STAT 216 Coursepack

Fall 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

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.

0.1 Activity 1: Intro to Data

0.1.1 Learning outcomes

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

0.1.2 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
- Types of sampling bias

For more on these concepts, read Chapter 1 and Section 2.1 in the textbook.

0.1.3 General information on labs

On Friday of each week you will complete a lab. Questions are selected from each lab to be turned in on Gradescope. The questions to be submitted on Gradescope are bolded in the lab. As you work through the lab have the Gradescope lab assignment open so that you can answer those questions as you go. Today's activity is Lab 0 in Gradescope for practice submitting as a group.

Steps of the statistical investigation process

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

1. Ask a research question.
2. Design a study and collect data.
3. Summarize and visualize the data. *Weeks 3–4*
4. Use statistical analysis methods to draw inferences from the data. *Weeks 6–13*
5. Communicate the results and answer the research question. *Weeks 6–13*
6. Revisit and look forward.

Today we will focus on the first two steps.

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

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

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

1. **What are the observational units or cases for today's study?**

2. 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.

One person from each group at each table, open the Google sheet linked in D2L and fill in the responses for the following questions for each group member. When creating a data set for use in R it is important to use single words or an underscore between words. Each outcome must be written the same way each time. Make sure to use all lowercase letters to create this data set to have consistency between responses. Do not give units of measure with the numerical values for the length of forearm. For `Residency` use in_state or out_state as the two outcomes.

- Major: what is your declared major?
- Residency: do you have in-state or out-of-state residency?
- Forearm_Length: what is the length of your forearm in inches from the end of your elbow to the end of your index finger?
- Num_Credits: how many credits are you taking this semester?

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

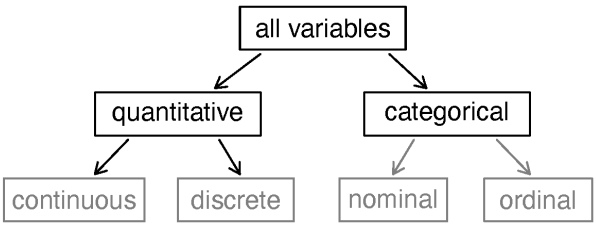


Figure 1: Types of variables.

3. For each column of data, fill in the following table to write out the variable we are collecting on each observational unit in this study and the type of each variable.

Column	Variable	Type of Variable
Major		
Residency		
Forearm Length		
Num Credits		

In the next few weeks we will look at how to summarize data both numerically and graphically. For now we will focus on sampling methods and the type of sampling bias that may be present.

- Sampling bias: a part of the target population is not included or underrepresented in the sample
- Non-response or non-participation bias: part of the already selected population does not respond or chooses not to participate
- Response bias: survey participant gives an untruthful or misleading response

To help determine the type of bias present, it is helpful to think about the observational units, the sample and the target population represented by the problem. The **target population** is the group of cases that makes up the population the researcher is interested in. If sampling bias is present, than the sample taken will not be representative of the actual target population. In these next questions, identify the target population, the sample selected, the variable collected and its type (categorical or quantitative), and the type of bias present.

4. **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.**

   Observational units or cases:

   Target population:

   Sample size:

   Sample taken:

   Variable:

   Type of Variable: categorical quantitative

   Type(s) of bias:

5. 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.

   Observational units or cases:

   Target population:

   Sample size:

   Sample taken:

   Variable:

   Type of Variable: categorical quantitative

   Type(s) of bias:

6. 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?"

   Observational units or cases:

   Target population:

   Sample size:

   Sample taken:

   Variable:

   Type of Variable: categorical quantitative

   Type(s) of bias:

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

0.1.4 Take-home messages

1. When creating a data set, each row will represent a single observational unit or case. Each column represents a variable collected. It is important to write each variable as a single word or use an underscore between words.
2. There are two types of variables: categorical (groups) and quantitative (numerical measures).
3. There are three types of bias to be aware of when designing a sampling method: selection bias, non-response bias, and response bias.

