistic:

Decision:

Significance level (α):

Statistically significant:

Confidence interval:

Margin of error:

Notes

What 'theory' is behind the theory-based methods of analysis?

Consider the US judicial system:

What is the null hypothesis?

What is the alternative hypothesis?

The jury is presented with evidence.

- If the evidence is strong (beyond a reasonable doubt), the jury will find the defendant:

- If the evidence is not strong (not beyond a reasonable doubt), the jury will find the defendant:

To create a simulation, which hypothesis (null or alternative) do we assume is true?

More on p-values:

Lower the p-value:

Interpretations require:

General steps of a hypothesis test:

Conclusions should include:

Decision:

If p-value $\leq \alpha$, the decision is to:

If p-value $> \alpha$, the decision is to:

True or False: If the p-value is above 0.10, that means the null hypothesis is true.

True or False: When conducting a simulation-based hypothesis test, the null hypothesis is assumed to be true to create the simulation.

Formulas

$SD(\hat{p}) =$

General form of a theory-based confidence interval:

Margin of error:

Example: Martian alphabet

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What are the two possible explanations for how these data could have occurred?

3. Of the two explanations, which is the null and which is the alternative hypothesis?

4. How could coins be used to create a simulation of what should happen if everyone in the class was just guessing?

5. How can we use the simulation to determine which of the two possibilities is more believable?

6. What decision should be made at an $\alpha = 0.05$ significance level? Justify your answer.

7. Are the results in this example statistically significant? Justify your answer.

8. Interpret the 95% confidence interval provided in the textbook.

9. The formula for the interval is $34/38 \pm (2 \times 0.08) = 0.89 \pm 0.16$. Calculating that, you should get (0.73, 1.05). Why was the interval shown in the textbook (0.73, 1) instead of (0.73, 1.05)?

Section 5.3 (Inference for one proportion)

You may skip Section 5.3.4, which will be covered next week.

Videos

- 5.3SimInf
- Bootstrapping

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test.'

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter; also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Vocabulary

Point estimate:

Test statistic:

Null value:

Null distribution:

One-sided hypothesis test:

Two-sided hypothesis test:

Bootstrapping:

Bootstrapped resample:

Bootstrapped statistic:

Notes

Which hypothesis must we assume is true in order to simulate a null distribution?

Explain the differences between a one-sided and two-sided hypothesis test.

How will the research questions differ?

How will the notation in the alternative hypothesis differ?

How does the p-value calculation differ?

How does the p-value in a two-sided test compare to the p-value in a one-sided test?

Should the default in research be a one-sided or two-sided hypothesis test? Explain why.

Purpose of bootstrapping:

How is bootstrapping used?

If we want to find a 90% confidence interval, what percentiles of the bootstrap distribution would we need?

Example: Organ donations

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

3. Write the null and alternative hypotheses in words, using the example in 5.3.1.

4. Write the null and alternative hypotheses in notation, using the example in 5.3.1.

5. To simulate the null distribution, we would not be able to use coins. Why not?

6. How could we use cards to simulate 1 sample which assumes the null hypothesis is true? How many blue cards — to represent what? How many red cards — to represent what? How many times would we draw a card and replace it back in the deck? What would you record once you completed the draw-with-replacement process?

7. How can we calculate a p-value from the simulated null distribution for this example in 5.3.1?

8. What was the p-value of the test from the example in 5.3.1?

9. At the 5% significance level, what decision would you make based on the p-value above?

10. What conclusion should the researcher make?

11. Are the results in this example statistically significant? Justify your answer.

12. How does the alternative hypothesis change, both in words and in notation, when the example changes to a two-sided hypothesis test in 5.3.2?

13. Explain how the p-value calculation changes between the example in 5.3.1 (one-sided hypothesis test) and the example in 5.3.2 (two-sided hypothesis test).

14. Why does doubling the p-value from the one-sided hypothesis test (your answer to question 8) not match the two-sided p-value calculated in Figure 5.12?

15. How could we use cards to simulate **one** bootstrapped resample? How many blue cards — to represent what? How many red cards — to represent what? How many times would we draw a card and replace it back in the deck? What would you record once you completed the draw-with-replacement process?

16. Interpret the 95% confidence interval provided in the textbook.

17. Are the results in this example statistically significant? Justify your answer.

6.2 Activity 6: Helperer-Hinderer — Simulation-based Hypothesis Test

6.2.1 Learning outcomes

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

6.2.2 Terminology review

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

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

To review these concepts, see Chapter 5 in your textbook, focusing on Sections 5.1 through 5.3.

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 Martian Alphabet 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

Do young children know the difference between helpful and unhelpful behavior? 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.

Ask a research question

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

Design a study and collect data

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

2. Are the cases independent? Justify your answer.

Summarize and visualize the data

```
 # 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
```

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

3. Using the `R` output, 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.

4. What type of plot should be used to represent these data? Sketch this plot.

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

Use statistical analysis methods to draw inferences from the data

When performing a hypothesis test, we must first identify the null hypothesis. The null hypothesis is written about the parameter of interest, or the value that summarizes the variable in the population. *For example, in the Martian Alphabet Activity, the parameter of interest is the true proportion of statistic students who would correctly identify Bumba.*

5. Write out the parameter of interest for this study.

6. If the children are just randomly choosing the toy, what proportion of infants would choose the helper toy? This is the null value for our study.

7. Using the parameter of interest in question 5, 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, $H_0 : \pi = \pi_0$.

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

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

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

10. Write out the alternative hypothesis in words.

11. 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 **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 cards to create one simulation of 16 infants under the assumption the null hypothesis is true. We will write the response variable outcomes on each card to represent the null hypothesis.

12. How many cards total do we need? On how many cards will we write **helper**? On how many cards will we write **hinderer**?

13. Next, we would mix the cards together and draw 1 card, write down if the card says helper or hinderer, and replace the card. How many times would we need to repeat this process to simulate one sample?

14. 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?

15. Create one simulation using the cards provided. Write down your simulated sample proportion. This is one simulation created under the assumption the null hypothesis is true. Is this value closer to 0.5 the null value or closer to the sample proportion (0.875)? Compare your simulated value to the other group's at your table.

16. Report your simulated sample proportion to your instructor. Sketch the distribution created by your class below.

17. Circle the observed statistic (value from question 3) on the distribution you drew in question 16. 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 likely to happen or unlikely to happen if the true proportion of infants who choose the helper is 0.5? Explain your answer using the plot.

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

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.

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 Module 6 Lab: 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

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

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

6.3.3 Helper-Hinderer

Do young children know the difference between helpful and unhelpful behavior? 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 activity 6.

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

In the last class each group created a single simulation assuming the null hypothesis is true. We plotted these simulations and compared our sample proportion calculated from the data to this simulated distribution.

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 1–15.

```
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?

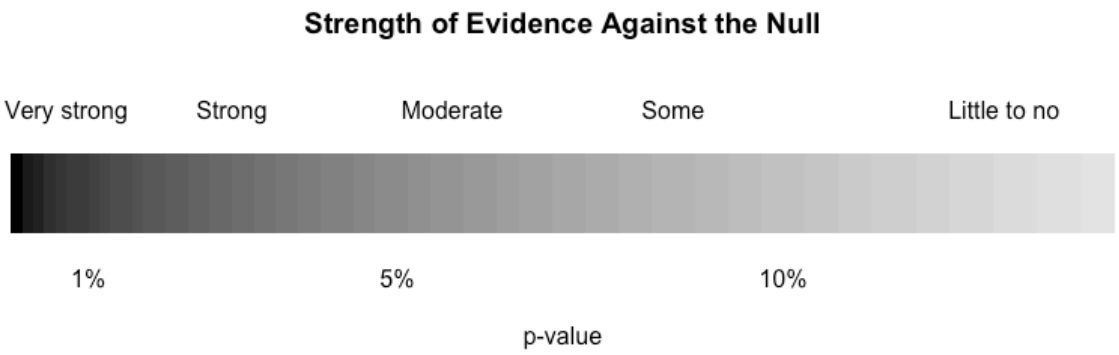
6. Circle the observed statistic (value from question 1) on the distribution you drew in question 4. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

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. **Complete your group's paragraph on Gradescope.**

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. 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.

MODULE 7

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

7.1 Module 7 Reading Guide: Categorical Inference

Section 5.1 (Foundations of inference: Hypothesis tests)

Review section 5.1.2, specifically the notes about the theory-based approach and the Central Limit Theorem.

Section 5.2 (The normal distribution)

Videos

- 5.2

Vocabulary

Normal distribution (Also known as: normal curve, normal model, Gaussian distribution):

Notation:

Standard normal distribution:

Notation:

Z-score:

Xth percentile:

68-95-99.7 rule:

Notes

Interpretation of a Z-score:

True or False: The more unusual observation will be the observation with the largest Z-score.

Approximately what percent of a normal distribution is in the interval

(mean – standard deviation, mean + standard deviation):

(mean – 2×(standard deviation), mean + 2×(standard deviation)):

(mean – 3×(standard deviation), mean + 3×(standard deviation)):

Formulas

Z =

R coding

Calculating normal probabilities When using the `pnorm()` R function, you will need to enter values for the arguments `mean`, `sd`, and `q` to match the question.

```
pnorm(mean = mu, sd = sigma, q = x, lower.tail = TRUE)
```

This function will return the proportion of the N(`mu`,`sigma`) distribution which is *below* the value `x`.

Example: `pnorm(mean = 5, sd = 2, q = 3, lower.tail = TRUE)` will give us the proportion of a N(5,2) distribution which is below 3, which equals 0.159:

```
pnorm(mean = 5, sd = 2, q = 3, lower.tail = TRUE)
#> [1] 0.1586553
```

Changing to `lower.tail = FALSE` will give the proportion of the distribution which is *above* the value `x`.

```
pnorm(mean = 5, sd = 2, q = 3, lower.tail = FALSE)
#> [1] 0.8413447
```

Displaying normal probabilities When using the `normTail()` R function, you will need to enter values for the arguments `m`, `s`, and `L` (or `U`) to match the question.

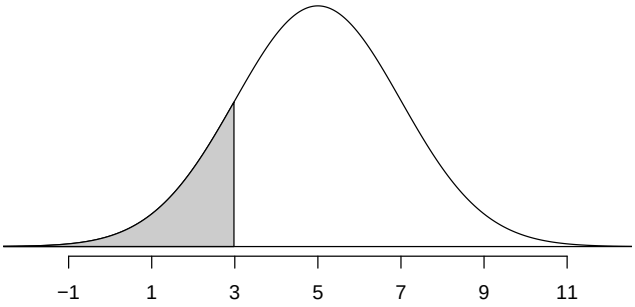
```
normTail(m = mu, s = sigma, L = x)
```

This function (in the `openintro` package) will plot a N(`mu`, `sigma`) distribution and shade the area that is below the value `x`.

Example: `normTail(m = 5, s = 2, L = 3)` creates the plot pictured below.

Changing `L` to `U` will shade the area *above* `x`.

Example: `normTail(m = 5, s = 2, U = 3)` plots a N(5,2) distribution with the area above 3 shaded.



Calculating normal percentiles When using the `qnorm()` `R` function, you will need to enter values for the arguments `mean`, `sd`, and `p` to match the question.

