STAT 216 Coursepack

Summer 2022
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

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

Contents

Preface

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

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

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

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

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

MODULE 1

Basics of Data

1.1 Module 1 Reading Guide: Basics of Data

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

Videos

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

Vocabulary

Data:

Summary statistic:

Case/Observational unit:

Variable:

Quantitative variable:

Discrete variables:

Examples of discrete variables using the `County` data:

Continuous variables:

Examples of continuous variables using the `County` data:

Example of a number which is NOT a numerical variable:

Categorical variable:

Ordinal variable:

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

Nominal variable:

Examples of nominal variables using the `County` data:

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

Data frame:

Scatterplot:

Each point represents:

Positive association:

Negative association:

Association or Dependent variables:

Independent variables:

Explanatory variable:

Response variable:

Observational study:

Experiment:

Placebo:

Notes

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

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

In a data frame, rows correspond to _____

and columns correspond to _____.

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

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

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

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

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

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

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

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

5. What is the purpose of the control group?

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

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

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

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

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

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

1.2 Activity 1: Martian Alphabet

1.2.1 Learning outcomes

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

1.2.2 Terminology review

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

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

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

1.2.3 General information labs

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

1.2.4 Can you read "Martian?"

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

1. Were you correct or incorrect in identifying Bumba?

Steps of the statistical investigation process

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

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

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

2. What are the observational units in this study?

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

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

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

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

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

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

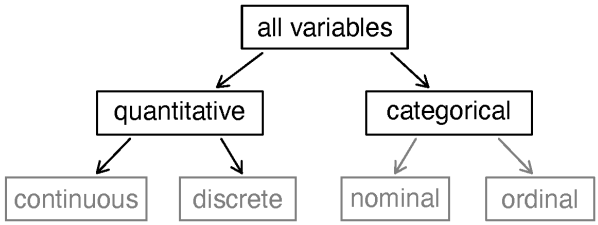


Figure 1.1: Types of variables.

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

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

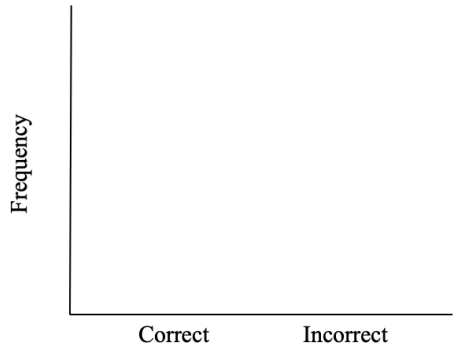
6. How many people in your class were correct in identifying Bumba? Using the class size from question 3, calculate the proportion of students who correctly identified Bumba.

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

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

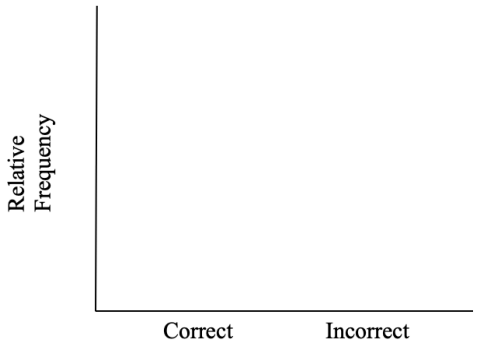
Looking at the data set and the summary statistic is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels or outcomes, correct or incorrect,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

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



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

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



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

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

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

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

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

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

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

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

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

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

    Probability of correct guesses: _____

    Sample size: _____

    Number of repetitions: _____

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

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

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

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

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

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

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

Introduction to R

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

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

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

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

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

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

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

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

The # sign is not part of the `R` code. It is used by these authors to add comments to the `R` code and explain what each call is telling the program to do. `R` will ignore everything after a # sign when executing the code.