0.1.5 Additional notes

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

0.2 OUt of Class Activity 2: American Indian Address

0.2.1 Learning outcomes

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

0.2.2 Terminology review

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

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

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

0.2.3 American Indian Address

Complete questions 1 - 4 in class. Your instructor will create and post the distribution needed for the remainder of the activity by 8 pm on Monday night. For this activity, you will read a speech given by Jim Becenti, a member of the Navajo American Indian tribe, who spoke about the employment problems his people faced at an Office of Indian Affairs meeting in Phoenix, Arizona, on January 30, 1947 (Moquin and Van Doren 1973). His speech is below:

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

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

By eye selection

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

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

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

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

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

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

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

   Where does the value of 3.95 fall in the plot created in question 4? Near the center of the distribution? In the tails of the distribution?

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

8. Using the graph created in question 4, estimate the proportion of students' computed sample means that were lower than the true mean of 3.95 letters?

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

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

11. 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?

0.2.4 Take-home messages

1. When we use a biased method of selection, we will over or underestimate the parameter.
2. To see if a method is biased, we compare the distribution of the estimates to the true value. We want our estimate to be on target or unbiased. When using unbiased methods of selection, the mean of the distribution matches or is very similar to our true parameter.
3. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.

0.2.5 Additional notes

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

0.3 Activity 2: American Indian Address (continued)

0.3.1 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.

0.3.2 Terminology review

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

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

To review these concepts, see Section 2.1 in the textbook.

Random selection

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

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

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

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

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

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

“

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

3. 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.

4. Where does the value 3.95, the true mean word length, fall in the distribution created in question 3? Near the center of the distribution? In the tails of the distribution?

5. How does the plot generated in question 3 compare to the plot generated in question 4 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 Samples".

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

7. What is the center value of the distribution created in question 6?

8. 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".

9. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from question 6.

Effect of sample size

We will now consider the impact of sample size.

10. First, consider if each student had selected 20 words, instead of 10, by eye. Do you think this would make the plot from question 4 in Activity 2A centered on 3.95 (the true mean word length)? Explain your answer.

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

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

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

12. Compare the distribution created in question 11 to the one created in question 6.

    Which features are similar?

    Which features differ?

13. Compare the spreads of the plots in question 11 and in question 6. You should see that in one plot all sample means are closer to the population mean than in the other. Which plot shows this?

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

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

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

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

0.3.3 Take-home messages

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

0.3.4 Additional notes

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

0.4 Week 2 Lab: Study Design

0.4.1 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.

0.4.2 Terminology review

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

- Scope of inference
- Explanatory variable
- Response variable
- Confounding variable
- Experiment
- Observational study

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

0.4.3 General information labs

Remember that each Friday 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.

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. In today's activity we will discover the purpose behind 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:

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:

Self-identified males: Paul, Antonio, Davieon, Chao, Aryan, Jabari, Tong, Andres, John, Liu, Lucas, Rashidi, Shiwoo, Jihoon, Alejandro, Daniel

Self-identified females: 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 females 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 females 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 females and 8 males to each group. Using this strategy, what proportion of males is in each group?

9. **Assume the scientists used the strategy in question 8, but they put the four tallest females 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)?

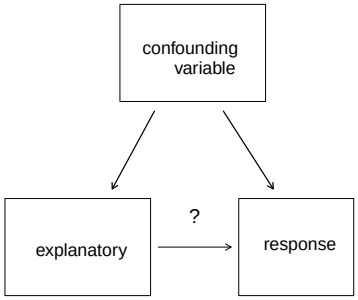
11. Notice the difference in the two proportions is shown as a dot in the plot at the bottom of the web page. Un-check the box for Animate above "Randomize" and click "Randomize" again. Did you get the same difference in proportion of males between the placebo and drug groups?

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

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.

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



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

16. 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.