```
qnorm(mean = mu, sd = sigma, p = x, lower.tail = TRUE)
```

This function will return the value on the N(`mu`, `sigma`) distribution which has `x` area of the distribution *below* it.

Example: `qnorm(mean = 5, sd = 2, p = 0.159, lower.tail = TRUE)` will give us the value on a N(5,2) distribution which has 0.159 (15.9%) of the distribution below it, which equals 3 (from the `R` output above).

Changing to `lower.tail = FALSE` will give the value which has `x` area of the distribution *above* it.

We would recommend you work through each of the examples in Section 5.2.4 using `R`.

Section 5.3.4 (Theory-based inferential methods for π)

Videos

- 5.3TheoryInf

Vocabulary

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.

2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test.'

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\Longrightarrow$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample proportion (or mean) will be approximately normal (bell-shaped and symmetric).

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter; also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Vocabulary

Standardized statistic:

Confidence level:

Notes

Conditions for the Central Limit Theorem to apply (for the sampling distribution of $\hat{p}$ to be approximately normal)

Independence:

Checked by:

Success-failure condition:

Checked by:

How can we determine the value of $z^\star$ to use as the multiplier in a confidence interval?

In `R`, use `qnorm(mean = __, sd = __, p = __)`.

Select one answer in each set of parentheses: The higher the confidence level, the (larger/smaller) the multiplier, meaning the confidence interval will be (wider/narrower).

If the success-failure condition for the Central Limit Theorem is not met, what is the appropriate method of analysis? Select one: A. Theory-based approach B. Simulation based approach.

Formulas

$SD(\hat{p}) =$

Null standard error of the sample proportion:

$SE_0(\hat{p}) =$

Standardized statistic (in this case, standardized sample proportion):

$Z =$

Standard error of the sample proportion when we do not assume the null hypothesis is true:

$SE(\hat{p}) =$

Theory-based confidence interval for a sample proportion:

Margin of error of a confidence interval for a sample proportion:

Example: Organ donations

1. What is the sample statistic presented in this example? What notation would be used to represent this value?

2. What is the sample size in this example?

3. Are the conditions met to use theoretical methods to analyze these data? Show your calculations to justify your answer.

Example: Payday loans

1. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

2. Write the null and alternative hypotheses in words.

3. Write the null and alternative hypotheses in notation.

4. Are the conditions met to use theoretical methods to analyze these data? Show your calculations to justify your answer.

5. Calculate the null standard error of the sample proportion.

6. What is the sample statistic presented in this example? What notation would be used to represent this value?

7. Calculate the standardized sample proportion.

8. How can we calculate a p-value from the normal distribution for this example?

9. What was the p-value of the test?

10. At the 5% significance level, what decision would you make?

11. What conclusion should the researcher make?

12. Are the results in this example statistically significant? Justify your answer.

13. Calculate the standard error of the sample proportion when we do not assume the null hypothesis is true.

14. Calculate the margin of error for a 95% confidence interval for π using 1.96 as the multiplier.

15. Calculate a 95% confidence interval for π using your margin of error calculated above.

16. Interpret the 95% confidence interval provided in the textbook.

17. Are the results in this example statistically significant? Justify your answer.

Section 5.4 (Errors, power, and practical importance)

Videos

- 5.4

Reminders from previous sections

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.
- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Vocabulary

Type 1 error:

Type 2 error:

Confirmation bias:

Power:

Practical importance:

Notes

Fill in the following table with whether the decision was correct or not, and if not, what type of error was made.

	Test conclusion (based on data)	
Truth (unknown)	Reject null hyp.	Fail to reject null hyp.
H_0 is true		
H_A is true (H_0 is false)		

How are the significance level and type I error rate related?

How are the significance level and type II error rate related?

After collecting data, a researcher decides to change from a two-sided test to a one-sided test. Why is this a bad idea?

1. It _____ (increases/decreases) the chance of a type I error.
2. This can result in _____.

How are power and type I error rate related?

How are power and type II error rate related?

How can we increase the power of a test?

1. _____ (Increase/Decrease) the significance level
2. _____ (Increase/Decrease) the sample size
3. Change from a _____ (one/two)-sided to a _____ (one/two)-sided test
4. Have a _____ (larger/smaller) standard deviation of the statistic
5. Have the alternative parameter value _____ (closer/farther) from the null value

Results are likely to be statistically significant (but may not be practically important) if the sample size is _____(large/small).

Results are unlikely to be statistically significant (but may be practically important) if the sample size is _____(large/small).

Examples:

1. In the Martian Alphabet study in the textbook and presented as an example in Reading Guide 5.1,

a. What was the p-value of the test?

b. At the 5% significance level, what decision would you make?

c. What type of error might have occurred in these data?

d. Interpret that error in the context of the problem.

2. In the Medical Consultant study in the textbook and presented as an example in the reading guide for sections 5.3.1–5.3.3,

   a. What was the p-value of the test?

   b. At the 5% significance level, what decision would you make?

   c. What type of error might have occurred in these data?

   d. Interpret that error in the context of the problem.

3. In the Payday Loans study in the textbook and presented as an example in the reading guide for section 5.3.4,

   a. What was the p-value of the test?

   b. At the 5% significance level, what decision would you make?

   c. What type of error might have occurred in these data?

   d. Interpret that error in the context of the problem.

7.2 Activity 7A: Helper-Hinderer — Simulation-based Confidence Interval

7.2.1 Learning outcomes

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

7.2.2 Terminology review

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

- Parameter of interest
- Bootstrapping
- Confidence interval

To review these concepts, see Chapter 5 in your textbook, focusing on Sections 5.1 through 5.3.

7.2.3 Helper-Hinderer

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

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

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

Activity intro

1. What is the value of the point estimate?

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

In today's activity, we will use bootstrapping to find a 95% confidence interval for π, the parameter of interest. See Section 5.3.3 in your textbook to review bootstrapping.

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

Use statistical analysis methods to draw inferences from the data

4. Write out the parameter of interest for this study in words. *Hint: this is the same as question 5 in Activity 6.*

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).

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

- Sample size:

- Number of successes:

- Number of repetitions:

- Confidence level (as a decimal):

We will use the `one_proportion_bootstrap_CI()` function in `R` (in the `catstats` package) to simulate the bootstrap distribution of sample proportions and calculate a confidence interval. Using the provided `R` script file, fill in the values/words for each `xx` with your answers from question 5 in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations. Then highlight and run lines 1–7.

```
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
```

6. Sketch the bootstrap distribution created below.

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

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

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

10. Interpret the 95% confidence interval in context.

Communicate the results and answer the research question

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

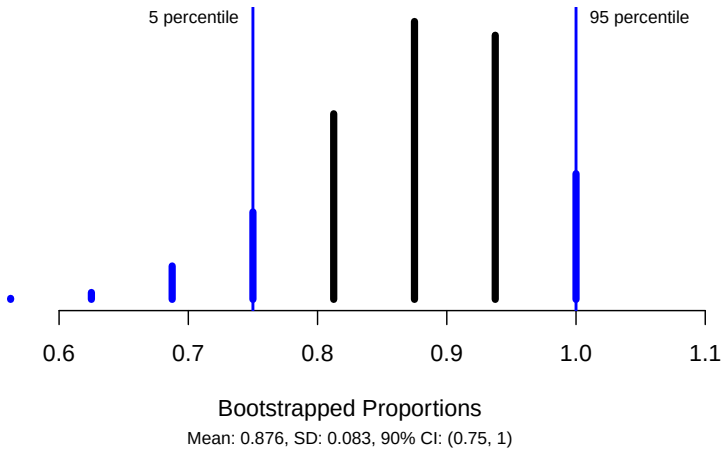
Explain how this indicates that the p-value provides strong evidence against the null.

Effect of confidence level

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

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

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



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

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

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?"

For this part of the activity we will assume that the the true proportion of infants that choose the helper toy is 0.75. *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.*

16. 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 0.75 for π (the true value), 16 for n, and 100 for 'Number of intervals.' Click 'sample.'

    a. 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 of 0.75?

    b. Now click on a red dot. Write down the confidence interval for this sample. Does this confidence interval contain the true value of 0.75?

    c. How many intervals out of 100 contain π, the true value of 0.75? *Hint*: This is given to the left of the graph of green and red intervals.

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

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

7.2.4 Take-home messages

1. The goal in a hypothesis test is to assess the strength of evidence for an effect, while the goal in creating a confidence interval is to determine how large the effect is. A **confidence interval** is a range of *plausible* values for the parameter of interest.

2. A confidence interval is built around the point estimate or observed calculated statistic from the sample. This means that the sample statistic is always the center of the confidence interval. A confidence interval includes a measure of sample to sample variability represented by the **margin of error**.

3. In simulation-based methods (bootstrapping), a simulated distribution of possible sample statistics is created showing the possible sample-to-sample variability. Then we find the middle X percent of the distribution around the sample statistic using the percentile method to give the range of values for the confidence interval. This shows us that we are $X\%$ confident that the parameter is within this range, where X represents the level of confidence.

4. When the null value is within the confidence interval, it is a plausible value for the parameter of interest; thus, we would find a larger p-value for a hypothesis test of that null value. Conversely, if the null value is NOT within the confidence interval, we would find a small p-value for the hypothesis test and strong evidence against this null hypothesis.

5. To create one simulated sample on the bootstrap distribution for a sample proportion, label n cards with the original responses. Draw with replacement n times. Calculate and plot the resampled proportion of successes.

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 7B: Handedness of Male Boxers — Theory-based Methods

7.3.1 Learning outcomes

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

7.3.2 Terminology review

In today's activity, we will introduce theory-based confidence intervals 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 5 in your textbook, focusing on Sections 5.1 through 5.3.

Activity 6 and the Week 6 Lab covered simulation-based methods for hypothesis tests involving a single categorical variable. This activity covers theory-based methods for testing a single categorical variable.

7.3.3 Handedness of male boxers

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

```
 # 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 for this 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, $n\pi \geq 10$ and $n(1-\pi) \geq 10$.

5. Verify that the independence condition is satisfied.

6. Is the success-failure condition met to model the data with the normal distribution? Show your work to support your answer. Hint: We don't know the true value of the parameter, π, so we use the null value, π_0, to check the success-failure condition.

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:

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

.

7. Calculate the null standard error of the sample proportion.

8. Calculate the standardized sample proportion.

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.

9. Using the 68-95-99.7 rule in Section 5.2.5 to guide you, fill in the percentages on the standard normal distribution displayed in Figure 7.1, and also mark the value of the standardized statistic calculated in question 8.

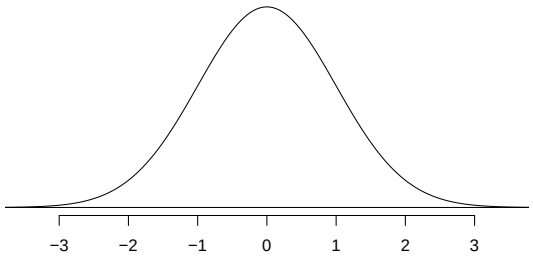


Figure 7.1: A standard normal curve.

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

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

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 calculated in question 8 at `xx` in line 7; highlight and run lines 7–9. Notice that in line 9 it says `lower.tail = FALSE`. `R` will calculate the p-value *greater* than the value of the standardized statistic.

Notes:

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

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

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

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

Validity conditions for a confidence interval

To check the success-failure condition to use theory-based methods for confidence intervals, we use $\hat{p}$ in the calculations since we are not assuming a value for π. That is, check that we have at least 10 successes and 10 failures in our **sample**: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

13. Verify that the success-failure condition is met to use theory based methods to find a 95% 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 $SD(\hat{p})$:

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

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

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

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

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

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

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

15. Fill in the normal distribution shown in figure 7.2 to show how `R` found the z^* multiplier.

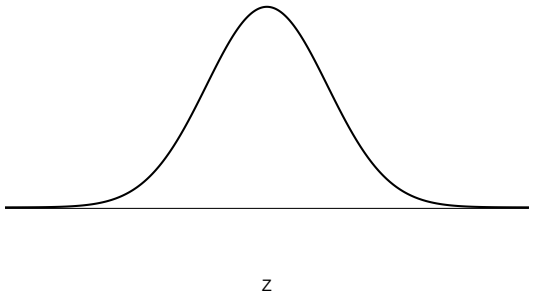


Figure 7.2: A standard normal curve.

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

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

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

16. What is the value of the multiplier needed to calculate the 95% confidence interval for the true proportion of male boxers that are left-handed?

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

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

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

20. Is the null value, 0.1, contained in the 95% confidence interval? Explain, based on your conclusion in question 12, why you expected this to be true.

7.3.4 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.

7.3.5 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 1 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 Sections 5.1, 5.2, and 5.4 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 ($\hat{p}$) 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 errors 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 Module 8 Reading Guide: Hypothesis Testing for a Difference in Proportions

Section 5.5 (Simulation-based inference for a difference in proportions)

You may skip section 5.5.3, which will be covered in the next module, as well as the material on relative risk in section 5.5.1, which will be covered in module 14.

Videos

- 5.5SimInf

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test (standardized) statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test.'

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\Longrightarrow$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Vocabulary

Randomization test:

Notes

In a randomization test involving two categorical variables, how many cards will you need and how will the cards be labeled?

Why, in the randomization test, are the cards all shuffled together and randomly dealt into two new groups?

After shuffling, how many cards are dealt into each pile?

To create a single bootstrap resample for two categorical variables, how many cards will you need and how will the cards be labeled?

What is done with the cards once they are labeled?

Interpretations of confidence level must include:

How do you determine if the results of a hypothesis test agree with a confidence interval?

How are the confidence level and the significance level related (for a two-sided test)?

Notation

Sample size of group 1:

Sample size of group 2:

Sample proportion of group 1:

Sample proportion of group 2:

Population proportion of group 1:

Population proportion of group 2:

Example: Gender discrimination

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample statistic presented in this example? What notation would be used to represent this value?

6. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

10. How can we calculate a p-value from the simulated null distribution for this example?

11. What was the p-value of the test?

12. At the 5% significance level, what decision would you make?

13. What conclusion should the researcher make?

14. Are the results in this example statistically significant? Justify your answer.

Example: Opportunity cost

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample statistic presented in this example? What notation would be used to represent this value?

6. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

10. How can we calculate a p-value from the simulated null distribution for this example?

11. What was the p-value of the test?

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

13. At the 5% significance level, what decision would you make?

14. What conclusion should the researcher make?

15. Are the results in this example statistically significant? Justify your answer.

Example: CPR and blood thinner

1. What is the research question?

2. What are the observational units?

3. What type of study design was used? Justify your answer.

4. What is the appropriate scope of inference for these data?

5. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

6. What is the parameter (using a difference in proportions) representing in the context of this problem? What notation would be used to represent this parameter?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

10. How can we calculate a p-value from the simulated null distribution for this example?

11. What was the p-value of the test?

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

13. At the 5% significance level, what decision would you make?

14. What conclusion should the researcher make?

15. Are the results in this example statistically significant? Justify your answer.

16. How could we use cards to simulate **one** bootstrap resample? How many blue cards — to represent what? How many red cards — to represent what? What would we do with the cards? What would you record once you have a simulated sample?

17. How can we calculate a 90% confidence interval from the bootstrap distribution for this example?

18. What was the 90% confidence interval?

19. Interpret the confidence *interval* in the context of the problem.

20. Interpret the confidence *level* in the context of the problem.

21. Does the conclusion of the hypothesis test match the confidence interval?

8.2 Activity 8A: The Good Samaritan — Simulation-based Hypothesis Test

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.
- 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.

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 5 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(Behavior) %>% count(Condition)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Behavior [2]
#>   Behavior Condition     n
#>   <chr>    <chr>     <int>
#> 1 Help     Hurry         2
#> 2 Help     No hurry     11
#> 3 No help  Hurry        18
#> 4 No help  No hurry      9
```

Vocabulary review

1. What is the name of the explanatory variable 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	
		Randomized Experiment	Observational Study
Selection of Cases	Random Sample		
	No Random Sample		

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 for this study. Let 1 = hurry condition, 2 = no hurry condition.

   π_1 —

   π_2 —

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 $\neq$):

   $H_A : \pi_1 - \pi_2$ _____ 0

Summarize and visualize the data

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

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

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

11. What is the notation used for the value calculated in question 10?

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.

12. 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?

13. 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?

14. 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?

15. Simulate one sample 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`.

16. 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")
```

17. Sketch the null distribution created here.

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

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

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

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

22. Write a conclusion to the test.

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. We use the same guidelines for the strength of evidence as we did in Activity 6.
3. 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 8B: The Good Samaritan (continued) — Simulation-based Confidence Interval

8.3.1 Learning outcomes

- Identify the parameter of interest 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:

- Parameter of interest
- Bootstrapping
- Confidence interval
- Types of errors

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

8.3.3 The Good Samaritan

In the last activity, we found a small p-value for the hypothesis test for a difference in proportions. There was very strong evidence that those in a hurry will be less likely to help people in need. In today's 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.

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

Vocabulary review

1. Report the point estimate for this study.

Use the provided `R` script file to create a segmented bar plot of those who helped others for those in the hurry condition and those in the no hurry condition. Enter the name of the explanatory variable for `explanatory` and the name of the response variable for `response`. **Make sure to title your plot**. Highlight and run lines 1–13.

```
good <- read.csv("https://math.montana.edu/courses/s216/data/goodsam.csv")
good %>%
  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 = "Title",  # Make sure to title your plot
       x = "Condition",   # Label the x axis
       y = "") +  # Remove y axis label
  scale_fill_grey()  # Make figure black and white
```

2. Sketch the segmented bar plot created here.

3. Based on the segmented bar plot, does there appear to be an association between the condition assigned and the behavior? Explain.

4. Write out the conclusion you made in Activity 8A.

5. 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

6. 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`.

7. 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
```

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

9. Report the bootstrap 99% confidence interval.

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

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

12. What conclusion to the research question can be made based on the 99% confidence interval?

13. Is the null value in the 99% confidence interval?

Table 8.3: Four different possible scenarios for hypothesis test decisions.

		Test conclusion	
		Fail to reject H_0	Reject H_0
	H_0 true	Good decision	Type 1 Error
Truth	H_A true	Type 2 Error	Good decision

Types of errors

Recall from a previous activity, hypothesis tests are not flawless. In a hypothesis test, there are two competing hypotheses: the null and alternative. We make a decision about which might be true, but we may choose incorrectly.

Shown in Table 8.3, a **Type 1 Error** happens when we reject the null hypothesis when H_0 is actually true. A **Type 2 Error** happens when we fail to reject the null hypothesis when the alternative is actually true.

14. Using a significance level of 0.01 and your answer to question 13, what decision do you make in regards to the null hypothesis?

15. What type of error could we have made?

16. Write this error in context of the problem.

8.3.4 Take-home messages

1. 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.
2. 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: Fatal Injuries in the Iliad

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 Fatal Injuries in the Iliad

Homer's Iliad is an epic poem, compiled around 800 BCE, that describes several weeks of the last year of the 10-year siege of Troy (Ilion) by the Achaeans. The story centers on the rage of the great warrior Achilles. But it includes many details of injuries and outcomes, and is thus the oldest record of Greek medicine. The data report 146 recorded injuries for which both injury site and outcome are provided in the Illiad (Hutchison and Hirthler 2013). For this activity we will focus on comparing injuries to the body and injuries to a limb and whether the injury resulted in a fatality. Are injuries to the body more lethal than injuries to a limb?

Upload and open the `R` script file for Week 8 lab. Upload and import the csv file, `iliad`. Enter the name of the data set (see the environment tab) for datasetname in the R script file in line 6. Highlight and run lines 1–11 to get the counts for each combination of categories.

```
injuries<- datasetname# Read data set in
injuries <- injuries %>%
  filter(Injury.Site != "Head/neck") #Removes the injuries to head and neck
injuries <- injuries%>%
  filter(Location != "Unknown") #Removes the unknown locations of injuries
injuries %>% group_by(Injury.Site) %>% count(Lethal) #finds the counts in each group
```

1. What is the explanatory variable?

2. What is the response variable?

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

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

	Injury Site		
Outcome	Body	Limb	Total
Fatal			
Non-fatal			
Total			

5. Write the parameter of interest for this study.

6. **Calculate the difference in proportion of fatal injuries for those inflicted on the body and those inflicted on a limb. Use body - limb for the order of subtraction. Use appropriate notation.**

Use the provided R script file to create a segmented bar plot of the data. Make sure to title your plot. Highlight and run lines 14–20.

```
injuries %>% # Data set piped into...
  ggplot(aes(x = Injury.Site, fill = Lethal)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Title",  # Make sure to title your plot
       x = "Location of Injury",   # Label the x axis
       y = "") +  # Remove y axis label
  scale_fill_grey()  # Make figure black and white
```

7. Based on the plot does there appear to be an association between the variables? Explain your answer.

8. Write the null hypothesis for this study in notation.

9. **Using the research question, write the alternative hypothesis in words.**

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 p-value for the test.

```
two_proportion_test(formula = response~explanatory, # response ~ explanatory
    data= injuries, # 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")
```

10. **Report and interpret the p-value in context of the problem.**

11. 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=injuries, # 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 = 0.9) # Enter the level of confidence as a decimal
```

12. Report and interpret the 90% confidence interval.

13. Write a conclusion to the research question in context of the study.

14. **What type of error could have occurred?**

15. Interpret this error in context of the study.

16. Write a paragraph summarizing the results of the study as if writing a press release. Be sure to describe:

- Summary statistic and interpretation
- P-value and interpretation
- Confidence interval and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

Upload your group's confidence interval interpretation and conclusion to Gradescope.

MODULE 9

Inference for Two Categorical Variables: Theory-based Methods

9.1 Module 9 Reading Guide: Hypothesis Testing for a Difference in Proportions

Section 5.5.3 (Theory-based methods for a difference in proportions)

Videos

- 5.5TheoryInf

Reminders from previous sections

n = sample size

$\hat{p}$ = sample proportion

π = population proportion

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test (standardized) statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is.

Also called a 'significance test.'

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\implies$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to 'reject' or 'fail to reject' a null hypothesis based on a p-value and a pre-set level of significance.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Notes

Conditions for the CLT to apply for a difference in proportions

Independence:

Checked by:

Success-failure condition:

Checked by:

Formulas

$SD(\hat{p_1} - \hat{p_2}) =$

Null standard error of the difference in sample proportions: $SE_0(\hat{p_1} - \hat{p_2}) =$

Standardized statistic/standardized difference in sample proportions: $Z =$

Standard error of the difference in sample proportions when we do not assume the null hypothesis is true: $SE(\hat{p_1} - \hat{p_2}) =$

Theory-based confidence interval for a difference in proportions:

Margin of error of a confidence interval for a difference in proportions:

Notation

Overall (pooled) proportion of successes:

Example: CPR and blood thinner

1. What are the observational units?

2. What type of study design was used? Justify your answer.

3. What is the appropriate scope of inference for these data?

4. What is the sample difference in proportions presented in this example? What notation would be used to represent this value?

5. What is the parameter (using a difference in proportions) representing in the context of this problem? What notation would be used to represent this parameter?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. Is it valid to use theory-based methods to analyze these data?

9. Calculate the pooled or overall proportion of successes. What notation would be used to represent this value?

10. Calculate the null standard error of the difference in sample proportions.

11. Calculate the standardized statistic.

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

Note: a p-value, p-value interpretation, decision, and conclusion for this example can be found in the Reading Guide solutions for Sections 5.5.1–5.5.2.

13. Calculate the standard error of the difference in sample proportions without assuming a null hypothesis.

14. Calculate the 90% confidence interval using $z^* = 1.65$ as the multiplier.

Note: A confidence interval interpretation and confidence level interpretation for this example can be found in the Reading Guide solutions for Sections 5.5.1–5.5.2.

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

9.2.1 Learning outcomes

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

9.2.2 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 5 in your textbook.

9.2.3 Helmet use and head injuries

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), we can see the summary results from a random sample of 3562 skiers and snowboarders involved in accidents 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?

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

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.

1. Write the null and alternative hypotheses in notation.

Ho:

Ha:

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

3. Interpret the difference in 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 Activity 8A. 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**: The success-failure condition holds for each group. Under the null hypothesis, the proportions π_1 and π_2 are equal, so we check the success-failure condition with our best estimate of these values under H_0, the pooled proportion from the two samples,

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

We then check that all four of the following inequalities hold:

$$\hat{p}_{pool} \times n_1 \geq 10, \qquad (1 - \hat{p}_{pool}) \times n_1 \geq 10,$$
$$\hat{p}_{pool} \times n_2 \geq 10, \qquad (1 - \hat{p}_{pool}) \times n_2 \geq 10$$

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

5. Is the success-failure condition met for each group? Show your work to verify your answer.

To calculate the standardized statistic we use:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE},$$

where the null standard error is calculated using the pooled proportion of successes:

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

6. Calculate $SE_0(\hat{p}_1 - \hat{p}_2)$.

7. Calculate the standardized statistic.

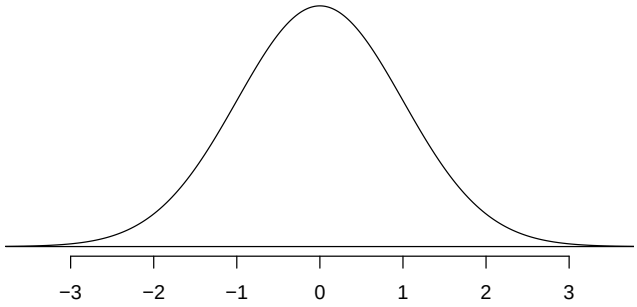


Figure 9.1: A standard normal curve.

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 8 at `xx` in line 2; highlight and run lines 2–4.

```
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.

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.

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

13. How would this impact the p-value?

Suppose in another sample of skiers and snowboarders 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

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

Use Rstudio find the p-value for this new sample. Enter the value of the standardized statistic found in question 14 for xx in line 7. Highlight and run lines 7–9.

```
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
```

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

16. 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 find the p-value for this new sample. Enter the value of the standardized statistic found in question 16 for xx in line 12. Highlight and run lines 12–14.

