STAT 216 Activity Coursepack

Fall 2020

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Introduction to Statistics at Montana State University. 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. Bring this workbook with you to class each week, and take notes in the workbook as you would your own notes. A well-written complete workbook will provide an optimal study guide for exams!

Fall 2020 Calendar of In-Class Activities

Week	Activity No.	Day	Date	Activity
1	1	M	8/17	Martian Alphabet
1	1	T	8/18	Martian Alphabet
1	1	W	8/19	Martian Alphabet
1	1	H	8/20	Martian Alphabet
1	1	F	8/21	Martian Alphabet
2	2	M	8/24	Study Design
2	2	T	8/25	Study Design
2	2	W	8/26	Study Design
2	2	H	8/27	Study Design
2	2	F	8/28	Study Design
3	3	M	8/31	Current Population Survey
3	3	T	9/1	Current Population Survey
3	3	W	9/2	Current Population Survey
3	3	H	9/3	Current Population Survey
3	3	F	9/4	Current Population Survey
4	-	M	9/7	No class − Labor Day
4	4	T	9/8	IMDb Movie Reviews
4	4	W	9/9	IMDb Movie Reviews
4	4	H	9/10	IMDb Movie Reviews
4	4	F	9/11	IMDb Movie Reviews
5	4	M	9/14	IMDb Movie Reviews
5	5	T	9/15	Movie Profits
5	5	W	9/16	Movie Profits
5	5	H	9/17	Movie Profits
5	5	F	9/18	Movie Profits
6	5	M	9/21	Movie Profits
6	—	T−F	9/22−9/25	Exam 1
7	—	M	9/28	Exam 1
7	6	T	9/29	Handedness of Male Boxers
7	6	W	9/30	Handedness of Male Boxers
7	6	H	10/1	Handedness of Male Boxers
7	6	F	10/2	Handedness of Male Boxers
8	6	M	10/5	Handedness of Male Boxers
8	7	T	10/6	Winter Sports Helmet Use and Head Injuries
8	7	W	10/7	Winter Sports Helmet Use and Head Injuries
8	7	H	10/8	Winter Sports Helmet Use and Head Injuries
8	7	F	10/9	Winter Sports Helmet Use and Head Injuries
9	7	M	10/12	Winter Sports Helmet Use and Head Injuries
9	8	T	10/13	COVID-19 and Air Pollution
9	8	W	10/14	COVID-19 and Air Pollution
9	8	H	10/15	COVID-19 and Air Pollution
9	8	F	10/16	COVID-19 and Air Pollution

Week	Activity No.	Day	Date	Activity
10	8	M	10/19	COVID-19 and Air Pollution
10	9	T	10/20	Weather Patterns and Record Snowfall
10	9	W	10/21	Weather Patterns and Record Snowfall
10	9	H	10/22	Weather Patterns and Record Snowfall
10	9	F	10/23	Weather Patterns and Record Snowfall
11	9	M	10/26	Weather Patterns and Record Snowfall
11	10	T	10/27	Hand Dexterity
11	10	W	10/28	Hand Dexterity
11	10	H	10/29	Hand Dexterity
11	10	F	10/30	Hand Dexterity
12	10	M	11/2	Hand Dexterity
12	−	T	11/3	No class − Election Day
12	−	W−F	11/4−11/6	Exam 2
13	−	M−T	11/9−11/10	Exam 2
13	−	W	11/11	No class − Veterans Day
13	−	H−W	11/12−11/18	Review

Martian Alphabet

1.1 Learning outcomes

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

1.2 Terminology review

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

- Observational units or cases
- Variables: categorical or quantitative
- 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.

1.3 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 the right.

1.3.1 Steps of the statistical investigation process

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!). To answer any research question, we must design a study and collect data. (This will normally be provided for you in class.) 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. To answer the research question we will simulate what *could* have happened in our class given random chance, repeat that many times to understand the expected variability

between different "randomly guessing" classes, then comparing our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) our class's result would occur if we were all merely guessing, allowing us to determine if we as a class can in fact read Martian.

Let's explore the data. **Observational units** or **cases** are the subjects data is collected on. In a data set the rows will represent a single observational unit.

1. What are the observational units in this study?

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

A **variable** is information collected or measured on each observational unit or case. Each column in a data set will represent a different variable.

3. Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?

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. Make sure you wrote the variable in question 3 as a variable **NOT** a summary statistic.

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

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

Categorical variables are data that are in groups or categories 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 a nominal variable. All categorical variables will be treated as nominal for analysis.

4. Is the variable identified in question 3 categorical or quantitative?

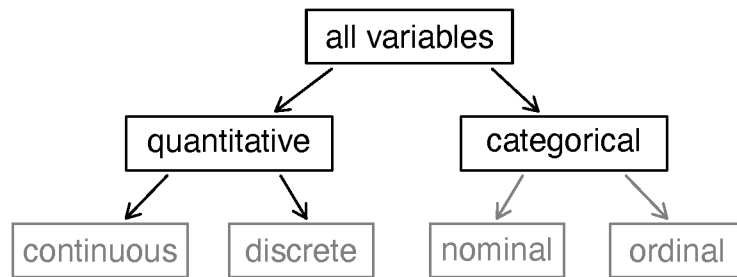5. Were you correct or incorrect in identifying Bumba?

Figure 1.1: Types of variables.

We will now collect the data from the entire class.

6. How many people in your class were correct in identifying Bumba? Using the class size from question 2, calculate the proportion of students who correctly identified Bumba.

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

Looking at the data set and the summary statistics 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, correct or incorrect, we will create two bars one for each level.

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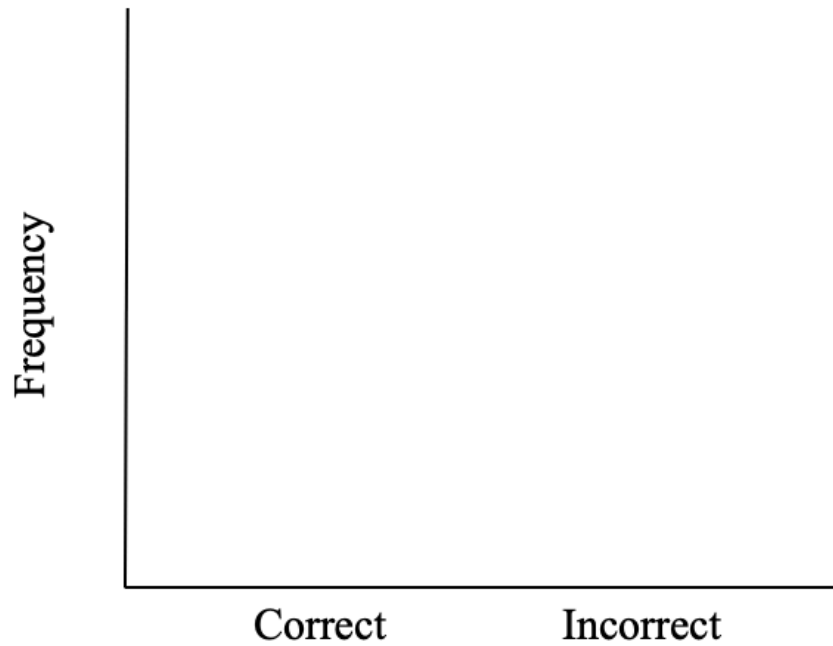
We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot.
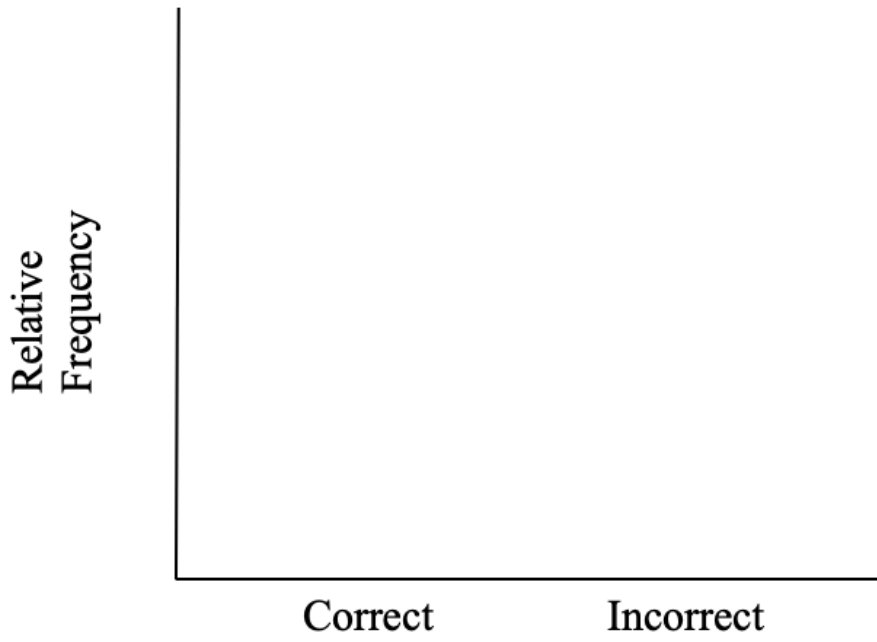
9. The next step is to analyze the data. If humans really don't know Martian and are just guessing which is Bumba, what are the chances of getting it right?

