Stat 120

Deepak Bastola

2023-05-25

Contents

Introduction to Statistics

Welcome to the captivating world of statistics! This course will provide you with a solid foundation in statistical theory while taking you on a journey through the practical aspects of the subject. Along the way, you'll gain experience with statistical software, learn to interpret and effectively communicate statistical findings, and explore a range of topics in data analysis, statistical inference, and randomness. Some of the areas we'll delve into include linear regression, experimental design, normal distribution, sampling distributions, confidence intervals, and the bootstrap method.

Although statistics is a field that relies on mathematics, it stands apart as a distinct discipline. At the heart of this course is the ability to interpret results and grasp underlying concepts, rather than merely obtaining numerical outcomes. By engaging with a diverse set of problems, you'll become well-versed in statistical methodologies. However, it's crucial to remember that a deep understanding of the concepts is the key to drawing meaningful conclusions.

0.1 Learning Objectives

- Learn basic principles of data analysis, and how data is produced and used in studies and experiments.
- Understand role of variation and randomness. Understand principles of inference: confidence intervals and hypothesis tests.
- Develop ability to examine statistical arguments critically.
- Learn how to use software (R/RStudio) to analyze data, create graphs, perform basic statistical tests

Basics R

Chapter 1

What is R?

R is a free and open source statistical programming language that facilitates statistical computation. There are a myriad of application that can be done in R, thanks to a huge online support community and dedicated packages. However, R has no graphical user interface and it has to be run by typing commands into a text interface.

1.1 What is RStudio?

RStudio provides graphical interface to R! You can think of RStudio as a graphical front-end to R that that provides extra functionality. The use of the R programming language with the RStudio interface is an essential component of this course.

1.2 R Studio Server

The quickest way to get started is to go to https://maize.mathcs.carleton.edu, which opens an R Studio window in your web browser. Once logged in, I recommend that you do the following:

- Step 1: Create a folder for this course where you can save all of your work. In the Files window, click on New Folder.
- Step 2: Click on Tools -> Global Options -> R Markdown. Then uncheck the box that says "Show output inline…"

(It is also possible to download RStudio on your own laptop. Instructions may be found at the end of this document.)

1.3 R/RStudio

The use of the R programming language with the RStudio interface is an essential component of this course. You have two options for using RStudio:

- The **server version** of RStudio on the web at (https://maize.mathcs.carleton.edu). The advantage of using the server version is that all of your work will be stored in the cloud, where it is automatically saved and backed up. This means that you can access your work from any computer on campus using a web browser. This server may run slow during peak days/hours. I also recommend you to download a local version of R server in your computer in case of rare outages.
- A **local version** of RStudio installed on your machine. This option is highly recommended due to the computational resources this course demands. Using this version you can only store your files in your local machine. Additionally, we can save our work on GitHub. We will learn how to use GitHub in the beginning of the course. Both R and RStudio are free and open-source. Please make sure that you have recently updated both R and RStudio.

1.4 Installing R/RStudio (not needed if you are using the maize server)

Download the latest version of R: https://cran.r-project.org/

Download the free Rstudio desktop version: https://www.rstudio.com/products/rstudio/download/

Use the default download and install options for each. For R, download the "precompiled binary" distribution rather than the source code

Updating R/RStudio (not needed if you are using the maize server)

If you have used a local version of R/RStudio before and it is still installed on your machine, then you should make sure that you have the most recent versions of each program.

- To check your version of R, run the command `getRversion()` and compare your version to the newest version posted on https://cran.r-project.org/. If you need an update, then install the newer version using the installation directions above.
- In RStudio, check for updates with the menu option `Help > Check for updates`. Follow directions if an update is needed.

Did it work? (A sanity check after your install/update)

Do whatever is appropriate for your operating system to launch RStudio. You should get a window similar to the screenshot you see here, but yours will be more boring because you haven't written any code or made any figures yet!

Put your cursor in the pane labeled *Console*, which is where you interact with the live R process. Create a simple object with code like `x <- 2 * 4` (followed by enter or return). Then inspect the `x` object by typing `x` followed by enter or return. You should see the value `8` printed. If this happened, you've succeeded in installing R and RStudio!

1.5 Install LaTeX (for knitting R Markdown documents to PDF):

You need a Latex compiler to create a pdf document from a R Markdown file. If you use the maize server, you don't need to install anything. If you are using a local RStudio, you should install a Latex compiler. Below are the recommended installers for Windows and Mac:

- MacTeX for Mac (3.2GB)
- MiKTeX for Windows (190MB)
- Alternatively, you can install the `tinytex` R package by running `install.packages("tinytex")` in the console.

1.6 Updating R/RStudio (not needed if you are using the maize2 server)

If you have used a local version of R/RStudio before and it is still installed on your machine, then you should make sure that you have the most recent versions of each program.

- To check your version of R, run the command `getRversion()` and compare your version to the newest version posted on https://cran.r-project.org/. If you need an update, then install the newer version using the installation directions above.
- In RStudio, check for updates with the menu option `Help > Check for updates`. Follow directions if an update is needed.

1.7 Opening a new file

If using Rstudio on your computer, using the **File>Open File** menu to find and open this .Rmd file.

If using Maize Rstudio from your browser:

- In the Files tab, select **Upload** and **Choose File** to find the .Rmd that you downloaded. Click *OK* to upload to your course folder/location in the maize server account.
- Click on the .Rmd file in the appropriate folder to open the file.

1.8 Running codes and knitting .Rmd files:

- You can run a line of code by placing your cursor in the line of code and clicking **Run Selected Line(s)**
- You can run an entire chunk by clicking the green triangle on the right side of the code chunk.
- After each small edit or code addition, **Knit** your Markdown. If you wait until the end to Knit, it will be harder to find errors in your work.
- Format output type: You can use any of pdf_document, html_document type, or word_document type.
- **Maize users**: You may also need to allow for "pop-up" in your web browser when knitting documents.