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

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

1.2.5 Take-home messages

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

1.2.6 Additional notes

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

MODULE 2

Study Design

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

Section 1.3 (Sampling principles and strategies)

Videos

- 1.3

Vocabulary

(Target) Population:

Sample:

Anecdotal evidence:

Bias:

Selection bias:

Non-response bias:

Response bias:

Convenience sample:

Simple Random Sample:

Non-response rate:

Representative:

Notes

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

Notes on types of sampling bias

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

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

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

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

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

Videos

- 1.4to1.6

Reminders from Section 1.2

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

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

Vocabulary

Observational study:

Observational data:

Prospective study:

Retrospective study:

Confounding variable:

Experiment:

Randomized experiment:

Blocking:

Treatment group:

Control group:

Placebo:

Placebo effect:

Blinding:

Scope of inference:

Generalizability:

Causation:

Notes

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

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

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

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

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

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

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

2.2 Activity 2A: American Indian Address

2.2.1 Learning outcomes

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

2.2.2 Terminology review

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

- Random sample
- Unbiased vs biased methods of selection
- Types of samping bias
- Generalization

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

Types of sampling bias.

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

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

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

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

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

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

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

   Target population:

   Sample:

   Variable:

   Type(s) of bias:

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

2.2.3 American Indian Address

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

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

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

By eye selection

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

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

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

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

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

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

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

    Where does the value of 3.95 fall in our plot above?

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

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

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

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

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

2.2.4 Take-home messages

1. There are three types of bias to be aware of when designing a sampling method: selection bias, non-response bias, and response bias.
2. When we use a biased method of selection, we will over or underestimate the parameter.
3. To see if a method is biased, we compare the distribution of the estimates to the true value. We want our estimate to be on target or unbiased. When using unbiased methods of selection, the mean of the distribution matches our true parameter.
4. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.

2.2.5 Additional notes

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

2.3 Activity 2B: American Indian Address (continued)

2.3.1 Learning outcomes

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

2.3.2 Terminology review

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

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

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

Random selection

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

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

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

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

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

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

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

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

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

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

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

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

   Which features are similar?

   Which features differ?

   Why didn't everyone get the same sample mean?

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

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

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

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

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

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

Effect of sample size

We will now consider the impact of sample size.

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

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

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

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

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

    Which features are similar?

    Which features differ?

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

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

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

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

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

2.3.3 Take-home messages

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

2.3.4 Additional notes

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

2.4 Module 2 Lab: Study Design

2.4.1 Learning outcomes

- Explain the purpose of random assignment and its effect on scope of inference.
- Identify whether a study design is observational or an experiment.
- Identify confounding variables in observational studies and explain why they are confounding.

2.4.2 Terminology review

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

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

To review these concepts, see Sections 1.2 through 1.6 in the textbook.

2.4.3 General information labs

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

2.4.4 Study design

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

For the next exercises, identify the explanatory variable, the response variable, and the study design (observational study or experiment).

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

   Explanatory variable:

   Response variable:

   Study design:

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

   Explanatory variable:

   Response variable:

   Study design:

2.4.5 Atrial fibrillation

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

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

   Explanatory variable:

   Response variable:

Suppose 24 subjects with permanent atrial fibrillation have volunteered to participate in this study:

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

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

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

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

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

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

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

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

What is the sample size of each group?

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

9. **Assume the scientists used the strategy in question 8, but they put the four tallest non-males and eight tallest males into the placebo group and the remaining subjects into the control group. They found that the proportion of patients whose heart rate stabilized is higher in the drug group than the placebo group.**

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

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

While the strategy presented in question 9 controlled for the sex of the subject, there are more potential **confounding variables** in the study. A confounding variable is a variable that is *both*

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

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

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

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

- Navigate to the "Randomizing Subjects" applet under the "Other Applets" heading at: http://www.rossmanchance.com/ISIapplets.html. This applet lists the sex and height of each of the 24 subjects. Click "Show Graphs" to see a bar chart showing the sex of each subject. Currently, the applet is showing the strategy outlined in question 7.
- Click "Randomize."

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

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

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

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

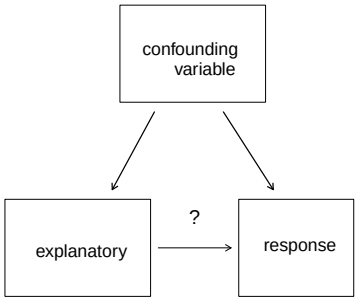
12. Change Repetitions under "Simulation" to 998 (for 1000 total). Sketch the plot of the distribution of difference in proportions from each of the 1000 random assignments here. Be sure to include a descriptive x-axis label.