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 of you will flip a coin one time to simulate your "guess". Let Heads = correct, Tails = incorrect. What was the result of your 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 data 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 seem to give enough information. The differences seen could just be due to that set of coin flips! Let's simulate another class. Each student should flip your 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 unfortunately 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.

   Probability of correct guesses: _____

   Sample size: _____

   Number of repetitions: _____

14. Sketch the distribution displayed by your instructor here, being sure to label each axis appropriately.

15. Is your class particularly good or bad at Martian? How can you use the plot in question 14 to tell?

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.

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?

1.4   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 create data (via computer simulation) 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.

1.5   Additional notes

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

Study Design

2.1 Learning outcomes

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

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

- Identify whether a study is observational or an experiment

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

- Identify the types of bias present in a study

2.2 Terminology review

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

- Population

- Sample

- Parameter

- Statistic

- Selection bias

- Response bias

- Non-response bias

- Scope of inference

- Explanatory variable

- Response variable

- Confounding variable

- Experiment

- Observational study

To review these concepts see Sections 1.3 through 1.6 in the textbook.

2.3 Types of sampling bias

There are two parts to study design: how the sample was selected and how the study was conducted. First we will look at sampling and types of bias.

In these next questions, identify the target population, the sample, 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 of Bias:

2. PEW Research surveys US adults about many different topics. Recently a survey was conducted to assess current presidential approval. A random sample of 6395 US adults was taken. Of those surveyed, 42% say they agree with President Trump on many or nearly all of the top issues facing the country today.

   Target population:

   Sample:

   Variable:

   Type of Bias:

3. 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 of this or opposed to this. Of the 2241 viewers who phoned in, forty-five percent were opposed to legalizing marijuana.

   Target population:

   Sample:

   Variable:

   Type of Bias:

4. To gauge the interest in a new swimming pool, a local organization stood outside of the Bogart Pool 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 of Bias:

5. The Bozeman school district is interested in surveying parents of students about their opinions on returning to school this fall following the COVID-19 pandemic. They divided the school district into 10 divisions based on location and randomly surveyed 20 households within each division.

   Target population:

   Sample:

   Variable:

   Type of Bias:

2.4 Study design

The two main study designs we will cover are observational studies and experiments. Both the sampling method and the study design will help to determine the **scope of inference** for a study. Remember that only in a randomized experiment can we conclude a **causal** (cause and effect) relationship between the explanatory and response variable.

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

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

Inferences to population can be made

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

Can draw cause-and-effect conclusions

Can only discuss association due to potential confounding variables

For the next exercises, identify the explanatory variable, the response variable, a potential confounding variable, and the study design.

6. The pharmaceutical company, Moderna Therapeutics is working in conjunction with the National Institute of Health towards a vaccine for COVID-19 and has recently begun Phase 3 clinical trials. US Clinical research sites will enroll 30,000 volunteers without COVID-19 to participate. Participants will be randomly assigned to receive either the candidate vaccine or a saline placebo. They will then be followed to assess vaccine related symptoms and development of COVID-19. The trial is blinded, so the investigators and the participants will not know who is assigned to which group.

   Explanatory Variable:

   Response Variable:

   Confounding Variable:

   Study design:

   What is the scope of inference for this study?

7. 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 1 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:

   Confounding Variable:

   Study design:

   What is the scope of inference for this study?

8. What is a potential confounding variable for the study in question 7? Explain how this meets the definition of a confounding variable.

2.5   Additional notes

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

Current Population Survey

3.1 Learning outcomes

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

- Plots for a single categorical variable: bar plot

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

- Recognize and simulate probabilities as long-run frequencies

- Construct two-way tables to evaluate conditional probabilities

3.2 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

- Probability

- Conditional probability

- Two-way tables

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

3.3 "Current" Population Survey: 1985

The data set we will use for this activity is from the Current Population Survey in 1985. The CPS is a survey sponsored by the Census Bureau and the Bureau of Labor Statistics to track labor force statistics for the United States population. The following table summarizes the data:

Variable	Description
educ	Number of years of education
south	Indicator variable for living in a southern region: S = lives in south, NS = does not live in south
sex	Gender: M = male, F = female
exper	Number of years of work experience (inferred from age and education)
union	Indicator variable for union membership: Union or Not
wage	Wage (dollars per hour)
age	Age (years)
race	Race: W = white, NW = not white
sector	Sector of the economy: clerical, const (construction), management, manufacturing, professional, sales, service, other
married	Marital status: Married or Single

3.3.1 Vocabulary review

1. What are the observational units?

2. Which variables are categorical?

3. What types of plots can be used to display categorical data?

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.

3.3.2 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, and tell R to treat the sector and sex variables as categorical variables ("factors").

```
cps <- read.csv("data/cps.csv") #This will read in the dataset
cps$sector <- factor(cps$sector) #When a variable is categorical, need to make it a factor
cps$sex <- factor(cps$sex)
```

3.3.3 Displaying a single categorical variable

If we wanted to know how many people in our data set were in each sector, we would create a bar plot of the variable sector.
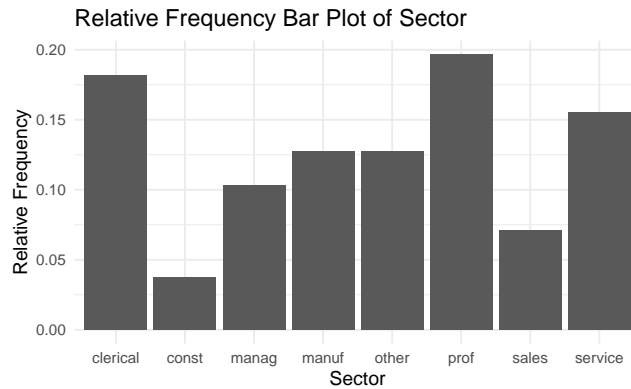
```
ggplot(data = cps,    #This specifies the dataset
       aes(y = sector)) +    #This specifies the variable
  geom_bar(stat = "count") +  #Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Sector",  #Give your plot a title
       x = "Frequency",    #Label the x axis
       y = "Sector")  + #Label the y axis
  coord_flip()  #Turn the bars so they are vertical
```
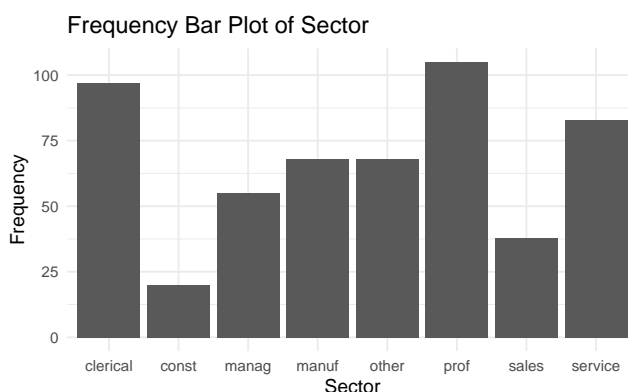

4. Which Sector has the largest number of people in it?

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 sector by the sample size. These are sample proportions.

```
ggplot(data = cps,    #This specifies the dataset
       aes(x = sector)) +    #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 Sector",  #Give your plot a title
       x = "Sector",    #Label the x axis
       y = "Relative Frequency")  #Label the y axis
```

Relative Frequency Bar Plot of Sector

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

3.3.4 Displaying two categorical variables

To see the differences in proportion of each sector between males and females we would create a segmented bar plot of sector segmented by sex.

```r
ggplot(data = cps,    #This specifies the dataset
       aes(x = sector, fill = sex)) +    #This specifies the variables
  geom_bar(stat = "count", position = "fill") +  #Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Sector by Sex",  #Make sure to title your plot
       x = "Sector",    #Label the x axis
       y = "") +   #Remove y axis label
    scale_fill_grey()  #Make figure black and white
```

Segmented Bar Plot of Sector by Sex
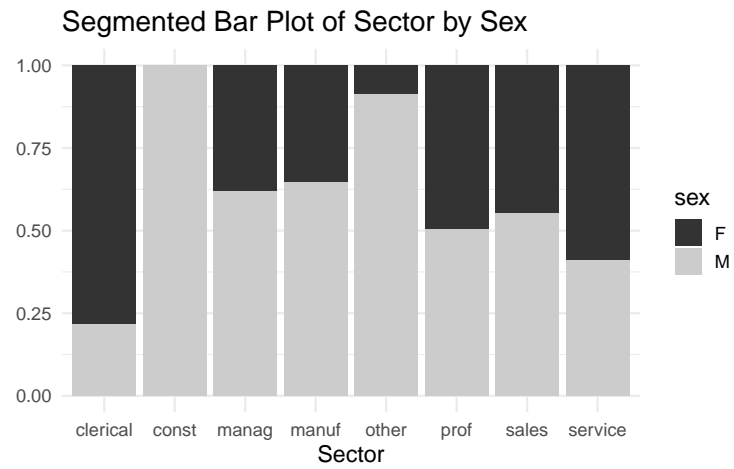
6. Using the segmented bar plot, which sector has about the same proportion of males and females?