1.9 Few More Instructions

The default setting in Rstudio when you are running chunks is that the "output" (numbers, graphs) are shown **inline** within the Markdown Rmd. If you prefer to have your plots appear on the right of the console and not below the chunk, then change the settings as follows:

1. Select Tools > Global Options.
2. Click the R Markdown section and uncheck (if needed) the option Show output inline for all R Markdown documents.
3. Click OK.

Now try running R chunks in the .Rmd file to see the difference. You can recheck this box if you prefer the default setting.

1.10 VPN

If you plan to do any work off campus this term, you need to install Carleton's VPN. This will allow you to access the **maize** server (if needed).

Installing the GlobalProtect VPN

Follow the directions here to install VPN.

Chapter 2

R Markdown Basics

An R Markdown file (.Rmd file) combines R commands and written analyses, which are 'knit' together into an HTML, PDF, or Microsoft Word document.

An R Markdown file contains three essential elements:

- Header: The header (top) of the file contains information like the document title, author, date and your preferred output format (pdf_document, word_document, or html_document).
- Written analysis: You write up your analysis after the header and embed R code where needed. The online help below shows ways to add formatting details like bold words, lists, section labels, etc to your final pdf/word/html document. For example, adding ** before and after a word will bold that word in your compiled document.
- R chunks: R chunks contain the R commands that you want evaluated. You embed these chunks within your written analysis and they are evaluated when you compile the document.

Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

2.1 R Markdown Syntax

2.1.1 Lists in R Markdown:

You can use asterisk mark to provide emphasis, such as `*italics*` or `**bold**`. You can create lists with a dash:

```
- Item 1
- Item 2
- Item 3
  + Subitem 1
* Item 4
```

to produce

- Item 1
- Item 2
- Item 3
  - Subitem 1
- Item 4

You can embed Latex equations in-line, `$\frac{1}{n} \sum_{i=1}^{n} x_{i}$` to produce $\frac{1}{n}\sum_{i=1}^{n} x_i$ or in a new line as `$$\text{Var}(X) = \frac{1}{n-1}\sum_{i-1}^{n} (x_{i} - \bar{x})^2$$` to produce

$$\text{Var}(X) = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

2.1.2 Embed an R code chunk:

Use the following

````
```r
Use back ticks to
create a block of code
```
````

to produce:

```
Use back ticks to
create a block of code
```

You can also evaluate and display the results of R code. Each tasks can be accomplished in a suitably labeled chunk like the following:

```
summary(cars)
```

```
     speed           dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
```

```
fit <- lm(dist ~ speed, data = cars)
fit
```

```

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

2.1.3 Including Plots:

You can also embed plots. See Figure 2.1 for example:

```
par(mar = c(0, 1, 0, 1))
pie(
  c(280, 60, 20),
  c('Sky', 'Sunny side of pyramid', 'Shady side of pyramid'),
  col = c('#0292D8', '#F7EA39', '#C4B632'),
  init.angle = -50, border = NA
)
```

(Credit: Yihui Xie)

2.1.4 Read in data files:

```
simple_data <- read.csv("https://deepbas.io/data/simple-1.dat", )
summary(simple_data)
```

```
   initials            state                age
 Length:3           Length:3           Min.   :45.0
 Class :character   Class :character   1st Qu.:47.5
```

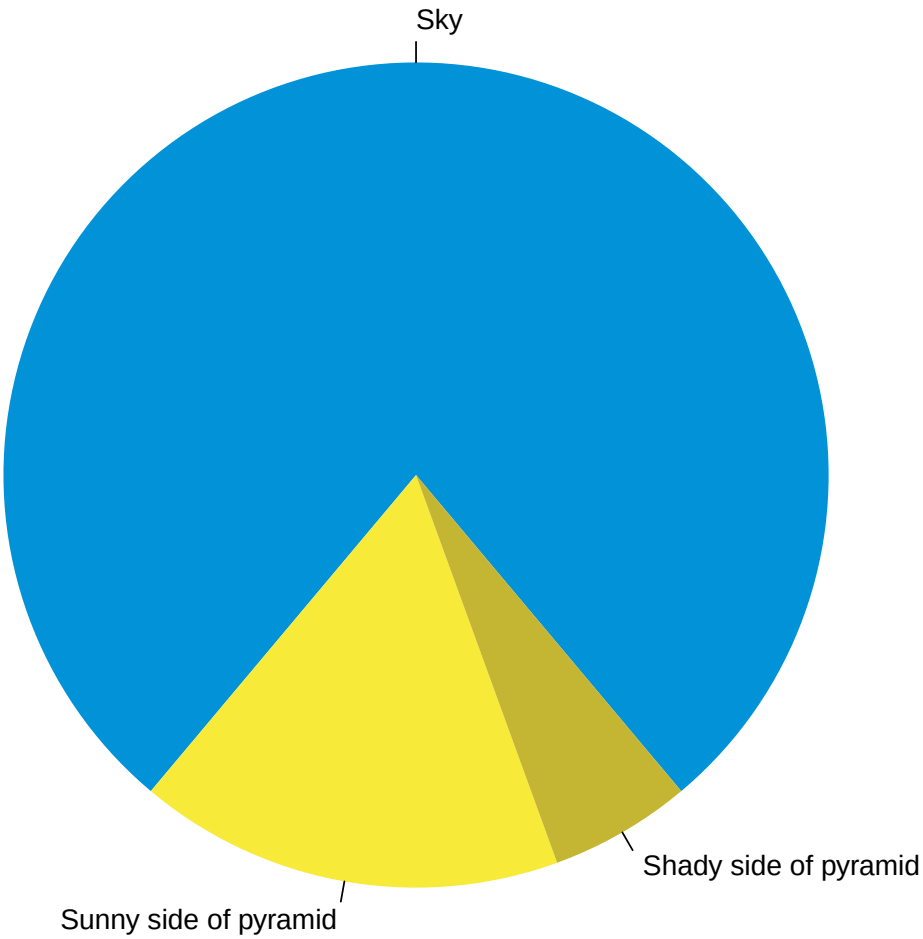


Figure 2.1: A fancy pie chart.