13. **Does random assignment *always* balance the placebo and drug groups based on the sex of the participants? Does random assignment *tend* to make the placebo and drug groups *roughly* the same with respect to the distribution of sex? Use your plot from question 12 to justify your answers.**

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

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

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



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

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

18. **Both the sampling method (which we covered earlier this week) and the study design will help to determine the *scope of inference* for a study: To *whom* can we generalize, and can we conclude *causation or only association*? Use the table below to determine the scope of inference of this trial study described in question 17.**

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

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

Inferences to population can be made

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

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

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

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

2.4.6 Take-home messages

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

2.4.7 Additional notes

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

MODULE 3

Exploring Categorical and Quantitative Data

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

Section 1.7 (Data in R)

Videos

- Starting_with_R

Notes

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

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

Tidy data: Data frames with

1 row per _____,

1 column per _____.

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

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

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

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

Functions

State what these introductory functions do in R:

`glimpse(data_set_name)`

`head(data_set_name)`

`data_set_name$variable_name`

`%>%`

`<-`

Section 2.1 (Exploring categorical data)

Videos

- 2.1
- MosaicPlots

Vocabulary

Frequency table:

Relative frequency table:

Contingency or two-way table:

Unconditional proportion:

Conditional proportion:

Row proportions:

Column proportions:

Statistic:

Sample proportion:

Notation:

Parameter:

Population proportion:

Notation:

Bar plot:

Segmented bar plot:

Simpson's Paradox:

Notes

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

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

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

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

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

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

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

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

Review of Simpson's Paradox

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

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

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

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

Section 2.3 (Exploring quantitative data)

Videos

- 2.3

Type of Plots

Scatterplot:

Dot plot:

Histogram:

Density plot:

Box plot:

Vocabulary

Four characteristics of a scatterplot:

Form:

Strength:

Direction:

Unusual observations or outliers:

Data density:

Tail:

Skew:

Symmetric:

Modality:

Distribution (of a variable):

Four characteristics of the distribution of one quantitative variable:

Center:

Variability:

Shape:

Outliers:

Point estimate:

Deviation:

Five number summary:

X^{th} percentile:

Interquartile range (IQR):

Robust statistics:

Notes

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

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

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

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

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

How are variance and standard deviation related?

Fill in the following table with the appropriate notation.

Summary Measure	Parameter	Statistic
Mean		
Variance		
Standard deviation		

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

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

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

Notes

Statistics summarize _____ .
Parameters summarize _____.

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

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

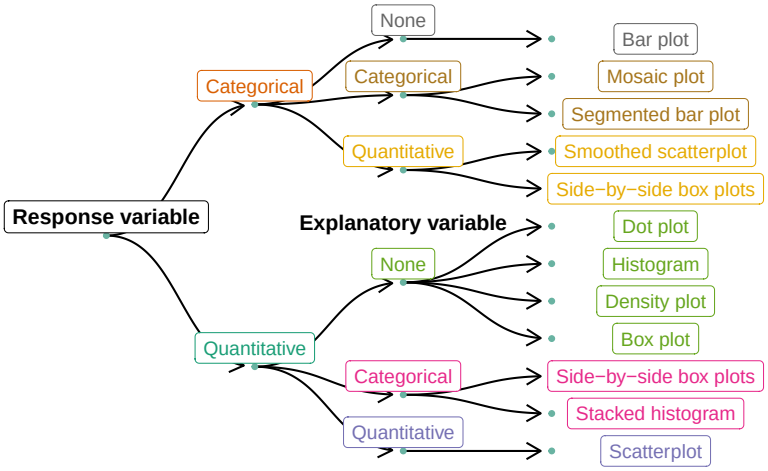
Look at the table of vocabulary terms. If there are any you do not know, be sure to review the appropriate section of your text.

Data visualization summary

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

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

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



3.2 Activity 3A: Graphing Categorical Variables

3.2.1 Learning outcomes

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

3.2.2 Terminology review

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

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

To review these concepts, see Sections 2.1 and 2.2 in the textbook.

3.2.3 Graphing categorical variables

Nightlight use and myopia

In a study reported in Nature (Quinn et al. 1999), a survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 2 had a higher incidence of nearsightedness (myopia) later in childhood.