7. Which sector has the highest proportion of females?

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

3.4 Probability

9. A study was reported in which ninth grade Minnesota teens were asked whether they had gambled at least once a week in the past year. The sample consisted of 49.1% boys. The proportion of boys who had gambled at least once per week during the past year was 0.229, while among non-boys this proportion was only 0.045.

Let B = the event the person is a boy, and C = the event the person is a weekly gambler.

a. Draw a segmented bar plot of sex segmented by gambling.

b. Identify what each numerical value represents in probability notation.

$0.491 =$

$0.229 =$

$0.045 =$

c. Create a two-way hypothetical table to represent the situation. Recall that in a two-way table, the explanatory variable should be your column headers (similar to the x-axis in a segmented bar graph!) while the response variable becomes the row headers.

		Total
Total		100,000

d. Find P(B and C). What does this probability represent in the context of the problem?

e. Find the probability that a selected non-gambler is a non-boy. What is the notation used for this probability?

10. In a computer store, 30% of the computers in stock are laptops and 70% are desktops. Five percent of the laptops are on sale, while 10% of the desktops are on sale. Let L = the event the computer is a laptop, and S = the event the computer is on sale.

a. Identify what each numerical value represents in probability notation.

   $0.30 =$

   $0.70 =$

   $0.05 =$

   $0.10 =$

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

	Total
Total	100,000

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

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

3.5 Additional notes

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

IMDb Movie Reviews

4.1 Learning objectives

- 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, inter-quartile range

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

4.2 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, inter-quartile range (IQR)

- Types of graphs: boxplots, dot plots, histograms

To review these concepts see Section 2.3 in the textbook.

4.3 Movies released in 2016

A data set was collected on movies released in 2016. Here is a list of some of the variables collected on these movies.

- Year: Year the movie was released

- Budget: The amount of money (in US $ millions) budgeted for the production of the movie

- Revenue: The amount of money (in US $ millions) the movie made after release

- Duration: The length of the movie (in minutes)

- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)

- IMDb Score: User rating score from 1 to 10

- Genre: Category the movie falls into

- Movie Facebook Likes: Number of likes a movie receives on Facebook

4.4 Vocabulary review

1. What are the observational units in this data set?

2. Which of the above listed variables are categorical?

3. Which of the above listed variables are quantitative?

4.5 Summarizing a single quantitative variable

The `favstats` function gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable 'IMDb'.

```
favstats(movies$imdb_score)   # [data set name]$[variable name]
#>   min  Q1 median  Q3 max     mean        sd  n missing
#>   3.4 5.9    6.6 7.1 8.2 6.459016 0.9218418 61       0
```

4. Give the values for the two measures of center.

5. Calculate the IQR.

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

4.6 Displaying a single quantitative variable

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

A histogram of the variable 'IMDb Score' is shown below. 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 of 4.75 will fall into the bin for 4.75 - 5.25. Visually this shows us the range of IMDb scores for Movies released in 2016.

```
ggplot(data = movies,    #Name data set
       aes(x = imdb_score)) +    #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
```
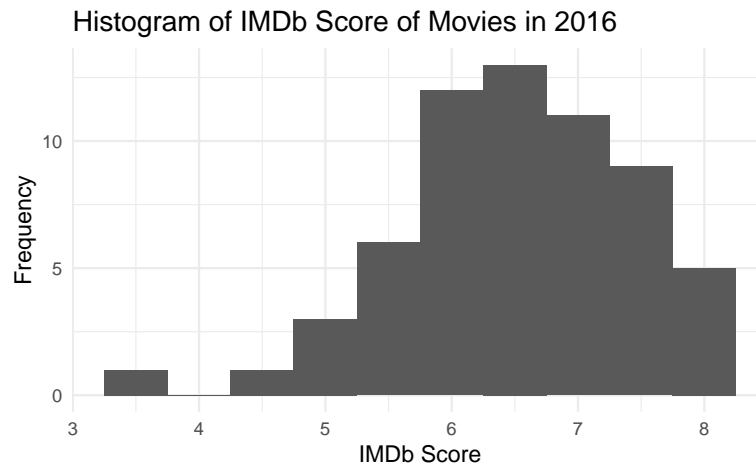

8. Which range of IMDb scores have the highest frequency?

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

The boxplot is created using the five number summary:

- Minimum value

- First quartile (Q1) - the 25th percentile

- Median - the 50th percentile

- Third quartile (Q3) - the 75th percentile

- Maximum value

10. The three smallest IMDb scores in the data set are 3.4, 3.5, and 3.7 and the three largest IMDb scores are 8.5, 8.7, and 9.1. Using the summary statistics above, sketch a boxplot of IMDb Score. Be sure to label the axes.

4.7 Displaying a single categorical and single quantitative variable

The boxplot of 'Budget' in millions by 'Content rating' is plotted using the code below. This plot helps to compare the budget for different levels of content rating.

```
ggplot(data = movies,  #Data set
       aes(y = budget_mil, x = content_rating))+  #Identify variables
  geom_boxplot()+  #Tell it to make a boxplot
  labs(title = "Side by side Boxplot of Budget by Content Rating",  #Title
       x = "Content Rating",   #x-axis label
       y = "Budget (in Millions)")  #y-axis label
```


Side by side Boxplot of Budget by Content Rating
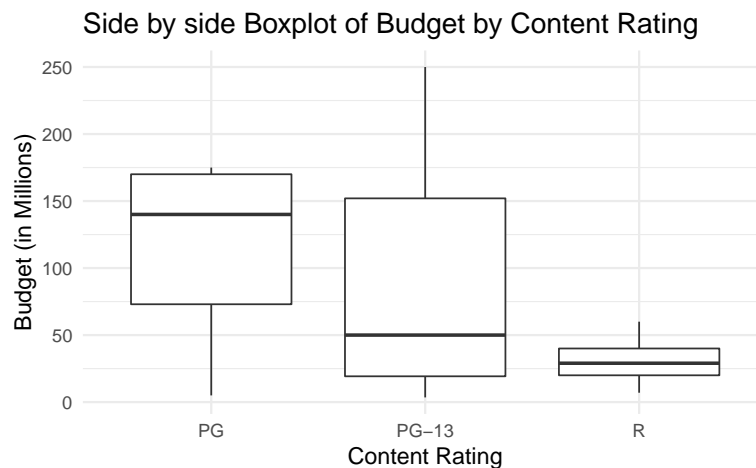
11. Answer the following questions about the boxplots above.

a. Which content rating has the highest center?

b. Which content rating has the largest spread?

c. Which content rating is the most symmetric distribution?

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

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

4.8 Additional notes

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

Movie Profits

5.1 Learning objectives

- 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 correlation coefficient

- Find the estimated line of regression using summary statistics and `R` linear model (`lm`) output

- Interpret the slope coefficient in context of the problem

- Interpret the coefficient of determination in context of the problem

5.2 Terminology review

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

- Scatterplot

- Correlation

- Slope

- Least-squares line of regression

- Coefficient of determination (r-squared)

To review these concepts see Chapter 3 in the textbook.