```
 Mode  :character   Mode  :character   Median :50.0
                                       Mean   :52.0
                                       3rd Qu.:55.5
                                       Max.   :61.0
    time
 Length:3
 Class :character
 Mode  :character
```

```
knitr::kable(simple_data)
```

initials	state	age	time
vib	MA	61	6:01
adc	TX	45	5:45
kme	CT	50	4:19

2.1.5 Hide the code:

If we enter the `echo = FALSE` option in the R chunk (see the .Rmd file). This prevents the R code from being printed to your document; you just see the results.

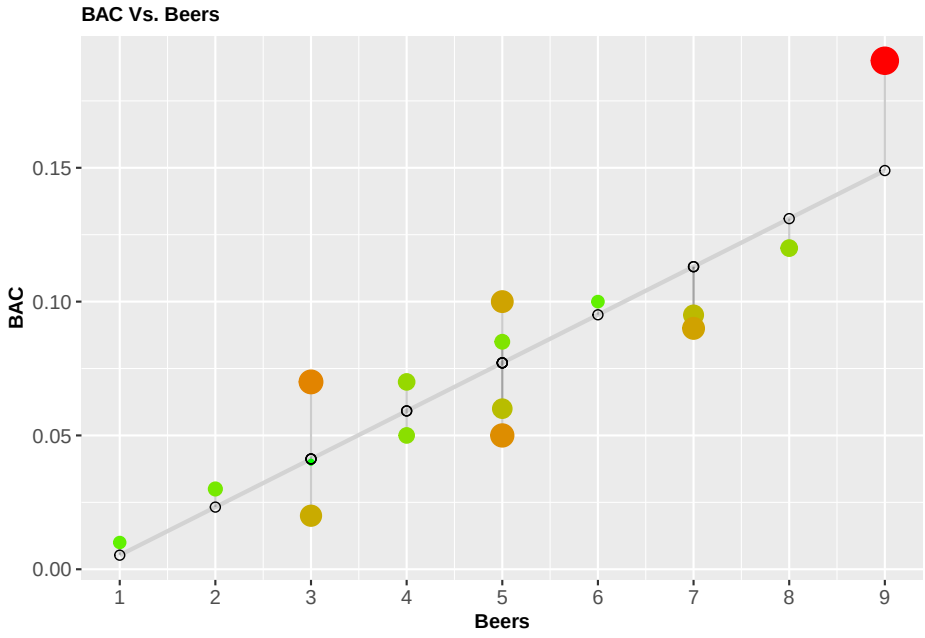
initials	state	age	time
vib	MA	61	6:01
adc	TX	45	5:45
kme	CT	50	4:19

Chapter 3

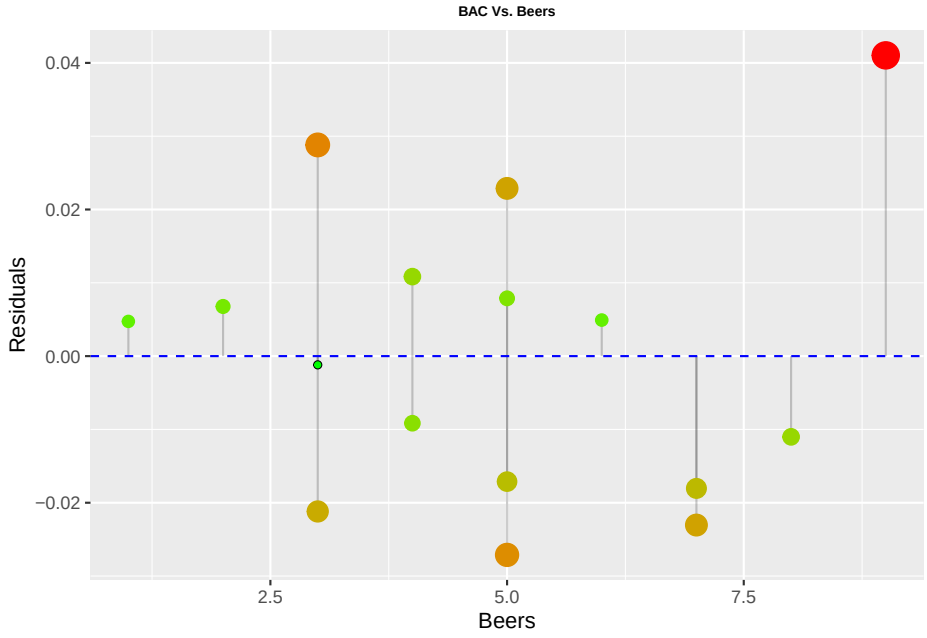
Helpful R codes

3.1 Residual Plots in ggplot2