```
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
```

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

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

a) Switching to a two-sided test.

b) Using a smaller sample size.

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

9.2.4 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. 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.

9.2.5 Additional notes

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

9.3 Week 9B: Winter Sports Helmet Use and Head Injuries — Theory-based Confidence Interval

9.3.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.

9.3.2 Terminology review

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

- Standard normal distribution
- Independence and success-failure conditions

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

9.3.3 Winter sports helmet use and head injury

In this activity we will focus on theory-based methods to calculate a confidence interval. Like with a single proportion, 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**: The success-failure condition holds for each group. Since we are not assuming a null hypothesis, we do not use the pooled sample proportion to check this condition as we did in Activity 9A. Instead, we use the individual sample proportions $\hat{p}_1$ and $\hat{p}_2$. Equivalently, we check that all cells in the table have at least 10 observations.

1. Explain why a theory-based confidence interval for the data set in Activities 8A and 8B would not be similar to the bootstrap interval created.

For this activity we will again use the Helmet Use and Head Injury data set. In Activity 9A we saw that there was evidence that helmet use is assoicated with a reduced risk of head injury. Today we will estimate the difference in proportion of head injuries for those who wore helmets and those who did not.

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., (Sulheim et al. 2017), we can see the summary results from a random sample of 3562 skiers and snowboarders involved in accidents 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. In the last activity we verified that the independence condtion was met. Is the success-failure condition to find the theory-based confidence interval met for each group? Explain your answer.

3. Write the parameter of interest for this study in context of the problem.

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.

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

Note that the formula changes when calculating the variability around the statistic in order to calculate a confidence interval from the formula used in Activity 9A! Here, 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.

4. Calculate the standard error for a difference in proportions to create a 90% confidence interval.

5. Interpret the value calculated in question 4 in context of the problem.

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) # Multiplier for 90% confidence interval
```

```
#> [1] 1.644854
```

6. Sketch a graph of the standard normal distribution and use the graph to explain how the `R` code above is used to find the z^* multiplier.

7. Using the multiplier of $z^* = 1.645$ and the standard error found in question 4, calculate the margin of error for a 90% confidence interval.

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

9. Interpret the confidence interval found in question 8 in context of the problem.

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

11. What decision would you make based on your confidence interval? Explain your answer.

9.3.4 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

12. 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?

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

14. Is the confidence interval calculated in question 13 with the smaller sample size wider or narrower than the confidence interval in question 8? Why?

9.3.5 Take-home messages

1. Simulation-based methods and theory-based methods should give the same 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.

9.3.6 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 5. Highlight and run lines 1–6 to get the counts for each combination of categories.

```
rosi <- datasetname
rosi %>% 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			
Insulin Required			
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= rosi, # 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 that 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. Interpret the confidence interval found in question 12 in context of the problem.

14. Write a conclusion to the research question.

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

- Summary statistic and interpretation
- P-value and interpretation
- Confidence interval and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

Upload a copy of your group's p-value interpretation and scope of inference to Gradescope.

MODULE 10

Exam 2 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the Exam 2 Review `R` script file to answer the following questions. Each adult (>21) islander was selected at random from all the adult islanders. Variables and their descriptions are listed below. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measure before listening to the music and then 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
`Gender`	Gender of Islander (M = male, F = Female)
`Age`	Age of Islander
`Married`	Marital status of Islander
`Smoking_Status`	Whether the Islander is a current smoker
`Children`	Whether the Islander has children
`weight_kg`	Weight measured in kg
`height_cm`	Height measured in cm
`respitory_rate`	Breaths per minute
`Type_of_Music`	Music type (Classical or Heavy Medal) Islander was randomly assigned to listen to
`Before_PuzzleCube`	Time to complete puzzle cube (minutes) before listening to assigned music
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed (note: missing data depicted by missing)
`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 provided Exam 2 Review `R` script file and analyze the following research question: The proportion of university graduates in the US is 42%. "Is there evidence that the proportion of university graduates in the Islands differs from the proportion in the US?"

   Parameter of Interest:

   Null Hypothesis:

   Notation:

   Words:

   Alternative Hypothesis:

   Notation:

   Words:

   Value of Statistic with Notation:

   Conditions:

   Independence:

   Success-Failure:

   Simulation P-value:

   Interpretation:

   Conclusion:

Decision:

Simulation Confidence Interval:

Interpretation:

Standardized Statistic:

Interpretation:

Theory-based p-value:

Theory-based Confidence Interval:

Does the theory-based p-value and CI match those found using simulation methods?

To what group can the results be generalized?

2. Use the provided 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?"

Parameter of Interest:

Null Hypothesis:

Notation:

Words:

Alternative Hypothesis:

Notation:

Words:

Value of Statistic with Notation:

Conditions:

Independence:

Success-Failure:

Simulation P-value:

Interpretation:

Conclusion:

Decision:

Simulation Confidence Interval:

Interpretation:

Standardized Statistic:

Interpretation:

Theory-based p-value:

Theory-based Confidence Interval:

Does the theory-based p-value and CI match those found using simulation methods?

What is the scope of inference for this study?

MODULE 11

Inference for a Quantitative Response with Paired Samples

11.1 Module 11 Reading Guide: Inference for a Single Mean or Paired Mean Difference

Section 6.1 (Inference for one mean)

Videos

- 6.1

Reminders from previous sections

n = sample size

$\overline{x}$ = sample mean

s = sample standard deviation

μ = population mean

σ = population standard deviation

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test.'

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\Longrightarrow$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.
- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample mean (or proportion) will be approximately normal (bell-shaped and symmetric).

Vocabulary

t-distribution:

- The variability in the t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The larger df, the closer the t distribution is to the standard normal distribution.

Degrees of freedom (df):

T-score:

Notes

To create a bootstrap distribution test, how many cards will you need and how will the cards be labeled?

What do you do with the cards after labeling them?

After resampling, what value will be plotted on the bootstrap distribution?

True or false: Bootstrapping can only be used if the sample size is small.

Why do we use a t-distribution rather than the normal distribution when analyzing quantitative data?

How do we calculate degrees of freedom for the t-distribution?

Conditions to use the CLT for means:

Independence:

Checked by:

Normality:

Checked by:

Formulas

$SE(\overline{x}) =$

$T =$

Confidence interval for a mean:

Notation

μ_0 represents

Example: Edinburgh rentals

1. What are the observational units?

2. What are the sample statistics presented in this example? What notation would be used to represent each value?

3. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

4. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

5. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

6. Based on Figure 6.3, give the confidence interval for the true mean for each of the following confidence levels.

90% confidence interval =

95% confidence interval =

99% confidence interval =

7. Interpret your 99% confidence interval in the context of the problem.

8. Use Figure 6.4 to determine a 90% confidence interval for the true standard deviation for three bedroom flats in Edinburgh.

Example: Mercury content of dolphin muscle

1. What is the research question?

2. What are the observational units?

3. Can the results of this study be generalized to a larger population? Why or why not?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. Are the independence and normality conditions satisfied?

7. Calculate the standard error of the sample mean.

8. What distribution should be referenced to find the multiplier for a 95% confidence interval?

9. Using $t^\star = 2.10$, calculate a 95% confidence interval for μ.

10. Interpret the interval calculated in the context of the problem.

Example: Cherry Blossom Race

1. What is the research question?

2. What are the observational units?

3. Can the results of this study be generalized to a larger population? Why or why not?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. Are the independence and normality conditions satisfied?

7. Write the null and the alternative hypotheses in words.

8. Write the null and the alternative hypotheses in notation.

9. Calculate the standard error of the sample mean.

10. Calculate the T-score (the standardized statistic for the sample mean).

11. What distribution should the T-score be compared to in order to calculate a p-value?

12. What was the p-value of the test?

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

14. At the 5% significance level, what decision would you make? What type of error might that be?

15. What conclusion should the researcher make?

16. Are the results in this example statistically significant? Justify your answer.

Section 6.2 (Inference for paired mean difference)

Videos

- 6.2

Vocabulary

Paired data:

Paired with repeated measures:

Paired with matching:

Notes

For each of the following scenarios, determine if the two sets of observations are paired or independent.

1. To test whether the IQ is related to genetics, researchers measured the IQ of two biological parents and the IQ of their first-born child. The average parent IQ was compared to the IQ of the first born child.

2. Hoping to see how exercise is related to heart rates, researchers asked a group of 30 volunteers to do either bicycle kicks or jumping jacks for 30 seconds. Each volunteer's heart rate was measured at the end of 30 seconds, then the volunteer sat for a 5 minute rest period. At the end of the rest period, the volunteer performed the other activity and their heart rate was measured again. Which activity was done first was randomly assigned.

3. Researchers hoping to look into the effectiveness of blended learning gathered two random samples of 50 8th graders (one at Belgrade Middle School which had 5 full-day instruction at the time of the study, the other from Chief Joseph Middle School which utilized a 2-day on, 3-day off blended learning structure). All 8th graders were given the same lessons and same homework, then asked to take the same end-of-unit test.

Conditions to use the CLT for paired mean difference:

Independence:

Checked by:

Normality:

Checked by:

Formulas

$SE(\overline{x_d}) =$

$T =$

Confidence interval for a paired mean difference:

Notation

$\overline{x_d} =$

$s_d =$

$\mu_d =$

$\sigma_d =$

Example: Tires

1. What are the observational units?

2. Why should we treat these data as paired rather than two independent samples?

3. What are the sample statistics presented in this example? What notation would be used to represent each value?

4. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

5. Write the null and alternative hypotheses in appropriate notation.

6. How could we use cards to simulate **one** bootstrap resample *which assumes the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting null distribution centered? Why does that make sense?

8. What was the p-value of the test? Interpret this p-value in the context of the problem.

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

Example: College textbook prices

1. What is the research question?

2. What are the observational units?

3. Why should we treat these data as paired rather than two independent samples?

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

8. Give the 95% confidence interval for μ_d.

9. Interpret your 95% confidence interval in the context of the problem.

10. Are the independence and normality conditions satisfied?

11. Write the null and the alternative hypotheses in words.

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

13. Calculate the T-score (the standardized statistic for the sample mean difference).

14. What distribution should the T-score be compared to in order to calculate a p-value?

15. What was the p-value of the test?

16. At the 5% significance level, what decision would you make? What type of error might that be?

17. What conclusion should the researcher make?

18. Are the results in this example statistically significant? Justify your answer.

19. Using $t^\star = 2.00$, calculate a 95% confidence interval for μ_d.

20. Interpret the interval calculated in the context of the problem.

11.2 Activity 11A: COVID-19 and Air Pollution

11.2.1 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 simulation-based hypothesis test for a paired mean difference.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a paired mean difference.
- Use bootstrapping to find a confidence interval for a paired mean difference.
- Interpret a confidence interval for a paired mean difference.
- Use a confidence interval to determine the conclusion of a hypothesis test.

11.2.2 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 6.2 in the textbook.

11.2.3 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 11.1, 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 11.1: The India Gate in New Delhi, India.

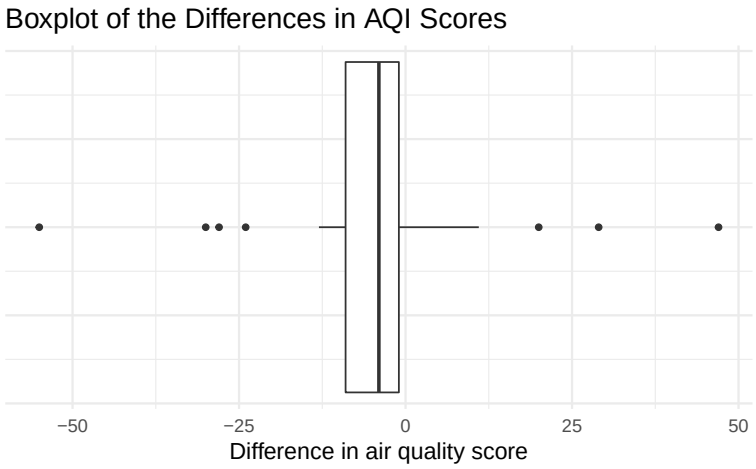


Table 11.1: Summary statistics for current AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

	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. What are the two competing possibilities to run a hypothesis test for this study?

5. Write the null hypothesis in words.

6. What is the research question?

7. Write the alternative hypothesis in notation.

Summarize and visualize the data

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

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

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.

10. 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 10 for `xx` to simulate the null distribution and enter the summary statistic from question 8 for `yy` to find the p-value. Highlight and run lines 1–21.