5.3 Movies released in 2016

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

- Year: Year the movie was released

- Budget: The amount of money (in US $ millions) budgeted for the production of the movie

- Revenue: The amount of money (in US $ millions) the movie made after release

- Duration: The length of the movie (in minutes)

- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)

- IMDb Score: User rating score from 1 to 10

- Genre: Category the movie falls into

- Movie Facebook Likes: Number of likes a movie receives on Facebook

5.3.1 Vocabulary review

1. What type of plot is used to display two quantitative variables?

2. What summary statistics are used to describe the relationship between two quantitative variables?

We will look at the relationship between 'Budget' and 'Revenue' for movies released in 2016. This shows a scatterplot of 'Budget' as a predictor of 'Revenue' (note: both variables are measures in "millions of dollars").

```
ggplot(data = movies,    #This is the data set
       aes(x = budget_mil, y = revenue_mil))+  #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 tile your plots
  geom_smooth(method = "lm", se = FALSE)  #Add regression line
```
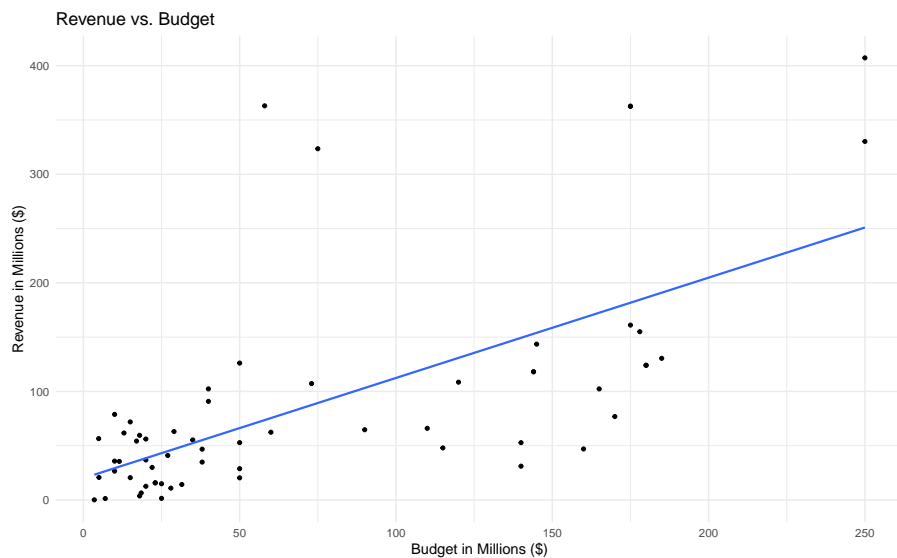
3. Assess the four features of the scatterplot that describe this relationship. Describe each feature using a complete sentence!

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual observations or outliers

4. Does there appear to be an association between 'Budget' and 'Revenue'? Explain.

5.3.2 Correlation

Correlation measures the strength and the direction between two quantitative variables. The closer the value of correlation to + 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.

```
# Take subset of variables
movies %>%
  select(c("budget_mil", "revenue_mil",  # Take subset of variables
          "duration", "imdb_score",
          "movie_facebook_likes")) %>%
  cor(use="pairwise.complete.obs") %>% # Calculate correlation matrix
  round(3)  # Round to 3 decimals
```

```
#>                       budget_mil revenue_mil duration imdb_score
#> budget_mil                1.000       0.647    0.527       0.308
#> revenue_mil               0.647       1.000    0.252       0.488
#> duration                  0.527       0.252    1.000       0.236
#> imdb_score                0.308       0.488    0.236       1.000
#> movie_facebook_likes      0.648       0.671    0.562       0.346
#>                       movie_facebook_likes
#> budget_mil                           0.648
#> revenue_mil                          0.671
#> duration                             0.562
#> imdb_score                           0.346
#> movie_facebook_likes                 1.000
```

5. Using the output above, which two variables have the strongest correlation?

6. What is the value of correlation between 'Budget' and 'Revenue'?

7. Based on the value of correlation what would the sign of the slope be? Positive or negative? Explain.

8. Does your answer to question 7 match the direction you choose in question 3?

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

5.3.3 Slope

The slope measures the change in y for each increase in x by 1. In other words, as the x variable increases by 1 unit, the y variable changes (increases/decreases) by the value of slope.

The linear model function in R 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 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) 20.0362329 14.3458255 1.396659 1.677479e-01
#> budget_mil   0.9236972  0.1418579 6.511426 1.806269e-08
```

10. Write out the least squares line using the summary statistics provided.

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

12. Using the least squares line from Question 10, predict the revenue for a movie with a budget of 165 million.

5.3.4 Residuals

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

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\text{Residual} = \text{Data - Model}$$

$$e_i = y_i - \hat{y}_i$$

13. The movie, *Independence Day: Resurgence*, had a budget of 165 million and revenue of 102.315 million. Find the residual for this movie.

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

5.3.5 Coefficient of determination (squared correlation)

The coefficient of determination, r^2, can also be used to describe the strength of the linear relationship between two quantitative variables. r^2 measures the proportion of variation in the response that is explained by the least squares line with the explanatory variable.

15. Use the correlation, r, to calculate the coefficient of determination between 'Budget' and 'Revenue', r^2.

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

5.3.6 Multivariate plots

What if we wanted to see if the relationship between 'Budget' and 'Revenue' differs if we add another variable into the picture? The following plot visualized three variables, creating a **multivariate** plot.

```r
ggplot(data = movies,    #This is the data set
       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
```
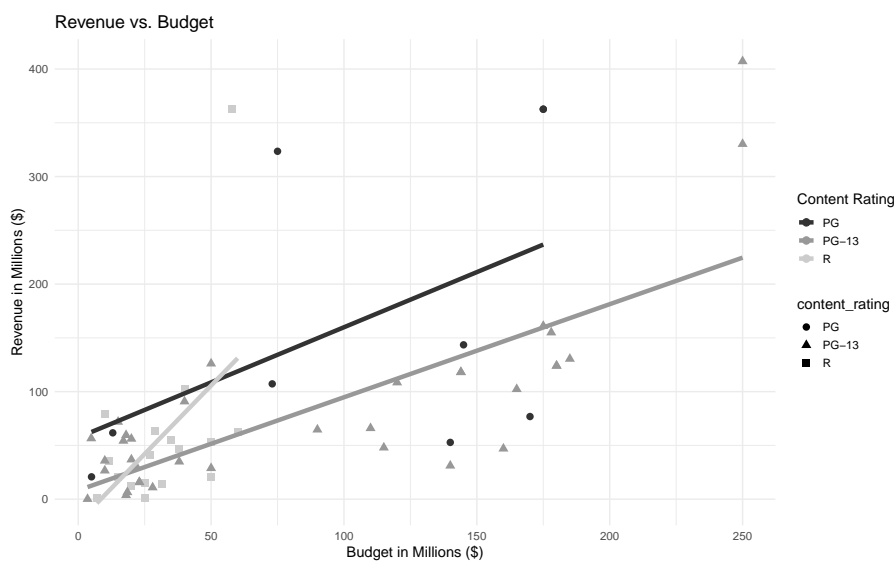

25. Identify the three varables plotted in this graph.

26. Does the relationship between 'Budget' and 'Revenue' differ among the different content ratings? Explain.

5.4   Additional notes

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

Handedness of Male Boxers

6.1   Learning objectives

- Identify the two possible explanations (one assuming the null hypothesis, and one assuming the alternative hypothesis) for a relationship seen in sample data

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

- Describe and perform simulation-based hypothesis tests for a single proportion

- Interpret and evaluate a p-value

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

6.2   Terminology review

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

- Parameter of interest

- Null hypothesis

- Alternative hypothesis

- Simulation

- Null distribution

- p-value

- Bootstrapping

- Confidence interval

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

6.3 Steps of the statistical investigation process

We will work through a six-step process to complete a hypothesis test for a single proportion.

- **Ask a research question** that can be addressed by collecting data. What are the researchers trying to show?

- **Design a study and collect data**. This step involves selecting the people or objects to be studied and how to gather relevant data on them.

- **Summarize and visualize the data**. Calculate summary statistics and create graphical plots that best represent the research question.

- **Use statistical analysis methods to draw inferences from the data**. Choose an analysis technique appropriate for the data and identify the p-value. In this study, we will focus on using randomization.

- **Communicate the results and answer the research question**. Using the p-value and confidence interval from the analysis, determine whether the data provide statistical evidence against the null hypothesis.

- **Revisit and look forward** to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study

6.4 Handedness of male boxers

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

6.4.1 Summary statistics review

1. What are the observational units?

2. What variable are we testing? Is it categorical or quantitative?

3. What type of plot would be appropriate to visually display the data?

4. Write out in context the statistic will we calculate to summarize the data.

6.4.2 Ask a research question

5. Identify the research question for this study.

6.4.3 Design a study and collect data

6. What is the target population for this study?

7. Did the researchers use a biased or an unbiased method of selection? Explain your answer.

6.4.4 Summarize and visualize the data

```
# Counts for Handedness
#Tally creates a table with a count for each level of the variable
tally(~Stance, data=handedness_sub, margins=T)
```