```
# residual size plot
library(ggplot2)
bac <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/BAC.csv")

fit <- lm(BAC ~ Beers, data = bac) # fit the model
bac$predicted <- predict(fit)   # Save the predicted values
bac$residuals <- residuals(fit) # Save the residual values

ggplot(bac, aes(x = Beers, y = BAC)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +     # regression line
  geom_segment(aes(xend = Beers, yend = predicted), alpha = .2) +      # draw line from point to
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +  # size of the points
  scale_color_continuous(low = "green", high = "red") +
  labs(title = "BAC Vs. Beers") +# color of the points mapped to residual size - green smaller, r
  guides(color = FALSE, size = FALSE) +                             # Size legend removed
  geom_point(aes(y = predicted), shape = 1, size = 2) +
  scale_x_continuous(breaks=1:9)+
  theme(axis.text=element_text(size=10),
        axis.title=element_text(size=10,face="bold"),
        plot.title = element_text(size = 10, face = "bold"))
```




```
ggplot(bac, aes(x = Beers, y = residuals)) +
  geom_point() +
  theme(legend.position = "none") +
  geom_segment(aes(xend = Beers, yend = 0), alpha = .2) +
  scale_color_continuous(low = "green", high = "red") +
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +  # size of the poin
  geom_hline(yintercept = 0, col = "blue", size = 0.5, linetype = "dashed") +
  labs(title = "BAC Vs. Beers",
       x = "Beers",
       y = "Residuals") +
  theme(plot.title = element_text(hjust=0.5, size=7, face='bold'))
```



3.2 Plotly codes

```
library(plotly)

cell_phone_data <- data.frame(
  Type = c("Android", "iPhone", "Blackberry", "Non Smartphone", "No Cell Phone"),
  Frequency = c(458, 437, 141, 924, 293)
)

data <- data.frame(
  Gender = c("Female", "Male"),
  In_a_relationship = c(32, 10),
  Its_complicated = c(12, 7),
  Single = c(63, 45)
)
```

```
plot_ly(cell_phone_data, labels = ~Type, values = ~Frequency, type = 'pie',
        textposition = 'inside', hoverinfo = 'label+value+percent',
        textinfo = 'label', insidetextfont = list(color = '#FFFFFF')) %>%
  layout(title = 'Cell Phone Usage',
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

```
plot_ly(data, x = ~Gender, y = ~In_a_relationship, type = 'bar', name = 'In a relations
  add_trace(y = ~Its_complicated, name = 'It\'s complicated') %>%
  add_trace(y = ~Single, name = 'Single') %>%
  layout(yaxis = list(title = 'Number of People'), barmode = 'group')
```

Chapter 4

Homework Guidelines

- You **can** discuss homework problems with classmates, but you must write up **your own** homework solutions and **do your own work in R (no sharing commands or output) unless explicitly told otherwise**.
- **Getting help**: You **can** use the following resources to complete your homework:
  - Carleton faculty (myself, other stat faculty, etc)
  - Discussions with classmates (see above) or knowledgeable friends
  - The math skills center
  - Lab assistants (in CMC 304)
  - Prefects
  - Student solutions provided in the back of your student textbook or in the student solution manual
  - It is okay to get coding help from prefects, tutors, classmates, online resources, but extra care should be done to write your own versions of the codes.
- You **cannot** use any resources other than the ones listed above to complete assignments (homework, reports, etc) for this class. E.g. you cannot use a friend's old assignments or reports, answers found on the internet, textbook (instructor) solutions manual, etc.

4.1 Format

- At the top of each assignment, provide the following details:
  - Class name (e.g., Stat xxx)
  - Homework number (e.g., "homework 1")
  - Your name

  - The names of classmates that you worked with on all or part of the assignment

- Turn in a neat, **correctly ordered**, and **legible** assignment with no ragged edges. If it can't be read, it will not be graded.
- **Staple** - no folded corners accepted!

4.2 Content

- You must **show all work** and formulas used to answer any question which requires a numerical answer. Be sure to show the natural sequence of work needed to answer the problem.
- Use **complete sentences** when answering any problem that requires an explanation or overall problem summary.

4.3 Problems using R

- These problems must be written up as Word or pdf document using R Markdown.
- **Label** all output with the problem number.
- First **give your answer to a problem in written form**, never just give R output as your answer. Follow your written answer with your "work" which contains **all relevant R commands and output** (numeric output or graphs) that are needed to answer a homework problem. Do not include typos or unnecessary commands/output.

Chapter 5

Graph Formatting

This worksheet provides a comprehensive guide on graph formatting in R using the ggplot2 package. We will explore various aspects of formatting, such as adding figure numbers, captions, titles, axes labels, customizing themes, and using different color scales.

5.1 Load the required packages and datasets

```
Cereals <- read.csv("http://people.carleton.edu/~kstclair/data/Cereals.csv")
```

5.2 Graph theme and colors

5.2.1 Adding figure numbers and captions

To automatically add figure numbers and captions, include the option fig_caption: true in the output options at the top of your markdown file. To add captions to the figures, use the fig.cap argument in the R code chunk that creates the figure.

```
ggplot(Cereals, aes(x = calgram)) +
  geom_histogram(binwidth = 0.3, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Calorie Content in Cereals",
       x = "Calorie Content (g)",
       y = "Count") +
  theme_minimal()
```