```
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
```

11. Sketch the null distribution created using the `R` output here.

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

13. 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?

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

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

16. 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.

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

18. Using the provided R script file, fill in the missing value at xx to find a 99% bootstrap confidence interval; highlight and run lines 24–27. 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

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

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

11.2.4 Take-home messages

1. The differences in a paired data set are treated like a single quantitative variable when performing a statistical analysis. Paired data (or paired samples) occur when pairs of measurements are collected. We are only interested in the population (and sample) of differences, and not in the original data.
2. When using bootstrapping to create a null distribution centered at the null value for both paired data and a single quantitative variable, we first need to shift the data by the difference $\mu_0 - \bar{x}_d$, and then sample with replacement from the shifted data.
3. When analyzing paired data, the summary statistic is the 'mean difference' NOT the 'difference in means'[1]. This terminology will be *very* important in interpretations.
4. To create one simulated sample on the null distribution for a sample mean or mean difference, shift the original data by adding $(\mu_0 - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Calculate and plot the sample mean or the sample mean difference.
5. To create one simulated sample on the bootstrap distribution for a sample mean or mean difference, label n cards with the original response values. Randomly draw with replacement n times. Calculate and plot the resampled mean or the resampled mean difference.

[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 Activity 11.

11.2.5 Additional notes

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

11.3 Activity 11B: Color Interference

11.3.1 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 theory-based hypothesis test for a paired mean difference.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a paired mean difference.
- Use theory-based methods to find a confidence interval for a paired mean difference.
- Interpret a confidence interval for a paired mean difference.
- Use a confidence interval to determine the conclusion of a hypothesis test.

11.3.2 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 Sections 6.1 and 6.2 in the textbook.

11.3.3 Color Interference

The abstract of the article "Studies of interference in serial verbal reactions" in the *Journal of Experimental Psychology* (Stroop 1935) reads:

> In this study pairs of conflicting stimuli, both being inherent aspects of the same symbols, were presented simultaneously (a name of one color printed in the ink of another color—a word stimulus and a color stimulus). The difference in time for reading the words printed in colors and the same words printed in black is the measure of interference of color stimuli upon reading words. … The interference of conflicting color stimuli upon the time for reading 100 words (each word naming a color unlike the ink-color of its print) caused an increase of 2.3 seconds or 5.6% over the normal time for reading the same words printed in black.

The article reports on the results of a study in which seventy college undergraduates were given forms with 100 names of colors written in black ink, and the same 100 names of colors written in another color (i.e., the word purple written in green ink). The total time (in seconds) for reading the 100 words printed in black, and the total time (in seconds) for reading the 100 words printed in different colors were recorded for each subject. The order in which the forms (black or color) were given was randomized to the subjects. Does printing the name of colors in a different color increase the time it takes to read the words? Use color - black as the order of subtraction.

Identify the scenario

1. Should these observations be considered paired or independent? Explain your answer.

2. Based on your answer to question 1, is the appropriate summary measure to be used to analyze these data the difference in mean times or the mean difference in times?

Ask a research question

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

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

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

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.
  - $n \geq 30$: If the sample size n is at least 30 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**!

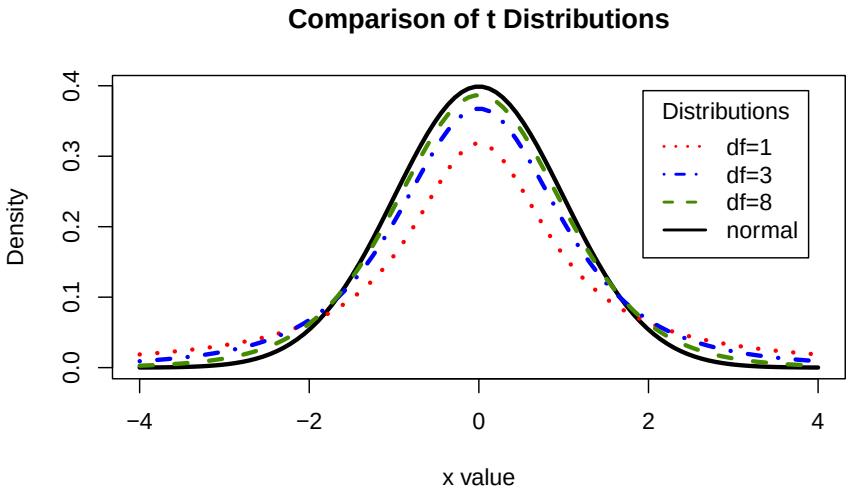


Figure 11.2: Comparison of the standard Normal vs t-distribution with various degrees of freedom

As you can seen from Figure 11.2, 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!)

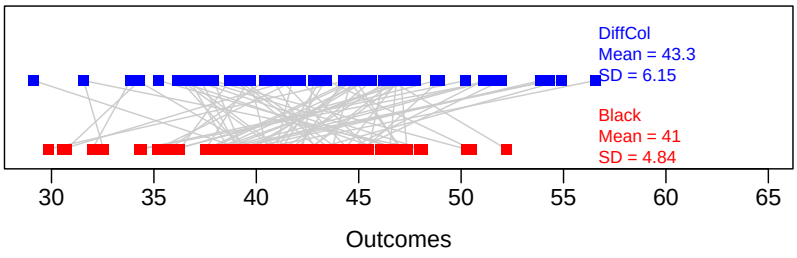
Summarize and visualize the data

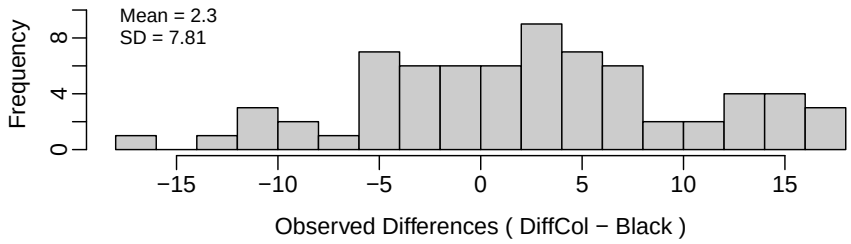
Since the original data from the study are not available, we simulated data to match the means and standard deviations reported in the article. We will use these simulated data in the analysis below.

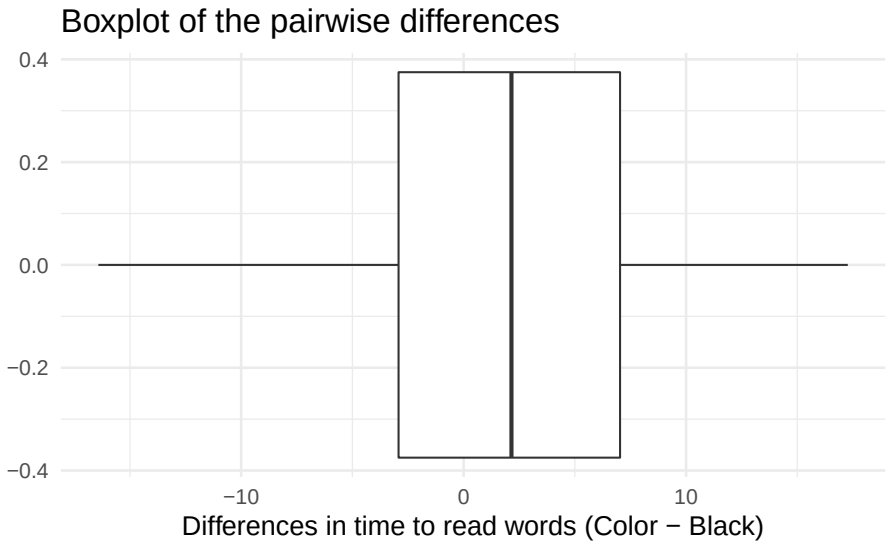
The following code plots each subject's time to read the colored words (above) and time to read the black words (below) connected by a grey line, a histogram of the differences in time to read words between the two conditions, and a boxplot displaying the pairwise differences in time (color − black).

```
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 pairwise differences",
       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

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

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

Use statistical inferential methods to draw inferences from the data

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

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

where the standard error of the sample mean difference is:

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

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

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

Using the provided `R` script file, enter the T-score (for `xx`) into the `pt()` function. For single sample or paired data, degrees of freedom are found by subtracting 1 from the sample size. You should therefore use `df` $= n_d - 1 = 70 - 1 = 69$ and `lower.tail = FALSE` to find the p-value. Highlight and run line 23.

```
pt(xx, df=69, lower.tail=FALSE)
```

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

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

To calculate a theory-based confidence interval for the paired mean difference, use the following formula:

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

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

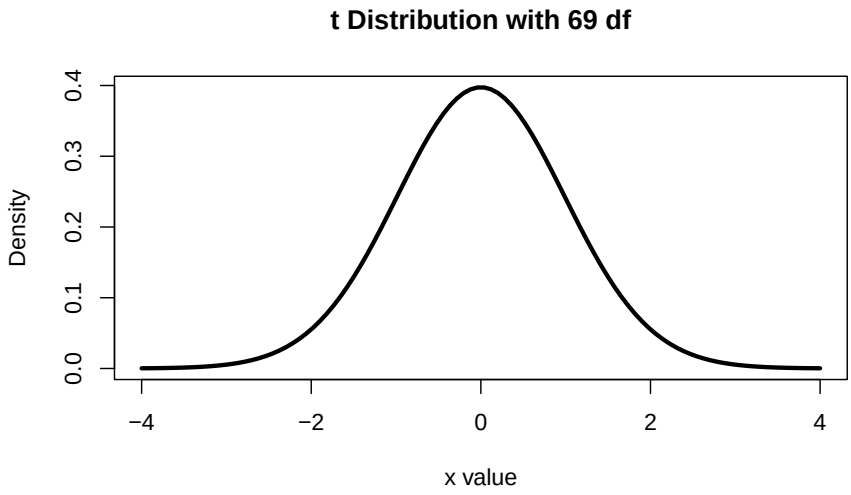


Figure 11.3: t-distribution with 69 degrees of freedom

11. In Figure 11.3, 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.

12. Calculate the margin of error for the true paired mean difference using theory-based methods.

13. Calculate the confidence interval for the true paired mean difference using theory-based methods.

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

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

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

17. The abstract states, that the conflicting color stimuli "caused an increase of 2.3 seconds or 5.6% over the normal time for reading the same words printed in black." Is this statement valid? Explain.

11.3.4 Take-home messages

1. In order to use theory-based methods for dependent groups (paired data), the independent observational units and normality conditions must be met.
2. A T-score is compared to a t-distribution with $n-1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.
3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

11.3.5 Additional notes

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

11.4 Module 11 Lab: Swearing

11.4.1 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.

11.4.2 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.

11.4.3 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.

4. What does μ_d represent in the context of this study?

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

6. 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 11 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 6. Highlight and run lines 1–7 to load the data and create a paired plot of the data.

```
swearing <- datasetname
paired_observed_plot(swearing)
```

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

Enter the outcome for group 1 (`Swear`) for `measurement_1` and the outcome for group 2 (`Neutral`) for `measurement_2` in line 10. Highlight and run lines 9–12 to get the summary statistics for the data.

```
swearing_diff <- swearing %>%
  mutate(differences = measurement_1 - measurement_2)
swearing_diff %>%
    summarise(favstats(differences))
```

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

9. **How far, on average, is each difference in pain tolerance from the mean of the differences in pain tolerance? What is the appropriate notation for this value?**

Use statistical inferential methods to draw inferences from the data

10. 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 means differences and compute a p-value.

11. 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:

12. Simulate a null distribution and compute the p-value. Using the `R` script file for this lab, enter your answers for question 11 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 15–21.

```
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

13. **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?**

14. 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.

15. 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 24–27. **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
```

Sketch the bootstrap distribution created using the code. Report the 98% confidence interval in interval notation.

16. Interpret the *confidence level* of the interval you calculated in question 15.

17. 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
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- Scope of inference

MODULE 12

Inference for a Quantitative Response with Independent Samples

12.1 Module 12 Reading Guide: Inference for a Difference in Two Means

Section 6.3 (Inference for a difference in two means)

Videos

- 6.3

Reminders from previous sections

n_1= sample size of group 1

n_2 = sample size of group 2

$\overline{x}$ = sample mean

s = sample standard deviation

μ = population mean

σ = population standard deviation

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test.'

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\Longrightarrow$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.
- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution.

Central Limit Theorem: For large sample sizes, the sampling distribution of a sample mean (or proportion) will be approximately normal (bell-shaped and symmetric).

t-distribution: A bell-shaped symmetric distribution, centered at 0, wider than the standard normal distribution.

- The variability in a t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The t-distribution gets closer to the standard normal distribution as df increases.

Degrees of freedom (df): describes the variability of the t-distribution.

T-score: the name for a standardized statistic which is compared to a t-distribution.

Notes

To create a **simulated null distribution** of differences in sample means,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of differences in sample means,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the bootstrap distribution?

Conditions to use the CLT for a difference in two means:

Independence:

Checked by:

Normality:

Checked by:

In a two-sample t-test, how are the degrees of freedom determined?

True or false: A large p-value indicates that the null hypothesis is true.