17. **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 16.**

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

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

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

0.4.4 Take-home messages

1. The study design (observational study vs, experiment) determines if we can draw causal inferences or not. If an association is detected, a randomized experiment allows us to conclude that there is a causal (cause-and-effect) relationship between the explanatory and response variable. Observational studies have potential confounding variables within the study that prevent us from inferring a causal relationship between the variables studied.
2. Confounding variables are variables not included in the study that are related to both the explanatory and the response variables. When there are potential confounding variables in the study we cannot draw causal inferences.
3. Random assignment balances confounding variables across treatment groups. This eliminates any possible confounding variables by breaking the connections between the explanatory variable and the potential confounding variables.
4. Observational studies will always carry the possibility of confounding variables. Randomized experiments, which use random assignment, will have no confounding variables.

0.4.5 Additional notes

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

0.5 Out of Class Activity 4: Movie Profits — Correlation and Coefficient of Determination

0.5.1 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.

0.5.2 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.

0.5.3 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

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. Upload and open the Movie Profits Out of Class Activity F22 Code `R` script file. Highlight and run lines 1–12. Highlight and run lines 1–12 to produce the same table as below.

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

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

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

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

In Figure 2, we see the data plotted with a horizontal line. Note that the line has a slope of zero, this shows no relationship between budget and revenue.

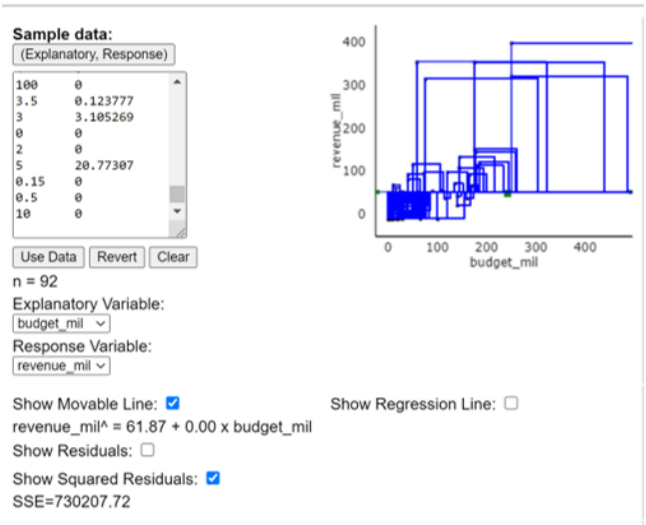


Figure 2: Plot of the data with no slope.

8. Write down the value of SSE given in this image. Since this is the sum of squared errors (SSE) for the horizontal line we call this the total sum of squares (SST).

In Figure 3, we see the data plotted with the regression line (we will learn more about the regression line in the next class). This is the line of best fit between budget and revenue.

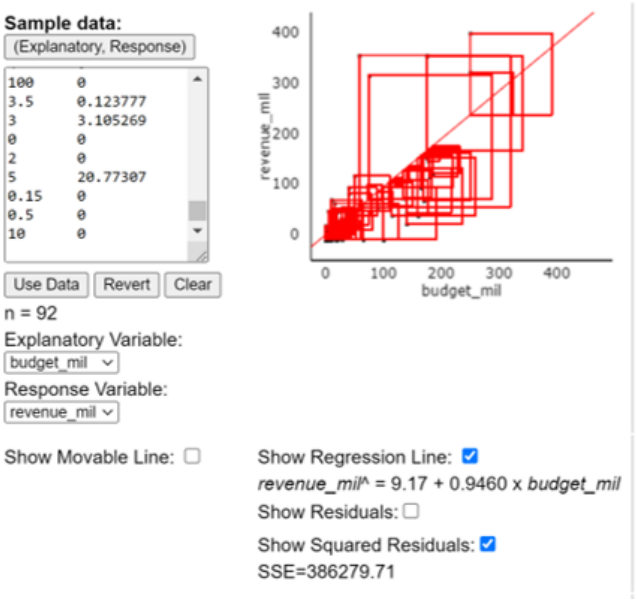


Figure 3: Plot of the data showing the regression line.

9. Write down the value for SSE from this image.

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

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

0.5.4 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.