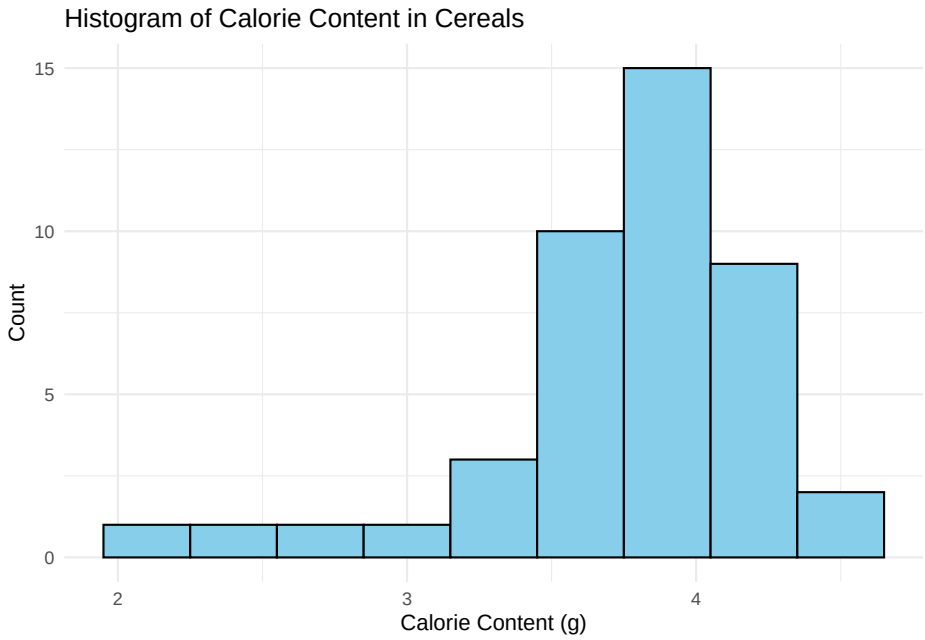
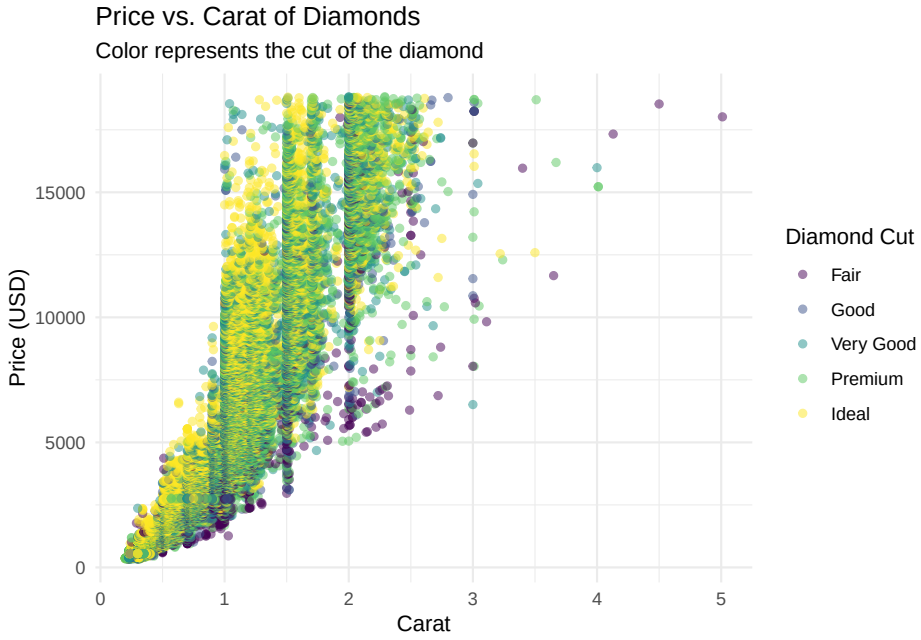


Figure 5.1: A nice figure

5.2.2 Customizing titles, axis labels, and legends

You can customize titles, axis labels, and legends using the labs() function.

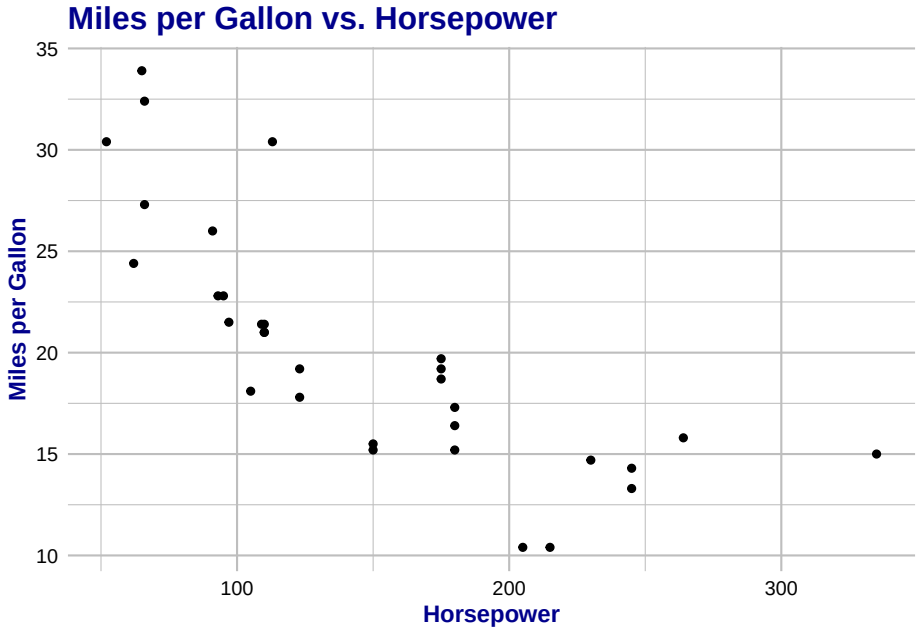
```
ggplot(data = diamonds, aes(x = carat, y = price, color = cut)) +
  geom_point(alpha = 0.5) +
  labs(title = "Price vs. Carat of Diamonds",
       subtitle = "Color represents the cut of the diamond",
       x = "Carat",
       y = "Price (USD)",
       color = "Diamond Cut") +
  theme_minimal()
```