```
#> Stance
#>  left-handed right-handed       Total
#>          81          419         500
```

8. Calculate the appropriate summary statistic that represents the research question. Use appropriate notation.

6.4.5 Use statistical analysis methods to draw inferences from the data

When testing data we must first identify the null hypothesis. The null hypothesis is written about the parameter of interest, the true value of interest.

9. Write out the parameter of interest for this study. (Hint: the true proportion of....)

10. Using the parameter of interest in question 9, write out the null hypothesis in words.

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

11. Write the null hypothesis in notation using the null value of 0.1 in place of π_0.

The **alternative hypothesis** is the claim to be tested and the direction is based on the research question.

12. Based on the research question from question 5, are we testing that the parameter is greater than 0.1, less than 0.1 or different than 0.1?

13. Write out the alternative hypothesis in words.

14. Write out the alternative hypothesis in notation.

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

- The true proportion of male boxers who are left handed is 0.1 and our results just occurred by random chance or

- The true proportion of male boxers who are left handed is greater than 0.1 and our results reflect this

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

The null distribution is created under the assumption the null hypothesis is true. In this case, we assume the true proportion of male boxers who are left handed is 0.1 so we will create 1000 different simulations of 500 boxers under this assumption.

Let's think about how to use cards to create one simulation of 500 boxers under the assumption the null hypothesis is true. Suppose blue cards represents left-handed and red cards represents right-handed.

15. How many cards total do we need? How many blue ones? How many red ones?

16. Next, we would mix the cards together and draw 1 card, write down if it's red or blue, and replace the card. How many times would we need to repeat this process to simulate our sample?

17. Once we have one simulated sample, what would we calculate and plot on the null distribution?

We will use the computer to simulate 1000 different samples of 500 male boxers, plotting the proportion who are left handed in each sample, based on the assumption that the true proportion of male boxers who are left handed is 0.1. This is called the null distribution because it is created based on the assumption that the null hypothesis is true.

To use the computer simulation, we will need to enter the "probability of success," "sample size," "number of repetitions," "as extreme as," and the "direction" (matches the direction of the alternative hypothesis).

18. What values should be entered into the simulation?

- Probability of success:

- Sample size:

- Number of repetitions:

- As extreme as:

- Direction:

The following R code produced the null distribution with 1000 simulations that follows. Check that your answers to question 18 match the code inputs.

```r
one_proportion_test(probability_success = 0.1, #Null hypothesis value
                    sample_size = 500, #Enter sample size
                    number_repetitions = 1000, #Enter number of simulations
                    as_extreme_as = 81/500, #observed statistic
                    direction = "greater", #specify direction of alternative hypothesis
                    report_value = "proportion") #Reporting proportion or number of successes?
```


Proportion of successes
Proportion of Samples = 0/1000 = 0
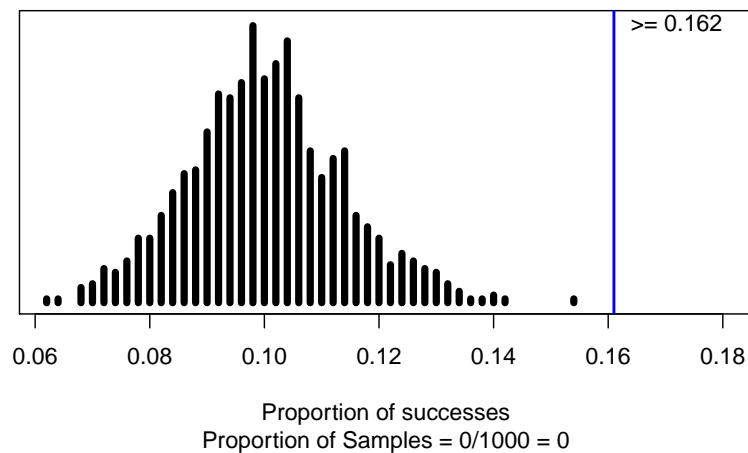
19. Around what value is the null distribution centered? Why does that make sense?

20. Where does the statistic (value from question 8) fall in the null distribution? Is it towards the center or in one of the tails?

21. Is the statistic likely to happen or unlikely to happen if the true proportion of male boxers is 0.1? Explain your answer.

22. Using the simulation, what is the proportion of samples at this summary statistic or greater, if the true proportion of male boxers is 0.1? *Hint: Look under the simulation.*

This is the **p-value**. The smaller the p-value the more evidence we have against the null hypothesis.

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

Strength of Evidence Against the Null

| Very strong	Strong	Moderate	Some	Little to no

| 1% | | 5% | | 10% | |
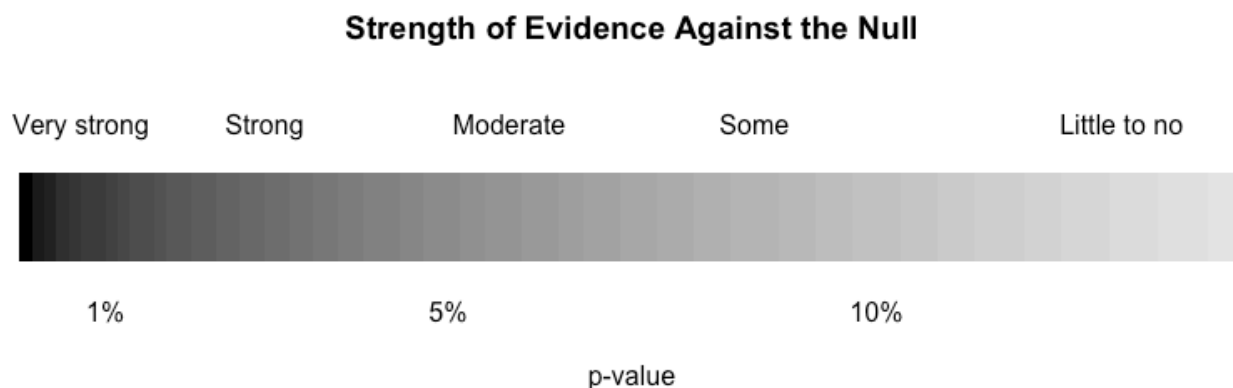
p-value

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

We will use bootstrapping to find the 95% confidence interval.

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

The following `R` code produced the following bootstrap distribution with 1000 simulations.