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

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

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

R code

Throughout these activities, we will often include the R code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into R using the `read.csv()` function. These lines of code read in the data set and name the data set `myopia`. Highlight and run lines 1–6 in the R script file to load the needed packages and the data.

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

Displaying a single categorical variable

If we wanted to know how many children in our data set were in each level of myopia, we would create a frequency bar plot of the variable `Sight`. Enter the variable name, `Sight`, for `variable` into the `ggplot` code at line 10 in the R script file. Highlight and run lines 9–15 to create the plot. Note: this is a **frequency** bar plot plotting counts (the number of children in each level of sight is displayed on the y-axis).

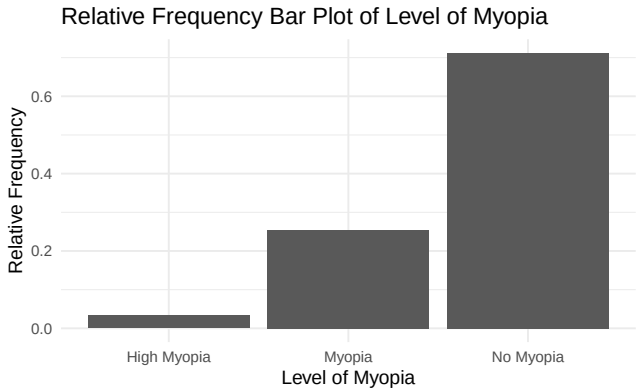
```
myopia %>% # Data set piped into...
ggplot(aes(y = variable)) +   # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Frequency",   # Label the x axis
       y = "Level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

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

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

We could also choose to display the data as a proportion in a **relative frequency** bar plot. To find the relative frequency, divide the count in each level of myopia by the sample size. These are sample proportions. Notice that in this code we told `R` to create a bar plot with proportions.

```
myopia %>% # Data set piped into...
ggplot(aes(x = Sight)) +   # This specifies the variable
  geom_bar(aes(y = ..prop.., group = 1)) +  # Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



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

Displaying two categorical variables

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

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

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

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

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

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

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

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

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

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

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

10. Fill in the following table with the values from the `R` output.

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

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

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

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

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

3.2.4 Take-home messages

1. Bar charts can be used to graphically display a single categorical variable either as counts or proportions. Segmented bar charts and mosaic plots are used to display two categorical variables.
2. Segmented bar charts always have a scale from 0 - 100%. The bars represent the outcomes of the explanatory variable. Each bar is segmented by the response variable. If the heights of each segment are the same for each bar there is no association between variables.
3. Mosaic plots are similar to segmented bar charts but the widths of the bars also show the number of observations within each outcome.

3.2.5 Additional notes

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

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

3.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

3.3.2 Terminology review

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

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

To review these concepts, see Section 2.3 in the textbook.

3.3.3 Movies released in 2016

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

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

Summarizing a single quantitative variable

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable `imdb_score`. The summary statistics give the two measures of center and two measures of spread for IMDb score. Highlight and run lines 1 – 8 in the provided `R` script file to load the data set. Check that the summary statistics match that printed in the coursepack.

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

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

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

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

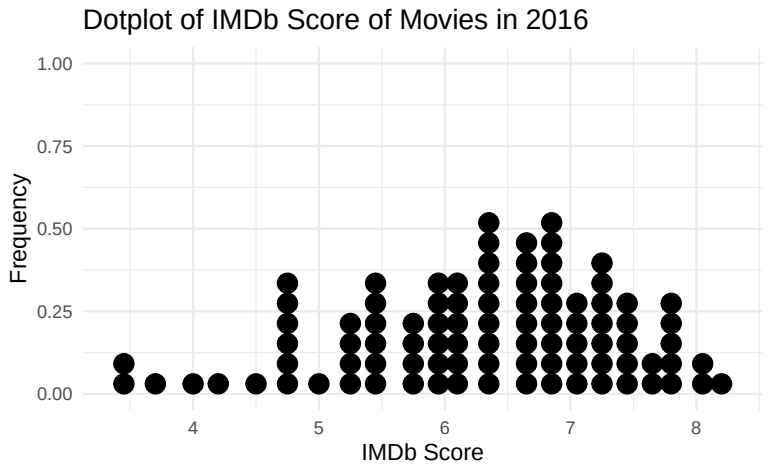
3. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