Formulas

$SE(\overline{x_1} - \overline{x_2}) =$

$T =$

Confidence interval for a difference in means:

Notation

μ_1 represents

μ_2 represents

σ_1 represents

σ_2 represents

$\overline{x_1}$ represents

$\overline{x_2}$ represents

s_1 represents

s_2 represents

Example: Test scores

1. What are the observational units?

2. What are the sample statistics presented in this example? What notation would be used to represent each value?

3. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

4. What is the research question?

5. Write the null and alternative hypothesis in appropriate notation.

6. How could we use cards to simulate **one** sample *which assumes the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 shuffles are generated, where is the resulting simulated distribution centered? Why does that make sense?

8. How was the p-value for this test found? The proportion of simulated null samples at _____ or _____.

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

10. From these data, can we conclude the exams are equally difficult?

11. What type of error may have occurred at the 5% significance level? Interpret that error in context.

Example: ESC and heart attacks

1. What is the research question?

2. What are the observational units?

3. What variables are recorded? Give the type (categorical or quantitative) and role (explanatory or response) of each.

4. What are the sample statistics presented in this example? What notation would be used to represent each value?

5. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

6. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

7. After 1000 resamples are generated, where is the resulting bootstrap distribution centered? Why does that make sense?

8. Does the 90% confidence interval provide evidence of a difference across the two treatments?

Example: NC births

1. What is the research question?

2. What are the observational units?

3. What variables will be analyzed? Give the type and role of each.

4. Can the results of this study be generalized to a larger population?

5. Are causal conclusions appropriate for these data?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. What are the sample statistics presented in this example? What notation would be used to represent each value?

9. Are the independence and normality conditions satisfied?

10. Calculate the standard error of the difference in sample means.

11. Calculate the T-score (the standardized statistic for the sample mean).

12. What distribution should the T-score be compared to in order to calculate a p-value?

13. What was the p-value of the test?

14. What conclusion should the researcher make?

15. Calculate a 95% confidence interval for the parameter of interest using `qt(0.975, df = 49) = 1.677` as the $t^\star$ value.

16. Interpret your interval in the context of the problem.

12.2 Activity 12: Weather Patterns and Record Snowfall

12.2.1 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.

12.2.2 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 Section 6.3 in the textbook.

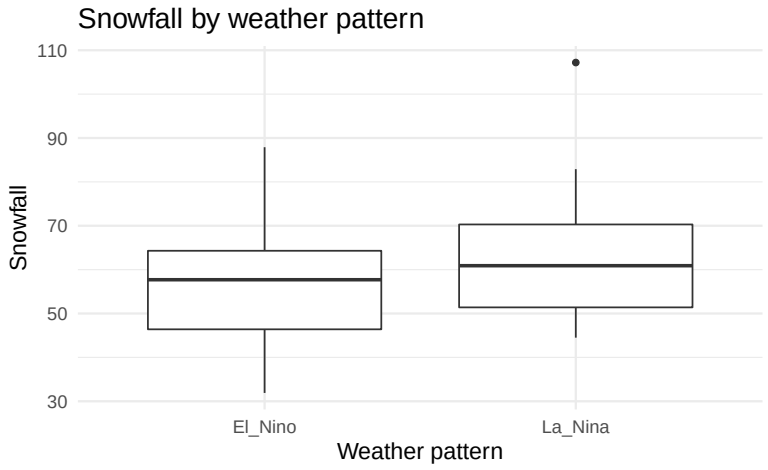
12.2.3 Weather patterns and record snowfall

In the winter of 2018–2019, Bozeman had a record snowfall which resulted in the collapse of two flat-roofed buildings on the MSU campus. A writer for the *Washington Post* predicted the heavy snowfall for 2018–2019 due to the El Niño weather pattern that occurred in that season. A meteorologist in Montana wanted to see if the weather pattern really was associated with total snowfall. She obtained historical data from 44 years on the weather pattern (El Niño or La Niña) and snowfall (in inches) at the Billings Weather Station (National Weather Service Corporate Image Web Team, n.d.). Side-by-side boxplots and summary statistics for each group are shown on the following page.

Notice from the `R` code that the name of the data set is `Snow`.

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

```
# Side-by-side box plots
Snow %>%
ggplot(aes(x = WeatherPattern, y = Snowfall)) +
    geom_boxplot() +
    labs(title = "Snowfall by weather pattern",
         x = "Weather pattern")
```




```
# Summary statistics
Snow %>%
     summarize(favstats(Snowfall ~ WeatherPattern))
```

```
#>   WeatherPattern  min   Q1 median   Q3   max     mean       sd  n missing
#> 1        El_Nino 31.9 46.4   57.7 64.3  87.9 56.23043 13.00823 23       0
#> 2        La_Nina 44.5 51.4   60.9 70.3 107.2 63.13333 15.48626 21       0
```

Quantitative variables review

1. The two variables assessed in this study are the type of weather pattern and snowfall. Identify the role for each variable (explanatory or response).

2. Which group (El Niño or La Niña) has the highest center in the distributions of snowfall? Explain which measure of center you are using.

3. Using the side-by-side box plots, which group has the largest spread in snowfall? 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. Use El Niño minus La Niña as the order of subtraction.

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). Use El Niño minus La Niña as the order of subtraction. 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 snowfall values observed in the data set would have been the same regardless of the weather pattern that year.

To demonstrate this simulation, your instructor will provide cards to for you to use to simulate a sample.

10. How many cards will we start with?

11. What will we write on each card?

12. Next, we will 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 is calculated from the cards and plotted on the null distribution? *Hint*: What statistic are we calculating from the data?

14. Create one simulation using the cards provided. Is your simulated statistic closer to the null value of zero than the difference in means calculated from the sample? Explain why this makes sense.

15. Once we create a null distribution of 1000 simulations, at what value do you expect the distribution to be centered? Explain your reasoning.

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.

16. When using the `two_mean_test()` function, we need to enter the name of the response variable, `Snowfall`, and the name of the explanatory variable, `WeatherPattern`, for the formula. The name of the data set as shown above is `Snow`. 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? `"El_Nino"` or `"La_Nina"`):

- Number of repetitions:

- As extreme as:

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

17. Simulate a null distribution and compute the p-value. 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 1–29.

```
two_mean_test(Snowfall ~ WeatherPattern, #Enter the names of the variables
              data = Snow,  # 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 code above.

18. 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 confidence interval.

19. Using bootstrapping find a 95% confidence interval. Using the provided `R` script file, enter the variable names as in the `two_mean_test()` function, outcome name for the first in subtraction, number of repetitions, and the confidence level as a decimal. Highlight and run lines 32–36. Report the 95% confidence interval in interval notation.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = Snow,  # 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)
```

20. Interpret the interval you calculated in question 19.

21. Would the results from a theory-based test match the results we saw with the simulation? Explain why or why not.

12.2.4 Take-home messages

1. This activity differs from Activities 11a and 11b because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.
2. To create one simulated sample on the null distribution for a difference in sample means, label cards with the response variable values from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2. Calculate and plot the difference in means.
3. To create one simulated sample on the bootstrap distribution for a difference in sample means, label n_1+n_2 cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.

12.2.5 Additional notes

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

12.3 Module 12 Lab: The Triple Crown

12.3.1 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.

12.3.2 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 Section 6.3 in the textbook.

12.3.3 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 for this 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.
  - $n \geq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Upload and open the `R` script file for Week 12 lab. 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 7. Write a title for the boxplots in line 11. Highlight and run lines 1–13 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 17. Highlight and run lines 16–17 to get the summary statistics for the data.

```
hikes %>%
  summarize(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:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)},$$

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 20. Report the p-value calculated.

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

11. **Explain why we multiplied by 2 in the code above.**

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

13. 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:

$$\bar{x}_1 - \bar{x}_2 \pm t^* SE(\bar{x}_1 - \bar{x}_2).$$

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 23 to find the appropriate t^* multiplier

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

14. Report the t^* multiplier for the 95% confidence interval.

15. Calculate the 95% confidence interval using theory-based methods.

16. Interpret the 95% confidence interval in context of the study.

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

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

19. What type of error may be possible?

20. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:

- Summary statistic
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- Scope of inference

MODULE 13

Inference for Two Quantitative Variables

13.1 Module 13 Reading Guide: Inference for Slope and Correlation

Sections 7.1 and 7.2 (Inference for regression and model conditions)

Videos

- 7.1and7.2

Reminders from previous sections

β_0: population y-intercept

β_1: population slope

ρ: population correlation

b_0: sample y-intercept

b_1: sample slope

r: sample correlation

Scatterplot: displays two quantitative variables; one dot = two measurements (x, y) on one observational unit.

Four characteristics of a scatterplot:

- *Form*: pattern of the dots plotted. Is the trend generally linear (you can fit a straight line to the data) or non-linear?
- *Strength*: how closely do the points follow a trend? Very closely (strong)? No pattern (weak)?
- *Direction*: as the x values increase, do the y-values tend to increase (positive) or decrease (negative)?
- Unusual observations or *outliers*: points that do not fit the overall pattern of the data.

Least squares regression line: $\hat{y} = b_0 + b_1 x$, where b_0 is the sample y-intercept (the estimate for the `(Intercept)` row in the `R` regression output), and b_1 is the sample slope (the estimate for the `x-variable_name` row in the `R`).

Sample slope interpretation: a 1 unit increase in the x variable is associated with a $|b_1|$ unit *predicted* increase/decrease in the y-variable.

General steps of a hypothesis test:

1. Frame the research question in terms of hypotheses.
2. Collect and summarize data using a test statistic.
3. Assume the null hypothesis is true, and simulate or mathematically model a null distribution for the test statistic.
4. Compare the observed test statistic to the null distribution to calculate a p-value.
5. Make a conclusion based on the p-value and write the conclusion in context.

Parameter: a value summarizing a variable(s) for a population.

Statistic: a value summarizing a variable(s) for a sample.

Sampling distribution: plot of statistics from 1000s of samples of the same size taken from the same population.

Standard deviation of a statistic: the variability of statistics from 1000s of samples; how far, on average, each statistic is from the true value of the parameter.

Standard error of a statistic: estimated standard deviation of a statistic.

Hypothesis test: a process to determine how strong the evidence of an effect is. Also called a 'significance test.'

Simulation-based method: Simulate lots of samples of size n under assumption of the null hypothesis, then find the proportion of the simulations that are at least as extreme as the observed sample statistic.

Theory-based method: Develop a mathematical model for the sampling distribution of the statistic under the null hypothesis and use the model to calculate the probability of the observed sample statistic (or one more extreme) occurring.

Null hypothesis (H_0): the skeptical perspective; no difference; no change; no effect; random chance; what the researcher hopes to prove is **wrong**.

Alternative hypothesis (H_A): the new perspective; a difference/increase/decrease; an effect; not random chance; what the researcher hopes to prove is **correct**.

Null value: the value of the parameter when we assume the null hypothesis is true (labeled as $parameter_0$).

Null distribution: the simulated or modeled distribution of statistics (sampling distribution) we would expect to occur if the null hypothesis is true.

P-value: probability of seeing the observed sample data, or something more extreme, assuming the null hypothesis is true.

$\Longrightarrow$ Lower the p-value the stronger the evidence AGAINST the null hypothesis and FOR the alternative hypothesis.

Decision: a determination of whether to reject or fail to reject a null hypothesis based on a p-value and a pre-set level of significance.

- If p-value $\leq \alpha$, then reject H_0.
- If p-value $> \alpha$, then fail to reject H_0.

Significance level (α): a threshold used to determine if a p-value provides enough evidence to reject the null hypothesis or not.

Common levels of α include 0.01, 0.05, and 0.10.

Statistically significant: results are considered statistically significant if the p-value is below the significance level.

Confidence interval: a process to determine how large an effect is; a range of plausible values for the parameter. Also called 'estimation.'

Margin of error: the value that is added to and subtracted from the sample statistic to create a confidence interval; half the width of a confidence interval.

Bootstrapping: the process of drawing with replacement n times from the original sample.

Bootstrapped resample: a random sample of size n from the original sample, selected with replacement.

Bootstrapped statistic: the statistic recorded from the bootstrapped resample.

Confidence level: how confident we are that the confidence interval will capture the parameter.

Bootstrap $X\%$ confidence interval: $((\frac{(1-X)}{2})^{th}$ percentile, $(X + (\frac{(1-X)}{2})^{th}$ percentile) of a bootstrap distribution

t-distribution: A bell-shaped symmetric distribution, centered at 0, wider than the standard normal distribution.

- The variability in a t-distribution depends on the sample size (used to calculate degrees of freedom — df for short).
- The t-distribution gets closer to the standard normal distribution as df increases.

Degrees of freedom (df): describes the variability of the t-distribution.

T-score: the name for a standardized statistic which is compared to a t-distribution.

Notes

To create a **simulated null distribution** of sample slopes or sample correlations,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the simulated null distribution?

To create a **bootstrap distribution** of sample slopes or sample correlations,

1. How many cards will you need and how will the cards be labeled?

2. What do you do with the cards after labeling them?

3. After shuffling, what value will be plotted on the bootstrap distribution?

Conditions to use the CLT for testing slope (or correlation):

Linearity:

Checked by:

Independent observations:

Checked by:

Nearly normal residuals:

Checked by:

Constant or equal variance:

Checked by:

In a theory-based test of slope or correlation, how are the degrees of freedom determined?

Explain why testing for slope is equivalent to testing for correlation.

Where in the `R` output can $SE(b_1)$ be found?

Formulas

$T =$

Confidence interval:

Example: Crop yields

1. What are the observational units?

2. What is the parameter representing in the context of this problem? What notation would be used to represent this parameter?

3. What is the research question?

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

5. How could we use cards to simulate **one** sample which assumes *the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

6. After 1000 shuffles are generated, where is the resulting simulated distribution centered? Why does that make sense?

7. What are the sample statistics presented in this example? What notation would be used to represent each value?

8. Write the least squares regression line for these data in appropriate notation.

9. How was the p-value for this test found? The proportion of simulated null samples at _____ or _____.

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

11. What conclusion can be drawn from these data?

12. How could we use cards to simulate **one** bootstrap resample *which does not assume the null hypothesis is true*? How many cards? What is written on the cards? What would we do with the cards? What would you record once you have a simulated sample?

13. Interpret the 95% confidence interval provided.

Example: Midterm elections and unemployment

1. What is the research question?

2. What are the observational units?

3. What variables will be analyzed? Give the type and role of each.

4. Can the results of this study be generalized to a larger population?

5. Are causal conclusions appropriate for these data?

6. Write the null and the alternative hypotheses in words.

7. Write the null and the alternative hypotheses in notation.

8. What are the sample statistics presented in this example? What notation would be used to represent each value?

9. Write the least squares regression line for these data in appropriate notation.

10. From the `R` output, what is the standard error of the slope estimate?

11. Calculate the T-score (the standardized statistic for the slope).

12. What distribution should the T-score be compared to in order to calculate a p-value?

13. What was the p-value of the test?

14. What conclusion should the researcher make?

15. Calculate a 95% confidence interval for the parameter of interest using `qt(0.975, df = 27) = 2.052` as the $t^\star$ value.

16. Interpret your interval in the context of the problem.

13.2 Activity 13A: Diving Penguins

13.2.1 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.

13.2.2 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 Chapters 3 and 7 in the textbook.

13.2.3 Diving Penguins

Emperor penguins are the most accomplished divers among birds, making routine dives of 5–12 minutes, with the longest recorded dive over 27 minutes. These birds can also dive to depths of over 500 meters! Since air-breathing animals like penguins must hold their breath while submerged, the duration of any given dive depends on how much oxygen is in the bird's body at the beginning of the dive, how quickly that oxygen gets used, and the lowest level of oxygen the bird can tolerate. The rate of oxygen depletion is primarily determined by the penguin's heart rate. Consequently, studies of heart rates during dives can help us understand how these animals regulate their oxygen consumption in order to make such impressive dives. The researchers equipped emperor penguins with devices that record their heart rates during dives. The data set reports Dive Heart Rate (beats per minute), the Duration (minutes) of dives, and other related variables. Can the dive heart rate be used to predict the duration of the dive for Emperor Penguins?

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

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.

Use the provided `R` script file to create a scatterplot to examine the relationship between the diving heart rate and duration of the dive by filling in the variable names (`Dive_HeartRate` and `Duration`) for `explanatory` and `response` in line 9. Highlight and run lines 1–15.