```
one_proportion_bootstrap_CI(sample_size = 500, #Sample size
                   number_successes = 81, #Observed number of successes
                   number_repetitions = 1000, #Number of bootstrap samples to use
                   confidence_level = 0.95) #Confidence level as a decimal
```

49

Bootstrapped values of the proportion
95% CI: (0.13, 0.196)
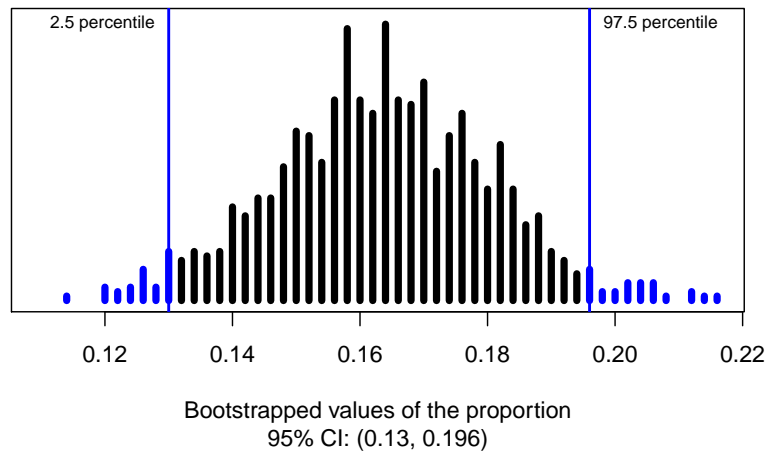
25. What is the value at the center of this distribution? Why does this make sense?

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

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

28. Interpret the 95% confidence interval in context.

6.4.6 Communicate the results and answer the research question

When we write a conclusion we answer the research question by stating how much evidence there is for the alternative hypothesis.

29. Write a paragraph summarizing the results. Be sure to include:

- Summary statistic
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- Generalization - to what group do the results apply?

6.4.7 Revisit and look forward

30. Suggest a new research question that you might investigate, building on what you learned in this study.

6.5 Additional notes

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

Winter Sports Helmet Use and Head Injuries

7.1   Learning objectives

- Write out the null and alternative hypothesis for two categorical variables

- Assess the conditions to use the standard normal distribution for a difference in proportions

- Calculate the Z test statistic for the difference in proportions

- Find the p-value and assess the strength of evidence

- Create and interpret a confidence interval for the difference in proportions

7.2   Terminology review

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

- Conditional proportion

- Z test

- $z*$ multiplier

- Null hypothesis

- Alternative hypothesis

- Test statistic

- Standard normal distribution

- Independence and Success-Failure conditions

- Type 1 and Type 2 errors

- Decisions

To review these concepts see Chapter 5 in your textbook.

7.3 Helmet Use and Head Injuries

In "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders" by Sullheim et. al., in the Journal of the American Medical Association, Vol. 295, No. 8, we can see the results from a random sample 3562 skiers and snowboarders involved in accidents.

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

Is there evidence that safety helmet use reduces the risk of head injury for skiers and snowboarders?

7.3.1 Vocabulary review

1. What is the explanatory variable?

2. What is the response variable?

3. Is this an experiment or observational study?

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

		Study Type	
		Randomized Experiment	Observational Study
Selection of Cases	Random Sample		
	No Random Sample		

5. What is the conditional proportion of skiers/snowboarders with a head injury that wore a helmet?

6. What is the conditional proportion of skiers/snowboarders with a head injury that did not wear a helmet?

7.3.2 Ask a research question

In this study we are looking at the relationship between two groups or two parameters (π_1 and π_2). Remember we define the parameter as the true proportion of observational units that represent the variable of interest.

7. What is the variable of interest in this study?

8. Write the two parameters of interest for this study. Let 1 = skier/snowboarder wore helmet, 2 = skier/snowboarder did not wear helmet.

   π_1 -

   π_2 -

When comparing two groups, we assume the two parameters are equal in the null hypothesis—there is no association between the variables.
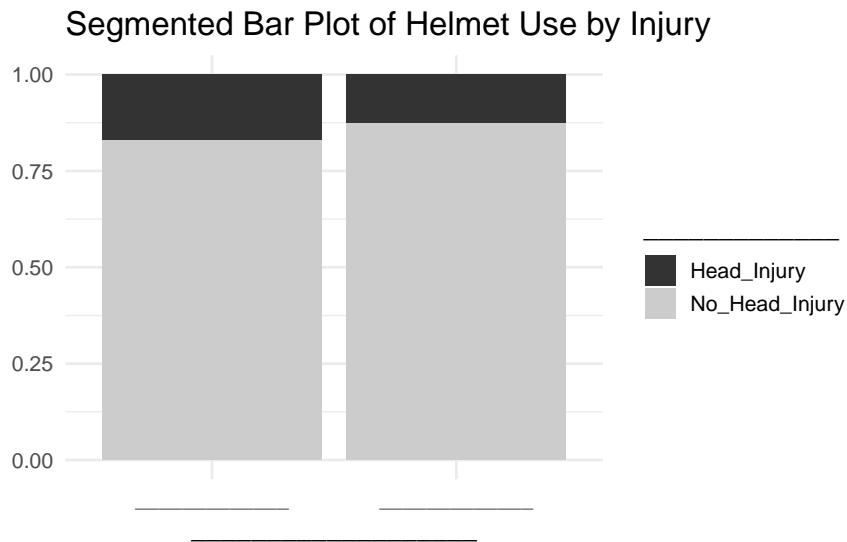
9. Write the null hypothesis out in words using your answers to question 8.

10. What is the research question?

11. Based on the research question fill in the appropriate sign for the alternative hypothesis:

$$H_A : \pi_1 - \pi_2 \underline{\hspace{3cm}} 0$$

7.3.3 Summarize and visualize the data

Segmented Bar Plot of Helmet Use by Injury

12. Fill in the blanks on the graph with the appropriate variables and values to complete the segmented bar plot showing the proportion of head injuries between those who use helmets and those who do not use helmets. *Hint: use the conditional proportions from questions 5 and 6.

13. Based on the bar plot, Does there appear to be an association between helmet use and head injury? Explain.

14. Calculate the point estimate for this study. Use helmet use minus no helmet use as the order of subtraction.

15. What is the notation used for the value calculated in question 14?

7.3.4 Use statistical analysis methods to draw inferences from the data

To test the null hypothesis we could use simulation methods as we did with a single categorical variable. In this activity we will focus on theory-based methods. Like with a single proportion, the difference in proportions can be mathematically modeled using the normal distribution if certain conditions are met.

Conditions for the sample distribution of $\hat{p}_1 - \hat{p}_2$

- Independence: The data are independent within and between the two groups.

- Success-Failure Condition: The success-failure condition holds for each group.

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

17. Is the success-failure condition met for each group? Explain your answer.

To calculate the test statistic we use:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE}$$

where the standard error is calculated using the pooled proportion of successes.

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})(\tfrac{1}{n_1} + \tfrac{1}{n_2})}, \text{where}$$

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

18. Calculate the $SE(\hat{p}_1 - \hat{p}_2)$.

19. Calculate the test statistic.

We will use the pnorm function in R to find the p-value. Use the provided R markdown file and enter the value of the test statistic at xx.

```
pnorm(xx #enter value of test statistic
  , m=0, s=1 #using the standard normal mean = 0, sd = 1
  , lower.tail=TRUE) # gives a p-value less than the test statistic
```

20. Report the p-value.

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

To find a confidence interval for the difference in proportions we will add and subtract the margin of error from the point estimate to find the two endpoints.

$$\hat{p}_1 - \hat{p}_2 \pm z^* SE(\hat{p}_1 - \hat{p}_2), \text{where}$$

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

Note that the formula changes when calculating the variability around the statistic in order to calculate a confidence interval! Here, use the sample proportions for each group to calculate the standard error for the difference in proportions.

The z^* multiplier is found under the Standard normal distribution. We find the values that encompass the middle 95% of the distribution. If 95% of the Standard normal distribution should be in the middle, that leaves 5% in the tails, or 2.5% in each tail. The qnorm function in R will tell us the z^* value for the desired percentile (in this case, $0.95 + 0.025$, 97.5% percentile).

```
qnorm(0.95 + 0.025) #multiplier for 95% confidence interval
```

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

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

23. Using the multiplier of $z^* = 1.96$, calculate the 95% confidence interval for the difference in true proportion of head injuries for those that used helmets minus those who did not.

24. Interpret the confidence interval found in question 23 in context of the problem.

25. Write a paragraph summarizing the results of the study. Be sure to include:

- Summary statistic

- Test statistic and interpretation

- P-value and interpretation

- Conclusion (written to answer the research question)

- Confidence interval and interpretation

- Scope of inference

7.3.5   Types of errors

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

		Test Conclusion	
		Reject H_0	Fail to reject H_0
Truth	H_0 true	good decision	Type 1 Error
	H_A true	Type 2 Error	good decision

A Type 1 Error is rejecting the null hypothesis when H_0 is actually true. A Type 2 Error is failing to reject the null hypothesis when the alternative is actually true.

26. Using a significance level of 0.05, what decision do you make in regards to the null hypothesis?

27. What type of error could we have made?

28. Write this error in context of the problem.

7.4 Additional notes

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

COVID-19 and Air Pollution

8.1 Learning outcomes

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

- Describe and perform a simulation-based hypothesis test for paired quantitative data

- Interpret and evaluate a p-value

- Find a confidence interval for the mean difference using bootstrapping

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

8.2 Terminology review

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

- Mean difference

- Paired data

- Independent groups

- Shifted null distribution

To review these concepts see Section 6.2 in the textbook.

8.3 COVID-19 and air pollution

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

Have high population-density U.S. cities seen the same improved air quality conditions? To study this question, data was gathered from the U.S. Environmental Protection Agency (EPA) AirData website which records the ozone (O3) and fine particulate matter (PM2.5) values for cities across the U.S. These measures are used to

Figure 8.1: The India Gate in New Delhi, India

calculate an air quality index (AQI) score for each city each day of the year. Thirty-three of the most densely populated U.S. cities were selected and the AQI score recorded for April 20, 2020 as well as the five-year median AQI score for April 20th (2015 - 2019). Note that higher AQI scores indicate worse air quality.

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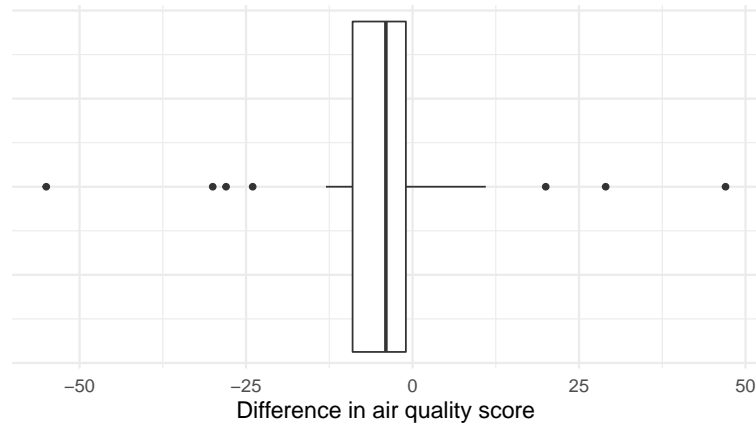
	Mean	Standard deviation	Sample size
Current	$\bar{x}_1 = 47.394$	$s_1 = 14.107$	$n_1 = 33$
5 Year Median	$\bar{x}_2 = 51.545$	$s_2 = 17.447$	$n_2 = 33$
Differences	$\bar{x}_d = $ -4.152	$s_d = 17.096$	$n_d = 33$

8.3.1 Vocabulary review

1. What is the sample size?

2. Identify the variables in this study. What role do each have?

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

4. Is this an experiment or observational study?

8.3.2 Ask a research question

5. What are the two competing possibilities to run a hypothesis test for this study?

6. Write the null hypothesis in words.

7. What is the research question?

8. Write the alternative hypothesis in notation.

8.3.3 Summarize and visualize the data

9. Report the summary statistic for the data.

10. What notation is used for the value in question 9?

8.3.4 Use statistical inferential methods to draw inferences from the data

To simulate the null distribution we will use a bootstrapping method - sampling with replacement from the data set. Before bootstrapping we will need to shift each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the simulated null distribution will be centered at the null value.

11. Calculate the difference $\mu_0 - \bar{x}$. Will we need to shift the data up or down?

12. Use the provided `R` markdown file and enter the calculated value from question 11 for xxx to simulate the null distribution and enter the summary statistic from question 9 for xxxx to find the p-value.