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

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

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



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

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

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

```
movies %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 0.5) +  # Create histogram with specified binwidth
  labs(title = "Histogram of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

6. Sketch the histogram created here.

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

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

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

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

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

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

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

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

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

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

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

Summary statistics for a single categorical and single quantitative Variable

Is there an association between content rating and budget for movies in 2016? To use the `favstats()` function in the mosaic package with two variables, we will enter the variables as a formula, response~explanatory. This function will give the summary statistics for budget for each content rating. Highlight and run lines 37–39 in the provided `R` script file and check that the summary statistics match those provided in the coursepack.

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

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

11. Which content rating has the largest IQR?

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

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

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

Displaying a single categorical and single quantitative variable

The boxplot of movie budgets (in millions) by content rating is plotted using the code below. Enter the variable `budget_mil` for `response` and the variable `content_rating` for explanatory at line 44, highlight and run code lines 42–48. This plot compares the budget for different levels of content rating.

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

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

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

    a. Which content rating has the highest center?

    b. Which content rating has the largest spread?

    c. Which content rating has the most skewed distribution?

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

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

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

3.3.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.
2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Values in the data set that are less than $Q_1 - 1.5 * IQR$ and greater than $Q_3 + 1.5 * IQR$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.
3. Data should be summarized numerically and displayed graphically to give us information about the study.
4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, IQR = Q3 - Q1.

3.3.5 Additional notes

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

3.4 Module 3 Lab: IPEDs

3.4.1 Learning outcomes

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

3.4.2 The Integrated Postsecondary Education Data System (IPEDS)

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

These data are on a subset of institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting
- United States only
- Title IV participating
- Not for profit
- 2-year or 4-year or above
- Has full-time first-time undergraduates
- Note that several variables have missing values for some institutions (denoted by "NA").

Variable Name	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Control	• Public • Private
Sector	• Public 2-year • Private 2-year • Public 4-year or higher • Private 4-year or higher
LandGrant	Is this a land-grant institution? (Yes/No)
Size	Institution size category based on total students enrolled for credit, Fall 2018: • Under 1,000 • 1,000 - 4,999 • 5,000 - 9,999 • 10,000 - 19,999 • 20,000 and above
Cost_OutofState	Cost of attendance for full-time, first-time degree/certificate seeking out-of-state undergraduate students living on campus for academic year 2018-19. It includes in-out-of-state tuition and fees, books and supplies, on campus room and board, and other on campus expenses.
Cost_InState	Cost of attendance for full-time, first-time degree/certificate seeking in-state undergraduate students living on campus for academic year 2018-19. It includes in-state tuition and fees, books and supplies, on campus room and board, and other on campus expenses.
Retention	The full-time retention rate is the percent of the (fall full-time cohort from the prior year minus exclusions from the fall full-time cohort) that re-enrolled at the institution as either full- or part-time in the current year
Percent_InState	Percent of first-time degree/certificate seeking undergraduate students who reside in the same state of the institution.
Enrollment	Total number of people enrolled for credit in the fall of the academic year.
Graduation_Rate	Graduation rate of first-time, full-time degree or certificate-seeking students - 2012 cohort (4-year institutions) and 2015 cohort (less-than-4-year institutions). This rate is calculated as the total number of completers within 150% of normal time divided by the revised cohort minus any allowable exclusions.
Percent_FinancialAid	Percentage of all full-time, first-time degree/certificate-seeking undergraduate students who were awarded any financial aid.

Summarizing a single quantitative variable

1. What are the observational units for this study?

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

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

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

```
IPEDS <- datasetname #Creates the object IPEDS
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") #Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") #Filters the data set to remove Private 2-year
IPEDS %>%
  summarise(favstats(variable)) #Gives the summary statistics
```

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

4. Calculate the interquartile range for this study.

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

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

Next we will create both a histogram and a boxplot of the variable `Retention`. Enter the name of the variable in both line 16 and line 23 for `variable` in the `R` script file. **Give each plot a descriptive title.** Highlight and run lines 15 – 27 to give the histogram and boxplot. **Export and upload both plots to Gradescope for your group.** To export the graphs: in the bottom right corner in the Plots tab, click on `Export`, then choose `Save as Image`. Save the image as a png. This will save your graph to the server. In the Files tab, click on the box next to your saved image file, click `More` and choose `Export`. This will save your file to your downloads folder on your computer.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 5) +  # Create histogram with specified binwidth
  labs(title = "Title", # Title for plot
       x = "Rentention Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

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

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

8. Identify any outliers in the data set.

Robust Statistics

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

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

10. Report the two measures of spread for retention given in the R output.

To show the effect of outliers on the measures of center and spread, the smallest values of retention rate in the data set were increased by 30%. Highlight and run lines 30–38.

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

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

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

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

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

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

Summarizing a single categorical and single quantitative variable

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

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

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

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