```
diving %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Heart Rate (bpm)",  # Label x-axis
       y = "Dive Duration (min)",  # Label y-axis
       title = "Scatterplot of Emperor Penguins Diving Heart Rate vs. Dive Duration") +
              # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

2. Sketch the plot created below. Based on your plot, does it appear that there is a relationship between dive heart rate and duration of the dive? Note: `Dive_HeartRate` should be on the x-axis.

3. 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

4. Write out the null hypothesis in words to test slope.

5. Using the research question, write the alternative hypothesis in notation using slope as the summary measure.

Summarize and visualize the data

Using the provided `R` script file, enter the response variable name, `Duration`, into the `lm()` (linear model) function for `response` and the explanatory variable name, `Dive_HeartRate`, for `explanatory` in line 18 to get the linear model output and value for the correlation coefficient. Highlight and run lines 18–19.

```
lm.diving <- lm(response~explanatory, data=diving) # lm(response~explanatory)
round(summary(lm.diving)$coefficients, 5)
cor(diving$Duration, diving$Dive_HeartRate)
```

6. 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.

7. Interpret the estimated slope in context of the problem.

8. Report the value of correlation between the diving heart rate and the duration of the dive.

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, Duration, 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 `diving`), 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 `Duration` and the explanatory variable name is `Dive_HeartRate` for these data.

9. What inputs should be entered for each of the following to create the simulation to test regression slope?

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

- Summary measure (choose `"slope"` or `"correlation"`):

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

- Number of repetitions:

Using the `R` script file for this activity, enter your answers for question 9 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 23–28.

```
regression_test(Duration ~ Dive_Heartrate, # response ~ explanatory
                data = diving, # Name of data set
                direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
                summary_measure = "xx", # "slope" or "correlation"
                as_extreme_as = x, # Observed slope or correlation
                number_repetitions = 1000) # Number of simulated samples for null distribution
```

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

11. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which two inputs in #8 would need to be changed to test for correlation? What inputs should you use instead?

12. Change the inputs in lines 23–28 to test for correlation instead of slope. Highlight and run those lines, then report the new p-value of the test.

13. The p-values from the test of slope (#10) and the test of correlation (#12) should be similar. Explain why the two p-values should match. *Hint: think about the relationship between slope and correlation!*

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in `R` (in the `catstats` package) to simulate the bootstrap distribution of sample slopes (or sample correlations) and calculate a confidence interval. Fill in the `xx`'s in the the provided `R` script file to find a 95% confidence interval for slope. Highlight and run lines 31–35.

```
regression_bootstrap_CI(Duration ~ Dive_Heartrate, # response ~ explanatory
   data = diving, # 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
```

14. Report the bootstrap 95% confidence interval in interval notation.

15. Interpret the interval in question 14 in context of the problem. *Hint: use the interpretation of slope in your confidence interval interpretation.*

Communicate the results and answer the research question

16. Based on the p-value, write a conclusion in context of the problem.

17. 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?

13.2.4 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.

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 13B: Golf Driving Distance

13.3.1 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.

13.3.2 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 Chapters 3 and 7 in the textbook.

13.3.3 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 was collecting on 354 PGA and LGPA 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 `xx` and `yy` in line 9. Highlight and run lines 1–15.

```
golf %>% # Pipe data set into...
ggplot(aes(x = xx, y = yy))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Driving Distance",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       title = "Scatterplot of Driving Distance by Percent Accuracy") +
              # 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.

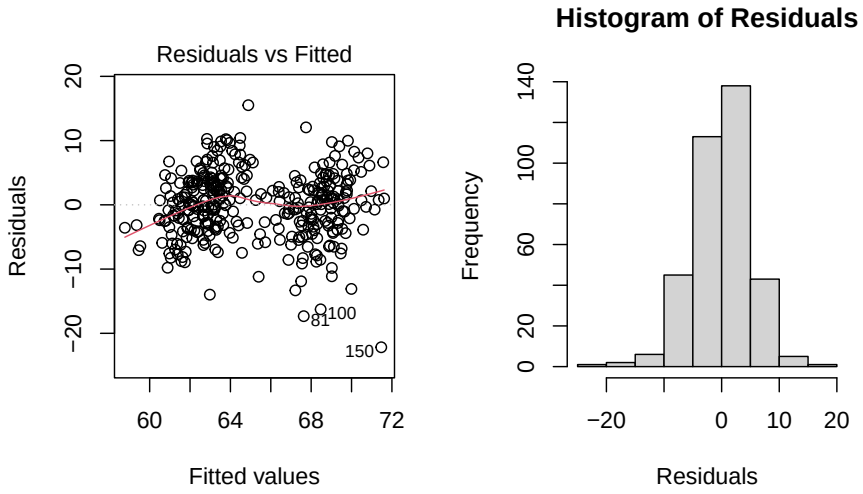
Conditions for the least squares line

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 26 to get the linear model output. Highlight and run lines 26–27.

```
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,

$$T = \frac{\text{slope estimate}}{SE} = \frac{b_1}{SE(b_1)}.$$

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 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.

$$\text{point estimate} \pm t^* SE(\text{estimate}).$$

When the point estimate is a regression slope, this formula becomes

$$b_1 \pm t^* SE(b_1).$$

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) # 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.

Multivariate plots

Another variable that may affect the percent accuracy is the sex of the golfer. We will look at how this variable may change the relationship between driving distance and percent accuracy. Highlight and run lines 33–39 to produce the multivariate plot.

```
golf %>%
  ggplot(aes(x = Driving_Distance, y = Percent_Accuracy, color=Sex))+  # Specify variables
  geom_point(aes(shape = Sex), size = 3) +  # Add scatterplot of points
  labs(x = "Driving Distance (m)",  # Label x-axis
       y = "Percent Accuracy",  # Label y-axis
       color = "Sex", shape = "Sex",
       # Be sure to title your plots
       title = "Scatterplot of Golf Driving Distance and Percent Accuracy by Sex") +
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```

15. Does the association between driving distance and percent accuracy change dependent on sex of the golfer? Explain your answer.

16. Explain the association between sex and each of the other two variables.

13.3.4 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.

13.3.5 Additional notes

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

13.4 Module 13 Lab: COVID Immunization and Infection Rates

13.4.1 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.

13.4.2 COVID immunization and infection rates

According to the *Washington Post* "States with higher vaccination rates now have markedly fewer coronavirus cases, as infections are dropping in places where most residents have been immunized and are rising in many places people have not." (Keating et al. 2021) In this article they found that there are differences in infection rates for different counties within a specific state. To check this claim, a random sample of 125 counties from different states was assessed. Vaccination rates and number of cases per 100,000 residents were found for each county. Researchers want to assess if counties with a high vaccination rate tend to have lower coronavirus case rates.

Upload and open the `R` script file for Week 13 lab. Upload and import the csv file, `covid_vaccinations`. 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 1–7 to load the data.

```
# Read in data set and remove NAs
covid <- datasetname
```

Summarize and visualize the data

To find the correlation between the variables, `PercentImmunized` and `Case_per_100K` highlight and run lines 10–13 in the `R` script file.

```
covid %>%
  select(c("PercentImmunized", "Case_per_100K")) %>%
  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 `PercentImmunized` and `Case_per_100K`.**

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 percent immunized and cases per 100,000. Enter the variable `Case_per_100K` for `response` and the variable `PercentImmunized` for `explanatory` in line 17. Highlight and run lines 17–18 to get the linear model output.

```
# Fit linear model: y ~ x
covidLM <- lm(response~explanatory, data=covid)
summary(covidLM)$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

Highlight and run lines 21–34 to produce the diagnostic plots needed to assess conditions to use theory-based methods. Use the scatterplot and the residual plots to assess the validity conditions for approximating the data with the t-distribution.

```
#Scatterplot
covid %>% # Pipe data set into...
  ggplot(aes(x = PercentImmunized, y = Case_per_100K))+  # Specify variables
  geom_point() +  # Add scatterplot of points
  labs(x = "Percent Immunized",  # Label x-axis
       y = "Number of cases per 100K",  # Label y-axis
       # Be sure to tile your plots
       title = "Scatterplot of Percent Immunized vs. Infection Rate of COVID in US Counties") +
  geom_smooth(method = "lm", se = FALSE)  # Add regression line

#Diagnostic plots
covidLM <- lm(Case_per_100K~PercentImmunized, data = covid) # Fit linear regression model
par(mfrow=c(1,2)) # Set graphics parameters to plot 2 plots in 1 row
plot(covidLM, which=1) # Residual vs fitted values
hist(covidLM$resid, xlab="Residuals", ylab="Frequency",
     main = "Histogram of Residuals") # Histogram of residuals
```

5. **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

6. Write out the null and alternative hypotheses in notation to test *correlation* between the percent immunized in US counties and the infection rate.

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 `covid`), 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 `Case_per_100K` and the explanatory variable name is `PercentImmunized`.

7. 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 7 in place of the `xx`'s to produce the null distribution with 1000 simulations. Highlight and run lines 37–42. **Upload a copy of your plot showing the p-value to Gradescope for your group.**

```
regression_test(Case_per_100K~PercentImmunized, # response ~ explanatory
                data = covid, # 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
```

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

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

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 45–49.