5.2.3 Customizing themes

You can customize themes using the `theme()` function and various `element_*()` functions.

```
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  labs(title = "Miles per Gallon vs. Horsepower",
       x = "Horsepower",
       y = "Miles per Gallon") +
  theme_minimal() +
  theme(plot.title = element_text(size = 16, face = "bold", color = "darkblue"),
        axis.title = element_text(size = 12, face = "bold", color = "darkblue"),
        axis.text = element_text(size = 10, color = "black"),
        panel.grid.major = element_line(color = "gray", size = 0.5),
        panel.grid.minor = element_line(color = "gray", size = 0.25))
```



5.2.4 Using different color scales

You can use different color scales for both continuous and discrete variables using the `scale_color_*()` and `scale_fill_*()` functions.

```
ggplot(data = diamonds, aes(x = carat, y = price, color = cut)) +
  geom_point(alpha = 0.5) +
  labs(title = "Price vs. Carat of Diamonds",
       subtitle = "Color represents the cut of the diamond",
       x = "Carat",
       y = "Price (USD)",
       color = "Diamond Cut") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```

5.2.5 Customizing plot elements

You can customize plot elements such as points, lines, and bars using the corresponding geom_*() functions and their arguments.

```
ggplot(mtcars, aes(x = hp, y = mpg, shape = factor(gear), size = gear)) +
  geom_point(aes(color = factor(gear))) +
```

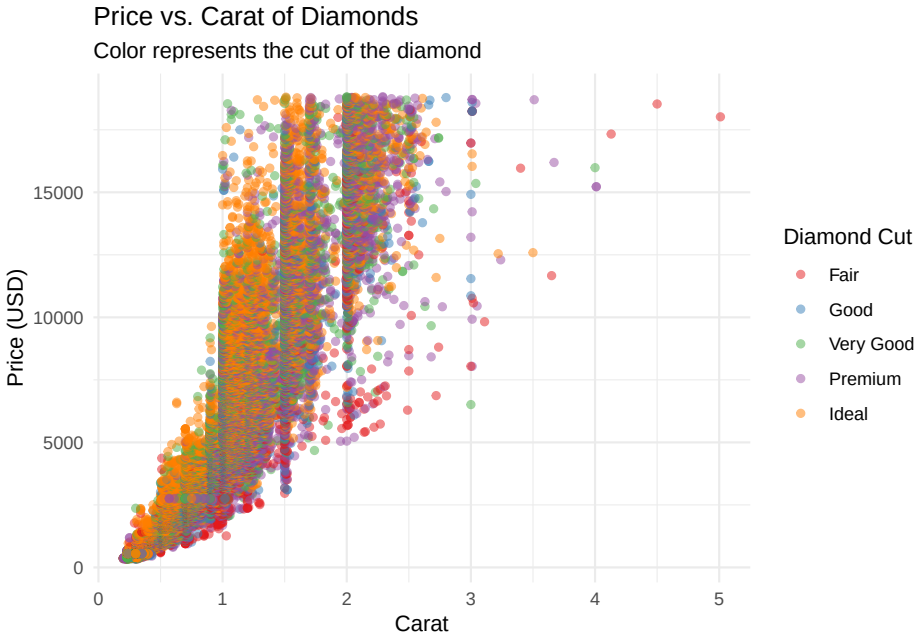
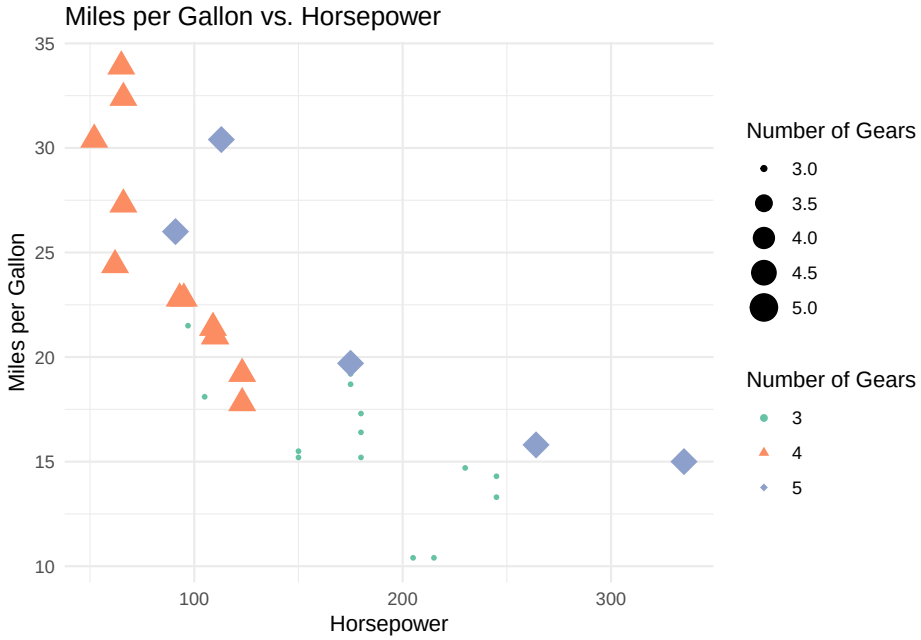


Figure 5.2: Figure 4: Scatterplot of price vs. carat of diamonds with color representing the cut and custom color scale.

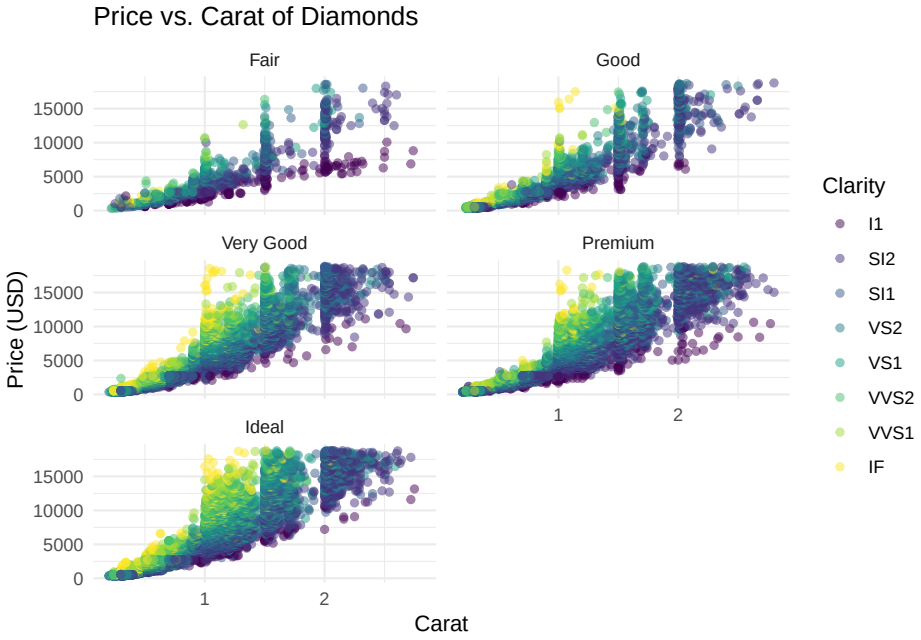
```
labs(title = "Miles per Gallon vs. Horsepower",
     x = "Horsepower",
     y = "Miles per Gallon",
     color = "Number of Gears",
     shape = "Number of Gears",
     size = "Number of Gears") +
theme_minimal() +
scale_shape_manual(values = c(16, 17, 18)) +
scale_color_brewer(palette = "Set2")
```



5.2.6 Faceting

You can create multiple plots based on a categorical variable using the `facet_wrap()` and `facet_grid()` functions.