```
IPEDS %>%  # Data set piped into...
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Create box plot
  labs(title = "Side by side box plot of retention rates by control",  # Title
       x = "Control",    # x-axis label
       y = "Retention Rates")  # y-axis label
```

16. **Compare the two boxplots.**

    Which type of university has the highest center?

    Largest spread?

    What is the shape of each distribution?

    Does either distribution have outliers?

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

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

Summarizing two categorical variables

Are private 4-year institutions smaller than public one? The following set of code will create a segmented bar plot of size of the institution by sector. Enter the variable `Sector` for explanatory and `Size` for response in line 53. Highlight and run lines 52 – 58 in the `R` script file.

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

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

MODULE 4

Exploring Multivariable Data

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

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

Videos

- Chapter3

Reminders from Section 2.3

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

Four characteristics of a scatterplot:

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

Vocabulary

Residual:

Formula:

Residual plot:

Correlation:

Notes

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

x represents

y represents

β_0 represents

β_1 represents

ϵ represents

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

x represents

$\hat{y}$ represents

b_0 represents

b_1 represents

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

Summary Measure	Parameter	Statistic
Correlation		
Slope		
y-intercept		

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

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

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

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

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

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

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

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

Example: Brushtail possums

1. What are the observational units?

2. Look at the scatterplot in Figure 3.5.

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

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

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

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

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

Section 3.2 (Least squares regression)

You may skip the special topic Sections 3.2.3.1 and 3.2.6.

Videos

- Chapter3

Vocabulary

Least squares criterion:

Least squares line:

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

slope:

y-intercept:

Extrapolation:

Coefficient of determination:

s_y^2 (or SST) represents

s_{RES}^2 (or SSE) represents

Notes

Two methods for determining the best line:

1.

2.

Notation for the coefficient of determination:

Formulas for calculating the coefficient of determination:

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

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

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

Example: Elmhurst College

1. What are the observational units?

2. Look at the scatterplot in Figure 3.13.

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

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

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

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

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

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

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

Section 3.3 (Outliers in linear regression)

Videos

- Chapter3

Vocabulary

Outlier:

Leverage:

Influential:

Notes

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

True or false: All high leverage outliers are influential.

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

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

Videos

- Chapter3

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

Notes

Statistics summarize:

Parameters summarize:

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

What is the formula for calculating a residual?

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

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

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

Fill in the blanks to correctly interpret:

- Slope:

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

- y-intercept:

  If _____, we predict the _____ to equal _____.

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

Section 4.1 (Gapminder world)

Videos

- Chapter4

Reminder from Section 3.1

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

Vocabulary

Interaction:

Aesthetic:

Notes

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

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

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

Section 4.2 (Simpson's Paradox, revisited)

Videos

- Chapter4

Reminder from Section 2.1

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

Notes

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

Example: SAT scores

1. What are the observational units?

2. Look at the scatterplot in Figure 4.5.

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

   b) What is the form of the scatterplot?

   c) What is the direction of the scatterplot?

   d) What is the strength of the scatterplot?

   e) Are there any outliers on the scatterplot?

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

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

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

Section 4.4 (Chapter 4 review)

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

Videos

- Chapter4

Notes

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

Simpson's Paradox:

4.2 Activity 4A: Movie Profits — Linear Regression

4.2.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- Use scatterplots to assess the relationship between two quantitative variables.
- Find the estimated line of regression using summary statistics and `R` linear model (`lm()`) output.
- Interpret the slope coefficient in context of the problem.

4.2.2 Terminology review

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

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

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

4.2.3 Movies released in 2016

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

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

Vocabulary review

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

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

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

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

3. Sketch the scatterplot created from the code.

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

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

5. Does there appear to be an association between budget and revenue? Explain.

Slope

The linear model function in `R` (`lm()`) gives us the summary for the least squares regression line. The estimate for `(Intercept)` is the y-intercept for the line of least squares, and the estimate for `budget_mil` (the x-variable name) is the value of b_1, the slope.

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

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

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

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

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

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

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

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

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

10. The prediction in question 9 is an example of what?

Residuals

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

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

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

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

4.2.4 Take-home messages

1. Two quantitative variables are graphically displayed in a scatterplot. The explanatory variable is on the x-axis and the response variable is on the y-axis. When describing the relationship between two quantitative variables we look at the form (linear or non-linear), direction (positive or negative), strength, and for the presence of outliers.
2. There are three summary statistics used to summarize the relationship between two quantitative variables: correlation (R), slope of the regression line (b_1), and the coefficient of determination (R^2).
3. We can use the line of regression to predict values of the response variable for values of the explanatory variable. Do not use values of the explanatory variable that are outside of the range of values in the data set to predict values of the response variable (reflect on why this is true.). This is called **extrapolation**.

4.2.5 Additional notes

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

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

4.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- Calculate and interpret R^2, the coefficient of determination, in context of the problem.
- Find the correlation coefficient from `R` output or from R^2 and the sign of the slope.

4.3.2 Terminology review

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

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

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

4.3.3 Movies released in 2016

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

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

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

Correlation

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