```
paired_test(data = Air$Difference,    #Vector of differences or data set with column for each group
            shift = xxx,    #Shift needed for bootstrap hypothesis test
            direction = "less",  #Direction of alternative
            as_extreme_as = xxxx,  #Observed statistic
            number_repetitions = 1000,  #Number of draws for null distribution
            which_first = 1)  #Not needed when using calculated differences
```

13. Sketch the null distribution created in Question 12 here.

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

15. What proportion of samples are beyond the sample mean difference in AQI Scores for current scores minus 5 year median scores?

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

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

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

The following `R` code creates a bootstrap distribution showing 1000 simulations of the mean difference.

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


Bootstrap Mean Difference
99% CI: (−12.091, 3.546)
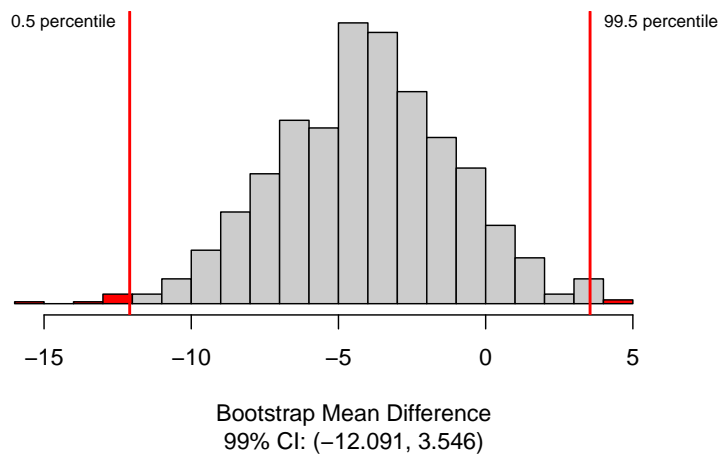
19. Use the bootstrapped distribution above to find a 99% confidence interval for the parameter of interest. Report the confidence interval in interval notation.

8.3.5 Communicate the results and answer the research question.

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

21. Write a paragraph summarizes the results of this study. Be sure to include:

- Summary statistic

- P-value and interpretation

- Conclusion (written to answer the research question)

- Confidence interval and interpretation

- Scope of inference

8.3.6 Revisit and look forward

22. Would it be possible to design an experiment to determine if the changed human behavior due to the COVID-19 pandemic causes a decrease in air pollution? Explain.

8.4 Additional notes

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

Weather Patterns and Record Snowfall

9.1 Learning objectives

- Write out the null and alternative hypothesis for one categorical and one quantitative variable
- Calculate and carry-out simulation based hypothesis test for a difference in means
- Interpret and evaluate a p-value
- Find a bootstap confidence interval for a difference in means
- Use a confidence interval to determine the conclusion of a hypothesis test

9.2 Terminology review

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

- Independent groups
- Difference in means

To review these concepts see Section 6.3 in the textbook.

9.3 Weather Patterns and Record snowfall

In the winter of 2018-2019, Bozeman had a record snowfall which resulted in the collapse of two flat-roofed buildings on the MSU campus. A writer for the Washington Post predicted the heavy snowfall for 2018-2019 due to the El Niño weather pattern that occurred in that season. A meteorologist in Montana wanted to see if the weather pattern really was associated with total snowfall. She obtained historical data from 44 years on the weather pattern (El Niño or La Niña) and snowfall (in inches) at the Billings Weather Station.

```
ggplot(data = Snow,
       aes(x = WeatherPattern, y = Snowfall)) +
    geom_boxplot() +
    labs(title = "Snowfall by weather pattern",
         x = "Weather pattern") +
    coord_flip()
```

Snowfall by weather pattern
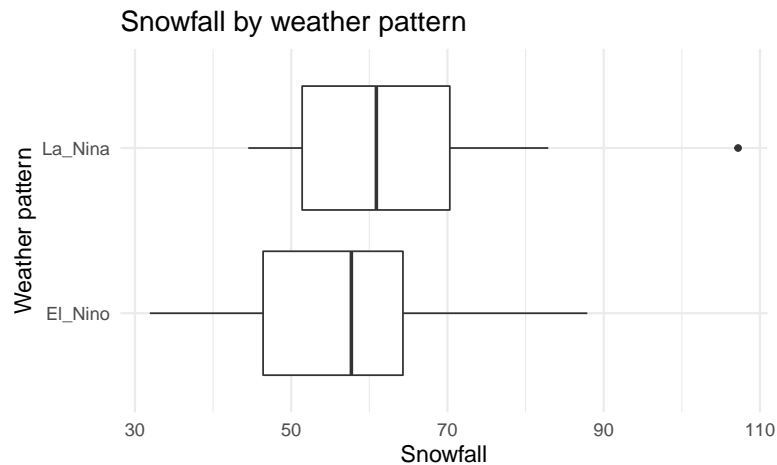
```
favstats(Snowfall~WeatherPattern, data=Snow)
```

```
#>   WeatherPattern  min   Q1 median   Q3   max     mean       sd  n missing
#> 1        El_Nino 31.9 46.4   57.7 64.3  87.9 56.23043 13.00823 23       0
#> 2        La_Nina 44.5 51.4   60.9 70.3 107.2 63.13333 15.48626 21       0
```

9.3.1 Quantitative variables review

1. The two variables assessed in this study are the type of weather pattern and snowfall. Identify the role for each variable (explanatory, response).

2. Which group (El Niño or La Niña) has the highest center? Explain which measure you are using.

3. Using the side-by-side boxplots, which group has the largest spread? How did you make that choice?

70

4. Is this an experiment or an observational study? Explain your reasoning.

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

9.3.2 Ask a research question.

6. Write out the parameter of interest in context of the study. Use proper notation and be sure to define your subscripts. Use El Niño minus La Niña as the order of subtraction.

7. What are the two competing possibilities we will evaluate in this study?

8. Identify which of your answers in question 7 is the null hypothesis and which is the alternative hypothesis.

9.3.3 Summarize and visualize the data

9. Calculate the summary statistic. Use El Niño minus La Niña as the order of subtraction. What is the appropriate notation for the statistic?

9.3.4 Use statistical inferential methods to draw inferences from the data

Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, we asssume there is no association between variables. This means that a snowfall value could be in either an El Niño year or a La Niña year.

To demonstrate this your instructor will use cards to represent the sample.

10. How many cards will we start with?

11. What will we write on each card?

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

13. What value is calculated from the cards and plotted on the null distribution?

14. Once we create a null distribution of 1000 simulations, at what value do you expect the distribution to be centered? Explain your answer.

Simulation method

15. Using the provided `R` markdown file, enter the values for the variables, dataset, first in subtraction, number of simulations, observed statistic, and direction of the alternative hypothesis.

```
two_mean_test(RESPONSE~PREDICTOR, data = DATASET,   #Variables and data
                      first_in_subtraction = "VALUE", #First value in order of subtraction
                      number_repetitions = ###,   #Number of simulations
                      as_extreme_as = ###,   #Observed statistic
                      direction = "??")   #Direction of alternative: "greater", "less", or "two-sided"
```

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

17. Using bootstrapping find a 90% confidence interval. Use the provided `R` markdown file. Enter the variables, first in subtraction, number of repetitions, and the confidence level.