0.5.5 Additional notes

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

0.6 Activity 4: Movie Profits — Linear Regression

0.6.1 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.

0.6.2 Terminology review

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

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

To review these concepts, see Chapter 6 & 7 in the textbook.

0.6.3 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
```

Vocabulary review

To look at the relationship between two quantitative variables we will create a scatterplot with the explanatory variable on the x-axis and the response variable on the y-axis. We can also find three summary measures for the linear relationship between the two variables: regression slope, correlation and the coefficient of determination.

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).) Upload and open the Movie Profits Activity 4 F22 Code `R` script file. 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
```

1. Sketch the scatterplot created from the code.

2. 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

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

Slope

The linear model function in `R` (`lm()`) gives us the summary for the least squares regression line. The estimate for `(Intercept)` is the y-intercept for the line of least squares, and the estimate for `budget_mil` (the x-variable name) is the value of b_1, the slope. Run lines 16 – 19 in the `R` script file to reproduce the linear model output found in the coursepack.

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

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

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

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

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

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

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

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

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

8. The prediction in question 7 is an example of what?

Residuals

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the variability in the response that hasn't been modeled by the line (model).

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

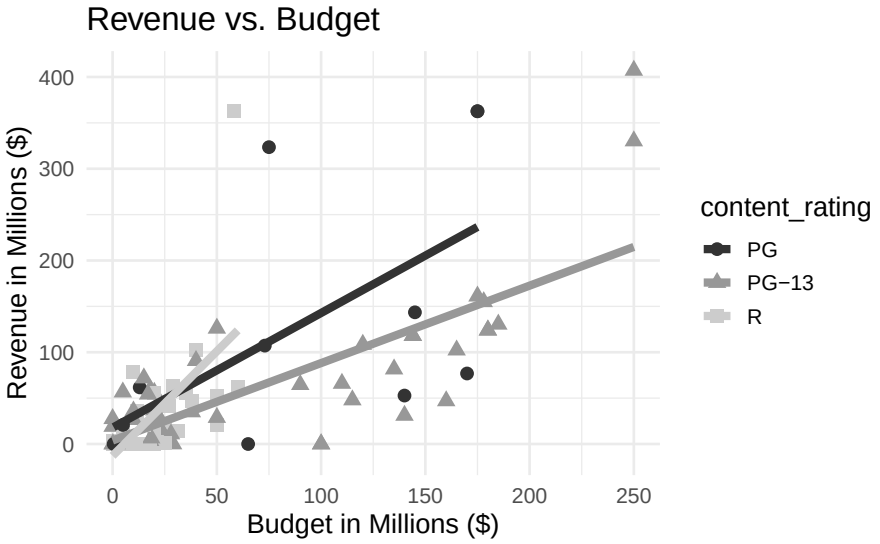
9. The movie *Independence Day: Resurgence* had a budget of 165 $MM and revenue of 102.315 $MM. Find the residual for this movie.

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

Multivariable plots

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

```
movies %>% # Data set pipes into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(x = budget_mil, y = revenue_mil, color = content_rating)) +  # Specify variables
  geom_point(aes(shape = content_rating), size = 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
```



11. Identify the three variables plotted in this graph.

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

0.6.4 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**.

0.6.5 Additional notes

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

0.7 Week 4 Lab: Penguins

0.7.1 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.