```
movies %>%  # Data set pipes into
  select(c("budget_mil", "revenue_mil",
           "duration", "imdb_score",
           "facebook_likes")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

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

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

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

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

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

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

Coefficient of determination (squared correlation)

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

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

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

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

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

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

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

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

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

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

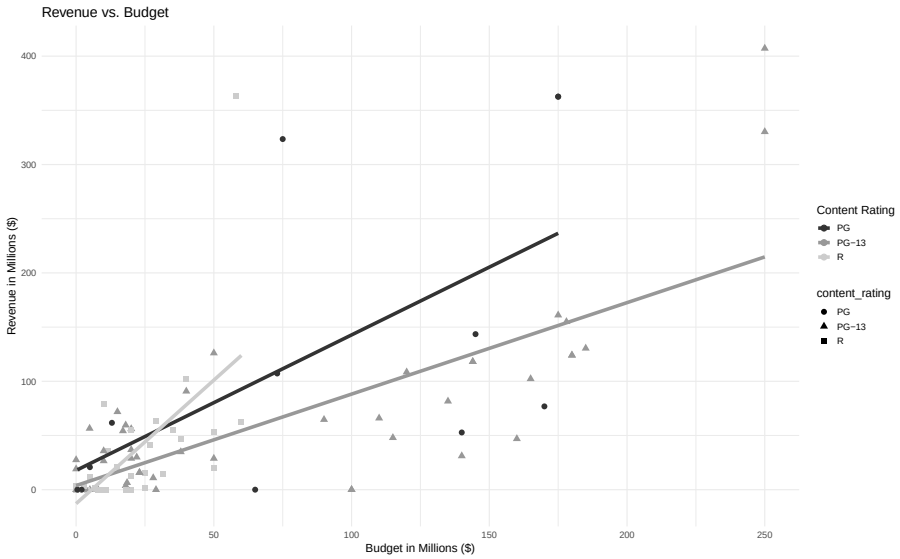
12. Calculate the value for R^2 using the values found for SST and SSE.

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

Multivariable plots

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

```
movies %>% # Data set pipes into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(x = budget_mil, y = revenue_mil, color = content_rating)) +  # Specify variables
  geom_point(aes(shape = content_rating), size = 3) +  # Add scatterplot of points
  labs(x = "Budget in Millions ($)",  # Label x-axis
       y = "Revenue in Millions ($)",  # Label y-axis
       color = "Content Rating",  # Label legend
       title = "Revenue vs. Budget") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE, lwd = 2) + # Add regression lines
  scale_color_grey() # Make black and white
```



14. Identify the three variables plotted in this graph.

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

4.3.4 Take-home messages

1. The sign of correlation and the sign of the slope will always be the same. The closer the value of correlation is to -1 or $+1$, the stronger the relationship between the explanatory and the response variable.
2. The coefficient of determination multiplied by 100 ($R^2 \times 100$) measures the percent of variation in the response variable that is explained by the relationship with the explanatory variable. The closer the value of the coefficient of determination is to 100%, the stronger the relationship.