```
ggplot(data = diamonds %>% filter(carat < 3), aes(x = carat, y = price)) +
  geom_point(aes(color = clarity), alpha = 0.5) +
  labs(title = "Price vs. Carat of Diamonds",
       x = "Carat",
       y = "Price (USD)",
       color = "Clarity") +
  facet_wrap(~ cut, ncol = 2) +
  theme_minimal()
```



5.3 Graph Sizing in R

This worksheet demonstrates how to adjust the size of various plots in R using ggplot2. We will explore different techniques to control the size of the plots and their elements.

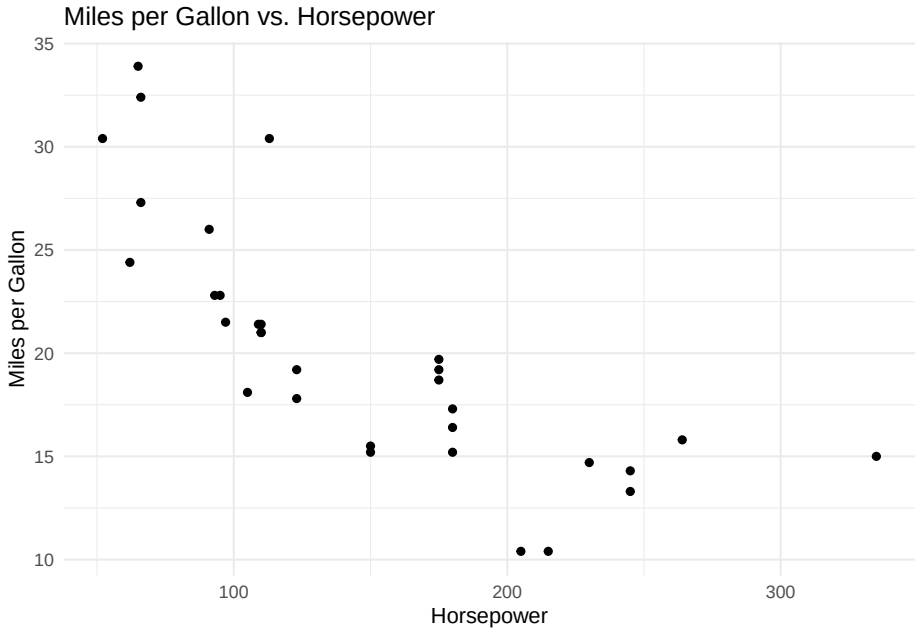
5.3.1 Load the necessary libraries and data

```
library(ggplot2)
library(dplyr)

# Load the built-in datasets
data("mtcars")
data("diamonds")
data("iris")
```

5.3.2 Adjusting the overall size of the plot

You can control the overall size of the plot using the width and height options within the R Markdown output settings. Another way is to use the ggsave() function when saving the plot as an image file.

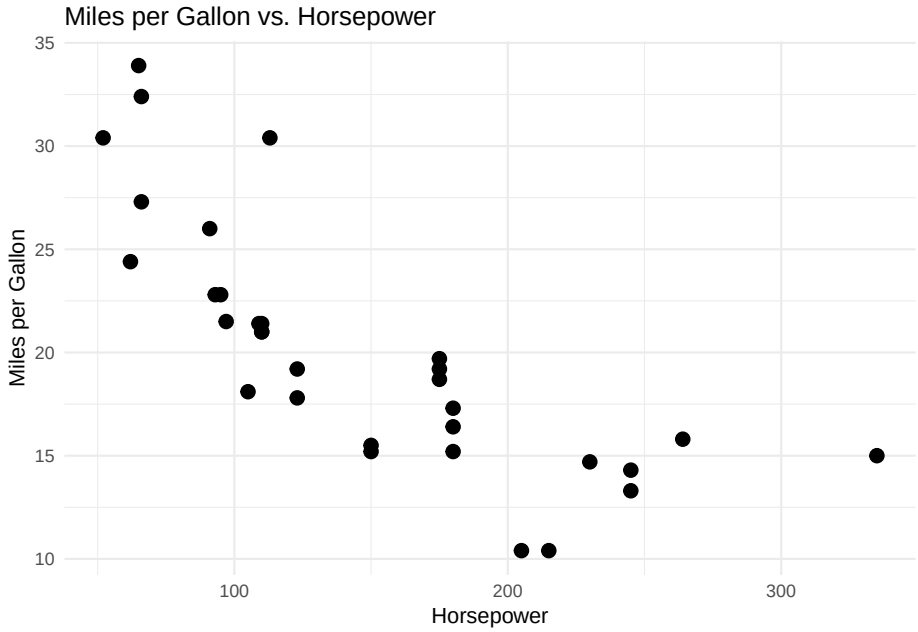
```
scatter_plot <- ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  labs(title = "Miles per Gallon vs. Horsepower",
       x = "Horsepower",
       y = "Miles per Gallon") +
  theme_minimal()