0.7.2 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 to get the linear model output.

```
# 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 your answer.**

MODULE 1

Exam 1 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the appropriate Exam 1 Review R script file to answer the following questions. Each adult (>21) islander was selected at random from all adult islanders. 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 measured after listening to music for each Islander. Heart rate and blood glucose levels were both measured before and then after drinking a caffeinated beverage.

Variable	Description
`Island`	Name of Island that the Islander resides on
`City`	Name of City in which the Islander resides
`Population`	Population of the City
`Name`	Name of Islander
`Consent`	Whether the Islander consented to be in the study
`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
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed
`Balance_Test`	Time balanced measured in seconds with eyes closed
`Blood_Glucose_before`	Level of blood glucose (mg/dL) before consuming assigned drink
`Heart_Rate_before`	Heart rate (bpm) before consuming assigned drink
`Blood_Glucose_after`	Level of blood glucose (mg/dL) after consuming assigned drink
`Heart_Rate_after`	Heart rate (bpm) after consuming assigned drink
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. What are the observational units?

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

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

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

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

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

c. Use the R script file to get the counts for each level of the variable. Fill in the following table with the success, failure, variable name, and counts using the values from the R output.

Variable	**Counts**
Success	
Failure	
Total	

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

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

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

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

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

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

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

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

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

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

c. Use the R script file to get the counts for each level and combination of variables. Fill in the following table with the variable names, levels of each variable, and counts using the values from the R output.

	Explanatory Variable		
Response variable	Group 1	Group 2	Total
Success			
Failure			
Total			

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

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

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

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

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

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

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

6. Use the appropriate Exam 1 Review R script file to find the appropriate summary statistic and graphical display of the data to assess the following research question: "Do Islanders who listen to classical music take less time to complete the puzzle cube after listening to the music than for Islanders that listen to heavy metal music?" Use classical - heavy metal as the order of subtraction.

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

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

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

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

c. Use the R script file to get the summary statistics for each level of the explanatory variable. Fill in the following table with the variable name, levels of the variable, and the summary statistics from the R output.

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

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

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

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

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

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

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

Shape:

Center:

Spread:

Outliers:

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

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

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

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

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

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

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

c. Use the R script file to get the summary statistics for this data. Fill in the following table using the values from the R output:

	y-intercept	slope	correlation
Summary value			

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

e. Interpret the value of correlation in context of the problem.

f. Calculate the value of the coefficient of determination.

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

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

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

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

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

Form:

Direction:

Strength:

Outliers:

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

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

1.1 Out of Class Activity 6: Helperer-Hinderer — Simulation-based Hypothesis Test

1.1.1 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.

1.1.2 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 Chapters 9 & 14 in your textbook.

1.1.3 Steps of the statistical investigation process

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

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

1.1.4 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

The following code reads in the data set and gives the number of infants in each level of the variable, whether the infant chose the helper or the hinderer. Remember to visually display this data we can use either a frequency bar plot or a relative frequency bar plot.

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

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

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

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

4. Sketch a relative frequency bar plot of these data.

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

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

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

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

6. 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.

7. 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?

8. Write out the alternative hypothesis in notation.

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

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

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

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

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

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

10. Flip a coin 16 times recording the number of times the coin lands on heads. This represents one simulated sample of 16 infants randomly choosing the toy.

11. Is the value from question 10 closer to 0.5, the null value, or closer to the sample proportion, 0.875?

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.

1.1.5 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.

1.1.6 Additional notes

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

1.2 Activity 6: Helper-Hinderer (continued)

1.2.1 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

1.2.2 Steps of the statistical investigation process

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

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

1.2.3 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 the out of class activity.

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

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

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

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

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

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

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

We will use the `one_proportion_test()` function in `R` (in the `catstats` package) to simulate the null distribution of sample proportions and compute a p-value. Using the provided `R` script file, fill in the values/words for each `xx` with your answers from question 3 in the one proportion test to create a null distribution with 1000 simulations. Then highlight and run lines 1–15.

```
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?