4.3.5 Additional notes

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

4.4 Module 4 Lab: Penguins

4.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- Use scatterplots to assess the relationship between two quantitative variables.
- Find the estimated line of regression using summary statistics and `R` linear model (`lm()`) output.
- Interpret the slope coefficient in context of the problem.
- Calculate and interpret R^2, the coefficient of determination, in context of the problem.
- Find the correlation coefficient from `R` output or from R^2 and the sign of the slope.

4.4.2 Penguins

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

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

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

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

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

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

Highlight and run lines 16–20 to get the correlation matrix in the R script file.

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

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

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

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

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

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

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

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

7. **Using the least squares regression line from question 5, predict the body mass for a penguin with a flipper length of 181 mm.**

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

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

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

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

10. What three variables are plotted on this plot?

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

MODULE 5

Exam 1 Review

Use the provided data set from the Islands (ExamReviewData.csv) and the Exam 1 Review `R` script file to answer the following questions. Each adult (>21) islander was selected at random from all the adult islanders. Variables and their descriptions are listed below. Music type (classical or heavy metal) was randomly assigned to the Islanders. Time to complete the puzzle cube was measure before listening to the music and then after listening to music for each Islander. Heart rate and blood glucose levels were both measured before and then after drinking a caffeinated beverage.

Variable	Description
`Island`	Name of Island that the Islander resides on
`City`	Name of City in which the Islander resides
`Population`	Population of the City
`Name`	Name of Islander
`Consent`	Whether the Islander consented to be in the study
`Gender`	Gender of Islander (M = male, F = Female)
`Age`	Age of Islander
`Married`	Marital status of Islander
`Smoking_Status`	Whether the Islander is a current smoker
`Children`	Whether the Islander has children
`weight_kg`	Weight measured in kg
`height_cm`	Height measured in cm
`respiratory_rate`	Breaths per minute
`Type_of_Music`	Music type (Classical or Heavy Medal) Islander was randomly assigned to listen to
`Before_PuzzleCube`	Time to complete puzzle cube (minutes) before listening to assigned music
`After_PuzzleCube`	Time to complete puzzle cube (minutes) after listening to assigned music
`Education_Level`	Highest level of education completed (note: missing data depicted by missing)
`Balance_Test`	Time balanced measured in seconds with eyes closed
`Blood_Glucose_before`	Level of blood glucose (mg/dL) before consuming assigned drink
`Heart_Rate_before`	Heart rate (bpm) before consuming assigned drink
`Blood_Glucose_after`	Level of blood glucose (mg/dL) after consuming assigned drink
`Heart_Rate_after`	Heart rate (bpm) after consuming assigned drink
`Diff_Heart_Rate`	Difference in heart rate (bpm) for Before - After consuming assigned drink
`Diff_Blood_Glucose`	Difference in blood glucose (mg/dL) for Before - After consuming assigned drink

1. What are the observational units?

2. List all the variables that are categorical.

3. List all the variables that are quantitative.

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

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

   Variable:

   Value of Summary Statistic (with notation):

   Interpretation:

   Type of Graph:

   Sketch of the graph:

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

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

   Explanatory Variable:

   Response variable:

   Value of Summary Statistic (with notation):

      Interpretation:

   Type of Graph:

   Sketch of the graph:

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

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

What is the scope of inference for this study?

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

   Explanatory Variable:

   Response Variable:

   Value of Summary Statistic (with notation):

      Interpretation:

   Type of Graph:

   Sketch of the graph:

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

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

   Shape:

   Center:

   Spread:

   Outliers:

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

What is the scope of inference for this study?

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

   Explanatory Variable:

   Response Variable:

   Value of Summary Statistic (with notation):

   Slope:

   Interpretation:

   Correlation:

   Interpretation:

   Coefficient of Determination:

   Interpretation:

   Type of Graph:

   Sketch of the graph:

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

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

Form:

Direction:

Strength:

Outliers:

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

What is the scope of inference for this study?