```
two_mean_bootstrap_CI(RESPONSE~EXPLANATORY, data = DATASET,   #Variables and data
                      first_in_subtraction = "VALUE", #First value in order of subtraction
                      number_repetitions = ###,   #Number of simulations
                      confidence_level = ##)
```

18. Interpret the interval you calculated in Question 17.

9.3.5 Communicate the results and answer the research question

19. Write a paragraph summarizing the results of the study. Be sure to include:

- Summary statistic

- P-value and interpretation

- Conclusion (written to answer the research question)

- Confidence interval and interpretation

- Scope of inference

9.3.6 Revisit and look rorward

20. Would the results from the theory-based test match the results we saw with the simulation? Explain why or why not.

21. If we had data on 45 La Niña years and 47 El Niño years and found a similar summary statistic, what would happen to the p-value? The width of the confidence interval? The power of the test?

9.4 Additional notes

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

Hand Dexterity

10.1   Learning outcomes

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

- Describe and perform theory-based hypothesis tests for the slope

- Interpret and evaluate a p-value

- Construct and interpret a theory-based confidence interval for slope

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

10.2   Terminology review

In today's activity, we will use theory-based tests of inference for regression slope. Some terms covered in this activity are:

- Correlation

- Slope

- Regression line

To review these concepts see Chapters 3 and 7 in the textbook.

10.3   Hand dexterity

Physical therapists often evaluate manual (hand) dexterity by having patients complete simple tasks, such as moving pegs on a board or threading objects through holes. Researchers want to examine the manual dexterity of children as part of a follow-up study of a test originally designed for adults to see how manual dexterity changes with age. In this test, 174 participants were given a board with 16 pegs, each in their own hole, arranged in a 4x4 grid. Participants were instructed to pick up the peg with one hand, flip it over by rotating their wrist, then reinsert it in the same hole. Using this test, researchers want to know if as people age the speed at which they can flip all 16 pegs increases.

The variables in this dataset consist of the following:

- **time:** Recorded time to flip all 16 pegs, measured in seconds.

- **speed:** The average speed to flip a peg for each participant (seconds per peg).

- **age:** Age of the participants, measured in years.

- **dominant:** Whether the participant's dominant hand was used, coded as 0 for no, 1 for yes.

- **gender:** The participant's gender, recorded as a binary variable, 0 for male, 1 for female.

- **HD:** The dominant hand of the participant, recorded as "R" for right hand, "L" for left hand.

- **handUsed:** Which hand the participant used to complete the test, recorded as "R" for right hand, "L" for left hand.

Data source: Hand Dexterity in Children: Administration and Normative Values of the Functional Dexterity Test (FDT), Gogola, G., et al., 2013

10.3.1 Vocabulary review

1. Explain why regression methods are appropriate to use to address the researchers' question. Make sure you clearly define the variables of interest in your explanation and their roles.

2. What is the scope of inference for this study? Explain your answer.

3. Use the provided `R` markdown file to create a scatterplot to examine the relationship between the speed at which a participant can flip a peg and the age of the participant by filling in the variable names ('speed' and 'age') for xxx and xxxx. Provide this plot. Based on your plot, does it appear that there is a relationship between `age` and `speed`? Note: `age` should be on the x-axis.

```
ggplot(data = hands,   #This is the data set
   aes(x = xxx, y = xxxx))+  #Specify variables
geom_point() +  #Add scatterplot of points
labs(x = "Age (yrs)",  #Label x-axis
   y = "Speed (sec/peg)",  #Label y-axis
   title = "Scatterplot of Age vs. Speed") + #Be sure to tile your plots
geom_smooth(method = "lm", se = FALSE)  #Add regression line
```

4. Describe the features of the plot you created in Question 3.

If you indicated there are potential outliers, which points are they?

10.3.2 Conditions for the least squares line

When performing inference on a least squares line, the follow conditions are generally required

- Linearity: the data should follow a linear trend
- Nearly normal residuals: residuals must be nearly normal
- Constant variability: the variability of points around the least squares line remains roughly constant
- Independent observations: individual data points must be independent

The scatterplot and the residual plots will be used to assess the conditions for approximating the data with the t-distribution.
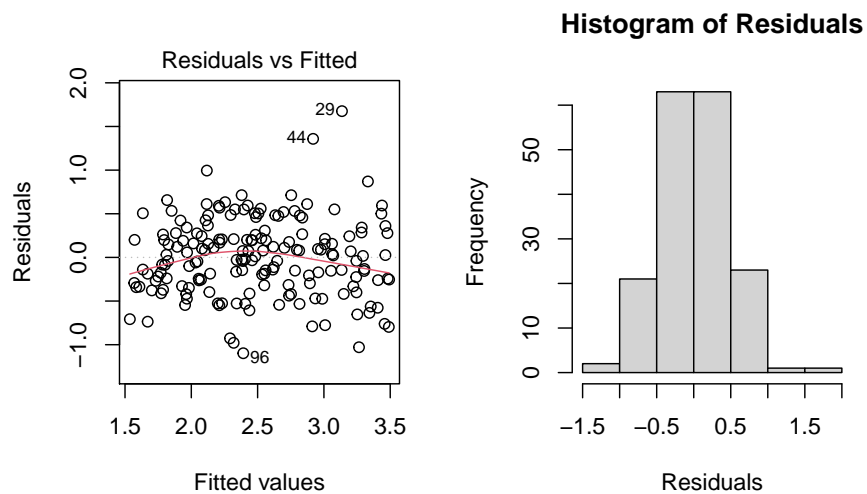
5. Are the conditions met to approximate the t-distribution?

77

10.3.3 Ask a research question

6. Write out the null hypothesis in words.

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

10.3.4 Summarize and visualize the data

Using the provided `R` markdown file, enter the response variable into the linear model function for xxx and the explanatory variable for xxxx to get the linear model output.

```
lm.hand <- lm(xxx~xxxx, data=hands) #lm(response~explanatory)
summary(lm.hand)$coefficients
```

8. Using the output from the evaluated `R` code above, write the equation of the regression line.

9. Interpret the slope in context of the problem.

10. Using your estimated line of best fit, predict the per peg speed for a participant who was 9.18 years old. Show all work.

11. Calculate the residual associated with the observation (9.18, 2.95), using your estimated regression line from question 8.

10.3.5   Use statistical inferential methods to draw inferences from the data

To find the value of the test statistic to test the slope we will use,

$$T = \frac{\text{slope estimate}}{SE} = \frac{b_1}{SE(b_1)}$$

We will use the linear model output above to get the estimate for slope and standard error.

12. Calculate the test statistic for slope. Identify where this calculated value is in the linear model output.

13. Interpret the test statistic in context of the problem.

14. Using the linear model output, report the p-value for the test of significance.

15. Based on the p-value, how much evidence is there against the null hypothesis?

Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^* SE(estimate)$$

$$b_1 \pm t^* SE(b_1)$$

The t^* multiplier comes from the t-distribution with $n - 2$ df.

```
qt(0.95+0.025, 172) #95% t* multiplier
#> [1] 1.973852
```

16. Calculate the 95% confidence interval for the true slope.

10.3.6 Communicate the results and answer the research question

17. Based on the p-value, write a conclusion in context of the problem.

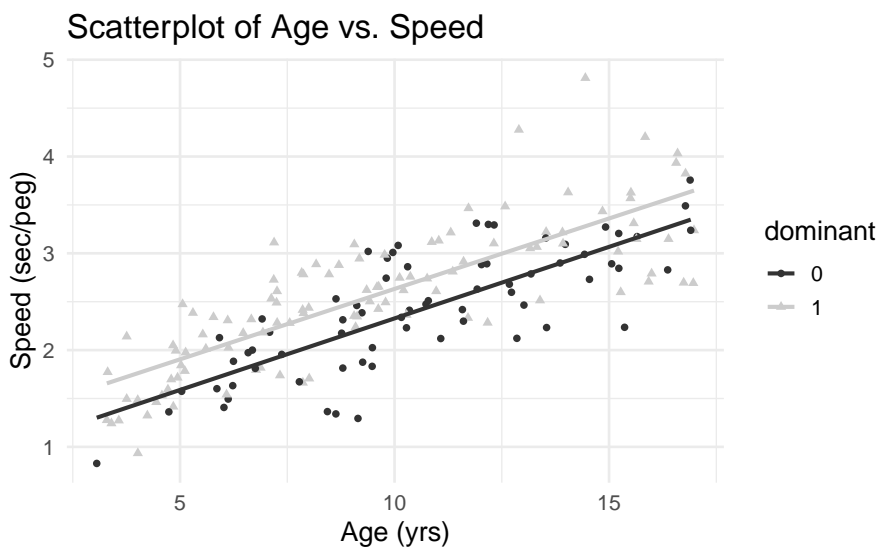18. Interpret the 95% confidence interval in context of the problem.

19. Summarize the results of the study in a written paragraph. Be sure to include.

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

10.3.7 Revisit and look forward

20. Is there an effect due to gender on the linear relationship between age and speed? Explain your answer using the scatterplot below.

```
ggplot(data = hands,    #This is the data set
       aes(x = age, y = speed, color = dominant))+  #Specify variables
  geom_point(aes(pch = dominant)) +  #Add scatterplot of points
  labs(x = "Age (yrs)",  #Label x-axis
       y = "Speed (sec/peg)",  #Label y-axis
       legend = "Dominant hand",  #Label your legend
       title = "Scatterplot of Age vs. Speed") + #Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE) +  #Add regression line
  scale_color_grey() #Make greyscale for printing
```


10.4 Additional notes

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