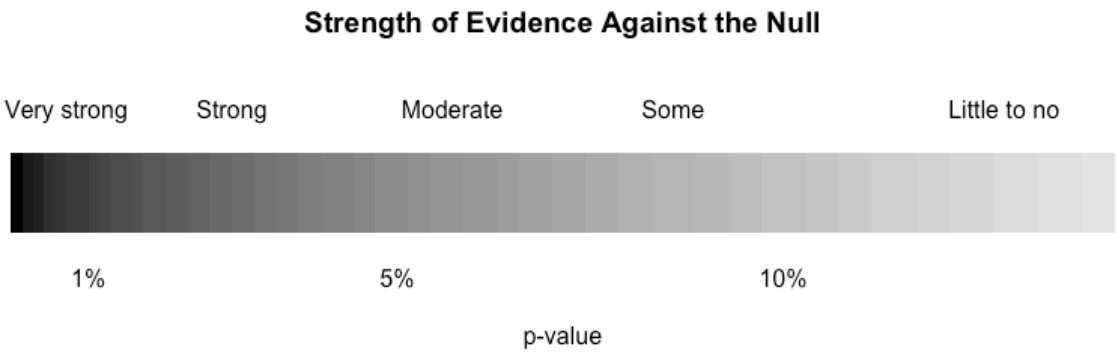
6. Circle the observed statistic (value from question 1) on the distribution you drew in question 4. Where does this statistic fall in the null distribution: Is it near the center of the distribution (near 0.5) or in one of the tails of the distribution?

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?

1.2.4 Take-home messages

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

1.2.5 Additional notes

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

1.3 Week 6 Lab: Helper-Hinderer — Simulation-based Confidence Interval

1.3.1 Learning outcomes

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

1.3.2 Terminology review

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

- Parameter of interest
- Bootstrapping
- Confidence interval

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

1.3.3 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.

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 in Activity 6A.*

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?**

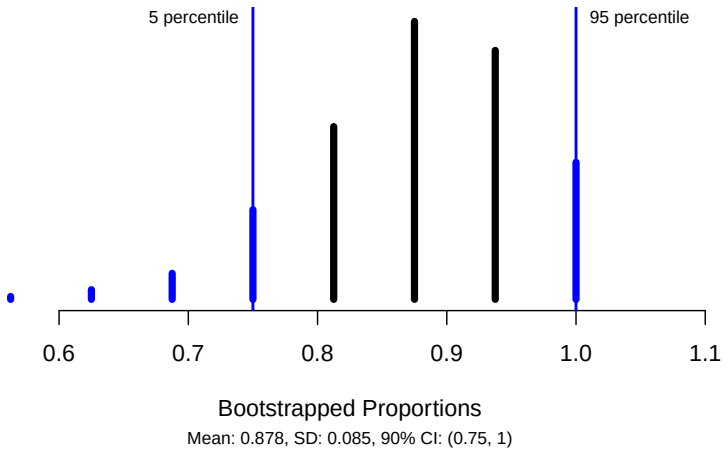
Explain how this indicates that the p-value provides strong evidence against the null.

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.

16. 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. In addition, a 95% confidence interval was found for the parameter of interest. We are 95% confident that the (true/sample)_____ proportion of infants age 6 to 10 months who will choose the helper toy is between (lower value)_____ and (upper value)_____. The results of this study can be generalized to (all infants age 6 to 10 months/infants similar to those in this study)_____ as the researchers (did/did not)_____ select a random sample.

1.3.4 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.

1.3.5 Additional notes

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

1.4 Out of Class Activity 7: Handedness of Male Boxers

1.4.1 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.

1.4.2 Terminology review

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

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

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

Activities 6A, 6B, 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.

1.4.3 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, $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

5. Verify that the independence condition is satisfied.

6. Is the success-failure condition met to model the data with the normal distribution? Show your work to support your answer.

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:

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

.

The standard error of the sample proportion measures the variability of possible sample proportions from the actual proportion. In other words, how far each possible sample proportion is from the actual proportion on average. For this study, the null standard error of the sample proportion is calculated using the null value, 0.1.

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

.

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

7. Using the null standard error of the sample proportion, 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.

8. 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 1.1, and also mark the value of the standardized statistic calculated in question 8.

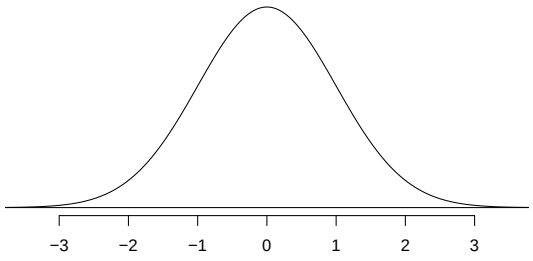


Figure 1.1: A standard normal curve.