```
regression_bootstrap_CI(Case_per_100K~PercentImmunized, # response ~ explanatory
   data = covid, # 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
```

10. Report the bootstrap 90% confidence interval in interval notation.

11. Interpret the 90% confidence interval in context of the problem.

Communicate the results and answer the research question

12. Based on the p-value, write a conclusion in context of the problem.

13. Using a significance level of 0.1, what decision would you make?

14. What type of error is possible?

15. Interpret this error in context of the problem.

16. 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
- P-value and interpretation
- Confidence interval and interpretation
- Conclusion (written to answer the research question)
- Scope of inference

MODULE 14

Probability and Relative Risk

14.1 Module 14 Reading Guide: Special Topics

Section 2.2 (Probability with tables)

Videos

- 2.2

Vocabulary

Random process:

Probability:

Hypothetical two-way table:

Unconditional probability:

Notation:

Conditional probability:

Notation:

Event:

Notation:

Complement:

Notation:

Sensitivity:

Specificity:

Prevalence:

Notes

Method for creating a hypothetical two-way table:

1. Start with

2. Fill in the column or row totals using

3. Fill in the interior cells using

4. Add/Subtract to fill in the row/column totals not filled in at step 2.

   To find unconditional probabilities from the table,

   To find conditional probabilities from the table,

Example: Baby Jeff

1. Let D be the event a child has CPK. What does D^C represent?

2. Let T be the event a child tests positive for CPK. What does T^C represent?

3. Write each of the following values in proper probability notation:

   a. $1/10000 = 0.0001 = P($ $)$
   b. $100\% = 1.0 = P($ $)$
   c. $99.98\% = 0.9998 = P($ $)$

4. Write out the steps for creating the hypothetical two-way table in section 2.2.4 of your textbook, then copy the table below.

   First,

   Next,

   After that,

Finally,

Hypothetical two-way table:

	Test Positive	Test Negative	Total
Has CPK			
Does not have CPK			
Total			100,000

5. What is the probability that a child who had a positive test result actually does have CPK? What probability notation should be used for this value?

6. Explain how the probability in #5 was calculated.

Section 5.5 revisited (Simulation-based inference for a relative risk)

Vocabulary

Relative risk:

Notes

Interpreting relative risk ($RR = \frac{\hat{p_1}}{\hat{p_2}}$)

The proportion of success in group 1 is _____ times the proportion of success in group 2.

The proportion of success in group 1 is _____ % higher/lower than in group 2.

Write the null hypothesis in notation for a test of relative risk.

Formulas

Relative risk =

Example: CPR and blood thinner

1. What is the sample relative risk? Interpret the value in the context of the study.

14.2 Activity 14A: What's the probability?

14.2.1 Learning outcomes

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

14.2.2 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 Sections 2.1 and 2.2 in the textbook.

14.2.3 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.

	Senior	Not a Senior	Total
Science	12,000	48,000	60,000
Liberal Arts	12,000	28,000	40,000
Total	24,000	76,000	100,000

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.

14.2.4 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.

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 14B: Titanic Survivors — Relative Risk

14.3.1 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.

14.3.2 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 5 in your textbook.

14.3.3 Percent increase or percent decrease?

1. Last season's skis are 30% off original sale price at REI. You want to buy a pair of skis that were originally \$100. How much will you pay?

2. What about a pair of skis that were originally \$593 at REI?

3. The same pair of skis are selling for \$650 at Chalet Sports. What percent higher is this price compared to the \$593 at REI?

4. You're on vacation in Spokane and decide to buy a \$450 pair of skis. The sales tax is 6.5%. How much do you pay in total?

14.3.4 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

5. 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			

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

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

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

9. 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:

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

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

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

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

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

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

14.3.5 Risk in the News

15. 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.

14.3.6 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.

14.3.7 Additional notes

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

14.4 Module 14 Lab: Efficacy of the COVID Vaccination

14.4.1 Learning outcomes

- Recognize and simulate probabilities as long-run frequencies.
- Use two-way tables to calculate conditional probabilities.
- Interpret the value of relative risk in terms of a percent increase or decrease.
- Evaluate the association between two categorical variables using relative risk.

14.4.2 Efficacy of the COVID vaccination

In November 2021, it was estimated that 59.1% of all US adults ($\geq$ 18 years old) were fully vaccinated against COVID-19 ("US COVID-19 Vaccine Tracker: See Your State's Progress" 2021). While vaccination is not 100% effective at protection against COVID-19, there are also other benefits to the vaccine. What impact does vaccination have on hospitalization rates for COVID? The following hypothetical two-way table was created based on CDC data on adult hospitalizations for COVID in the US ("Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status" 2021) in the same time period.

Let A = the event the US adult is vaccinated, and B = the event the US adult is hospitalized with COVID.

	Vaccinated	Not Vaccinated	Total
Hospitalized with COVID	2.3049	27.7302	30.0351
Not hospitalized with COVID	59,097.6951	40,872.2698	99,969.9649
Total	59,100	40,900	100,000

1. What is the probability that a US adult is both hospitalized with COVID-19 and vaccinated? Use proper probability notation.

2. **What is the probability that a US adult hospitalized with a COVID infection is vaccinated? Use proper probability notation.**

3. What is the probability that a US adult is hospitalized with a COVID infection in November 2021? Use proper probability notation.

4. **Give the probability notation for the calculation $\frac{27.7302}{30.0392} = 0.923$. Write out what this probability measures in words.**

5. What is the probability that a vaccinated US adult is hospitalized with COVID?

6. What is the probability that a un-vaccinated US adult is hospitalized with COVID?

7. **Calculate the relative risk for hospitalization with COVID in November 2021 for US adults fully vaccinated compared US adults not vaccinated.**

8. Calculate the percent increase (or decrease) in hospitalization rate for US adults fully vaccinated compared to US adults not vaccinated.

9. **Interpret the relative risk as a percent increase/decrease in context of the problem.**

10. **Does it appear that there is an association with the risk of hospitalization due to COVID-19 and vaccination status? Explain.**

11. **Explain why a hypothesis test would not be appropriate in this case.**

MODULE 15

Semester Review

15.1 Final Exam Review

Use the provided data set from the Islands (ExamReviewData.csv) and the Exam 2 Review `R` script file to answer the following questions. Each adult (>21) islander was selected at random from all the adult islanders. Variables and their descriptions are listed below. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measure before listening to the music and then 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
`Gender`	Gender of Islander (M = male, F = Female)
`Age`	Age of Islander
`Married`	Marital status of Islander
`Smoking_Status`	Whether the Islander is a current smoker
`Children`	Whether the Islander has children
`weight_kg`	Weight measured in kg
`height_cm`	Height measured in cm
`respitory_rate`	Breaths per minute
`Type_of_Music`	Music type (Classical or Heavy Medal) Islander was randomly assigned to listen to
`Before_PuzzleCube`	Time to complete puzzle cube (minutes) before listening to assigned music
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed (note: missing data depicted by missing)
`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 provided Final Exam Review `R` script file and analyze the following research question, "Does drinking a caffeinated drink increase blood glucose levels, on average?" Use before – after as the order of subtraction.

   Parameter of Interest:

   Null Hypothesis:

   Notation:

   Words:

   Alternative Hypothesis:

   Notation:

   Words:

   Value of Statistic with Notation:

   Conditions:

   Independence:

   Normality:

   Simulation P-value:

   Interpretation:

   Conclusion:

Decision:

Simulation Confidence Interval:

Interpretation:

Standardized Statistic:

Interpretation:

Theory-based p-value:

Theory-based Confidence Interval:

Does the theory-based p-value and CI match those found using simulation methods?

What is the scope of inference for this study?

2. Use the provided Final Exam Review `R` script file and 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.

Parameter of Interest:

Null Hypothesis:

Notation:

Words:

Alternative Hypothesis:

Notation:

Words:

Value of Statistic with Notation:

Conditions:

Independence:

Normality:

Simulation P-value:

Interpretation:

Conclusion:

Decision:

Simulation Confidence Interval:

Interpretation:

Standardized Statistic:

Interpretation:

Theory-based p-value:

Theory-based Confidence Interval:

Does the theory-based p-value and CI match those found using simulation methods?

What is the scope of inference for this study?

3. Use the provided Final Exam Review `R` script file and analyze the following research question: "Is there an association between height and balance time for Islanders?"

Parameter of Interest:

Null Hypothesis:

Notation:

Words:

Alternative Hypothesis:

Notation:

Words:

Value of Statistic with Notation:

Conditions:

Independence:

Linearity:

Constant Variance:

Normality of Residuals:

Simulation P-value:

Interpretation:

Conclusion:

Decision:

Simulation Confidence Interval:

Interpretation:

Standardized Statistic:

Interpretation:

Theory-based p-value:

Theory-based Confidence Interval:

Does the theory-based p-value and CI match those found using simulation methods?

What is the scope of inference for this study?

15.2 Golden Ticket to Descriptive and Inferential Statistical Methods

In this course, we have covered descriptive (summary statistics and plots) and inferential (hypothesis tests and confidence intervals) methods for five different scenarios:

- one categorical response variable
- two categorical variables
- one quantitative response variable or paired differences in a quantitative variable
- two quantitative variables
- one quantitative response variable and one categorical explanatory variable

The "golden ticket" shown on the next page presents a visual summary of the similarities and differences across these five scenarios.

Scenario	One Categorical Response	Two Categorical Variables	One Quantitative Response OR Paired Differences	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 or Mean Difference	Slope or correlation	Difference in means
Parameter notation	π	$\pi_1 - \pi_2$	μ or μ_d	β_1 or ρ	$\mu_1 - \mu_2$
Statistic notation	$\hat{p}$	$\widehat{p_1} - \widehat{p_2}$	$\bar{x}$ or $\bar{x}_d$	b_1 or r	$\bar{x}_1 - \bar{x}_2$
Null hypothesis	$H_0: \pi = \pi_0$	$H_0: \pi_1 - \pi_2 = 0$	$H_0: \mu = \mu_0$ or $H_0: \mu_d = 0$	$H_0: \beta_1 = 0$ 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 case, 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 $(\mu_o - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Plot sample mean. Repeat 1000's of times. Centered at μ_0 (single mean) or 0 (paired mean difference).	Hold the x values constant; shuffle y's to new 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. Repeat 1000's of times. Centered at $\bar{x}$ or $\bar{x}_d$.	Label n cards with the original (explanatory, response) 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	Independent cases; Number of expected successes and number of expected failures both at least 10. $\pi_o * n \geq 10, (1 - \pi_o) * n$	Independence (within and between groups); Number of expected successes and number of expected failures in each group is at least 10. $\hat{p}_{pool} * (n_1) \geq 10, (1 - \hat{p}_{pool}) * (n_1)$ $\hat{p}_{pool} * (n_2) \geq 10, (1 - \hat{p}_{pool}) * (n_2)$	Independent cases; $n < 30$ with no clear outliers OR $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$	Independent cases; Linear form; 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 $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$
Theory-based standardized statistic (test statistic)	$z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})}$ $SE_0(\hat{p}) = \sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}$	$z = \frac{\widehat{p_1} - \widehat{p_2}}{SE_0(\widehat{p_1} - \widehat{p_2})}$ $SE_0(\widehat{p_1} - \widehat{p_2}) = \sqrt{\widehat{p_{pool}} \times (1 - \widehat{p_{pool}}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, where $\widehat{p_{pool}} = \frac{total\ successes}{total\ sample\ size} = \frac{n_1 \times \widehat{p_1} + n_2 \times \widehat{p_2}}{n_1 + n_2}$	$t = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$ $SE(\bar{x}) = \frac{s}{\sqrt{n}}$	$t = \frac{b_1}{SE(b_1)}$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.	$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Conditions for theory-based 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 $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$	Independent cases; Linear form; 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 $30 \leq n < 100$ with no extreme outliers OR $n \geq 100$
Theory-based confidence interval	$\hat{p} \pm z^* \times SE(\hat{p})$ $SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$	$\widehat{p_1} - \widehat{p_2} \pm z^* \times SE(\widehat{p_1} - \widehat{p_2})$ $SE(\widehat{p_1} - \widehat{p_2}) = \sqrt{\frac{\widehat{p_1} \times (1 - \widehat{p_1})}{n_1} + \frac{\widehat{p_2} \times (1 - \widehat{p_2})}{n_2}}$	$\bar{x} \pm t^* \times SE(\bar{x})$ $SE(\bar{x}) = \frac{s}{\sqrt{n}}$	$b_1 \pm t^* \times SE(b_1)$ $SE(b_1)$ 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/%20and%20https://www.lpga.com/.

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

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.

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

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%20from%20https://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/.

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.

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.

"Welcome to the Navajo Nation Government: Official Site of the Navajo Nation." 2011.%20Retrieved%20from%20https://www.navajo-nsn.gov/.