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

9. Interpret the standardized sample proportion from question 7 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 7 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
```

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

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

1.4.4 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.

1.4.5 Additional notes

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

1.5 Activity 7: Handedness of Male Boxers — Theory CI

1.5.1 Learning objectives

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

1.5.2 Terminology review

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

- Parameter of interest
- Multiplier
- Normal distribution

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

1.5.3 Handedness of Male Boxers

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

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

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

Theory-based confidence interval

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

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

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

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

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

$$\text{point estimate} \pm \text{margin of error}$$
$$\hat{p} \pm z^* SE(\hat{p})$$
$$ME = z^* SE(\hat{p})$$

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.

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

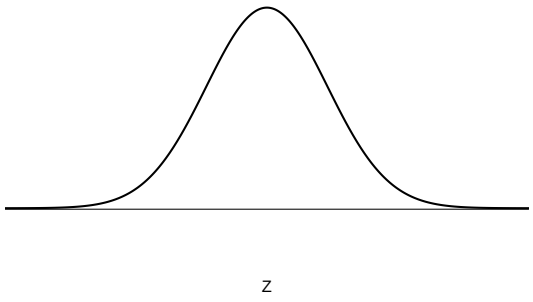


Figure 1.2: A standard normal curve.

The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, 95% + 2.5% = 97.5% percentile). Enter the value of 0.975 for xx in the provided R script file. This will give the value of the multiplier for a 95% confidence interval.

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

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

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

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

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

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

Simulation Methods

In activity 7A, we found that the success-failure condition was met to use theory-based methods. Here we will use simulation methods to find a 95% confidence interval for the parameter of interest.

Use the `one_proportion_bootstrap_CI()` function in R to simulate the bootstrap distribution of sample proportions and calculate a confidence interval. Using the provided R script file, fill in the values/words for each `xx` in the one proportion bootstrap confidence interval (CI) code to create a bootstrap distribution with 1000 simulations. Make sure to run the library(catstats) function before running the one_proportion_bootstrap_CI function.

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

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

What does *confidence* mean?

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

For this part of the activity we will assume that the the true proportion of male boxers that are left-handed is 0.1. *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.*

9. 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 0.1 for π (the true value), 500 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 0.1?

    b. Now click on a red dot. Write down the confidence interval for this sample. Does this confidence interval contain the true value of 0.1?

    c. How many intervals out of 100 contain π, the true value of 0.1? *Hint*: This is given to the left of the graph of green and red intervals.

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

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

1.5.4 Take-home messages

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

3. If repeat samples of the same size are selected from the population, approximately 95% of samples will create a 95% confidence interval that contains the parameter of interest.

1.5.5 Additional notes

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

1.6 Week 7 Lab: Errors and Power

1.6.1 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.

1.6.2 Terminology review

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

- Significance level
- Type 1 error
- Type 2 error
- Power

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

1.6.3 ACL recovery

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

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

2. **Use proper notation to write the null and alternative hypothesis the company would need to test in order to check their advertisement claim.**

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

As you have already learned in this course, sample size of a study is extremely important. Often times, researchers will conduct what is called a power analysis to determine the appropriate sample size based on the goals of their research, including a desired **power** of their test. Power is the probability of correctly rejecting the null hypothesis, or the probability of the data providing strong evidence against the null hypothesis *when the null hypothesis is false*.

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

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

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

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

4. Check the box for "Show Critical Value(s) and Rejection Region(s)". You will now see a vertical line on the plot indicating the *minimum* sample proportion which would lead to reject the null hypothesis. What is this value?

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

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

level. Therefore the significance level is the maximum probability of rejecting the null hypothesis when the null hypothesis is true, so the significance level IS the probability of making a type 1 error in a hypothesis test!

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

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

7. Where is the blue distribution centered?

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

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

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

"Average Driving Distance and Fairway Accuracy." 2008. https://www.pga.com/ and https://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 from https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm.

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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