scatter_plot
```



5.3.3 Adjusting the size of points, lines, and bars

Use the size parameter within the geom_*() functions to control the size of points, lines, and bars.

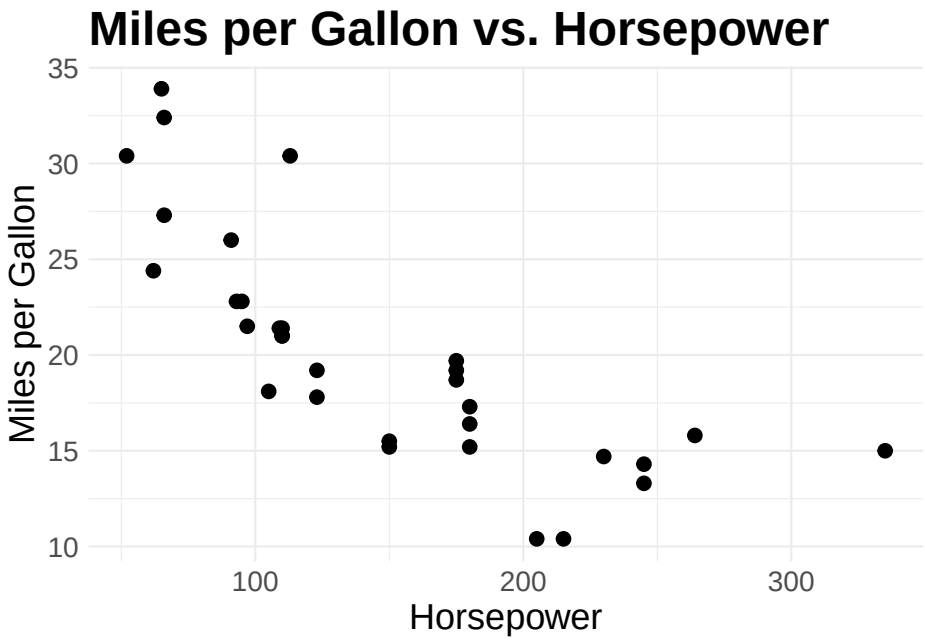
```
scatter_plot_large_points <- scatter_plot +
  geom_point(size = 3)

scatter_plot_large_points
```



5.3.4 Adjusting the size of text elements

You can change the size of text elements, such as axis labels and titles, using the theme() function.

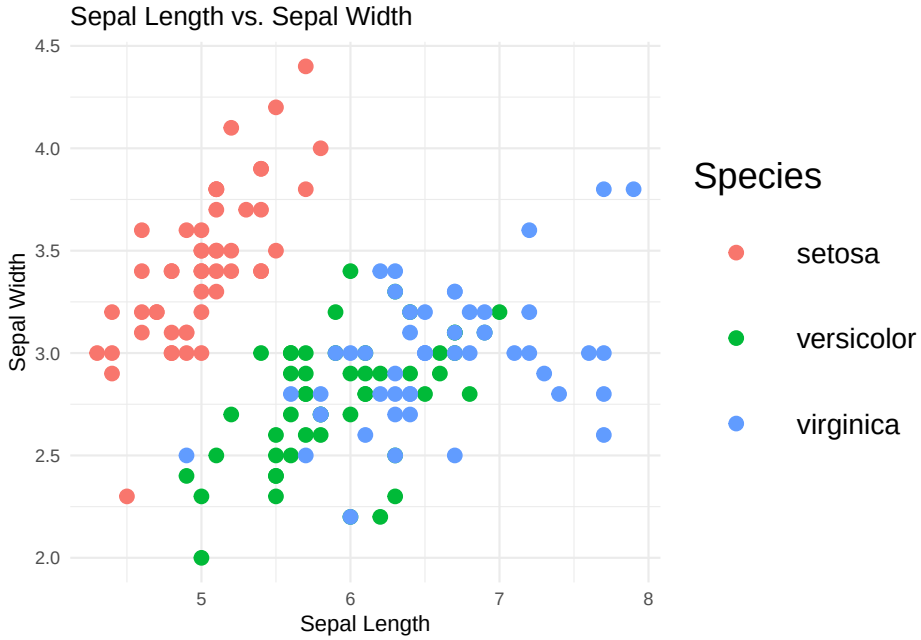
```
scatter_plot_custom_text <- scatter_plot_large_points +
  theme(plot.title = element_text(size = 24, face = "bold"),
        axis.title.x = element_text(size = 18),
        axis.title.y = element_text(size = 18),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14))

scatter_plot_custom_text
```



5.3.5 Adjusting the size of legend elements

You can modify the size of the legend elements using the theme() function along with element_text() and element_rect().

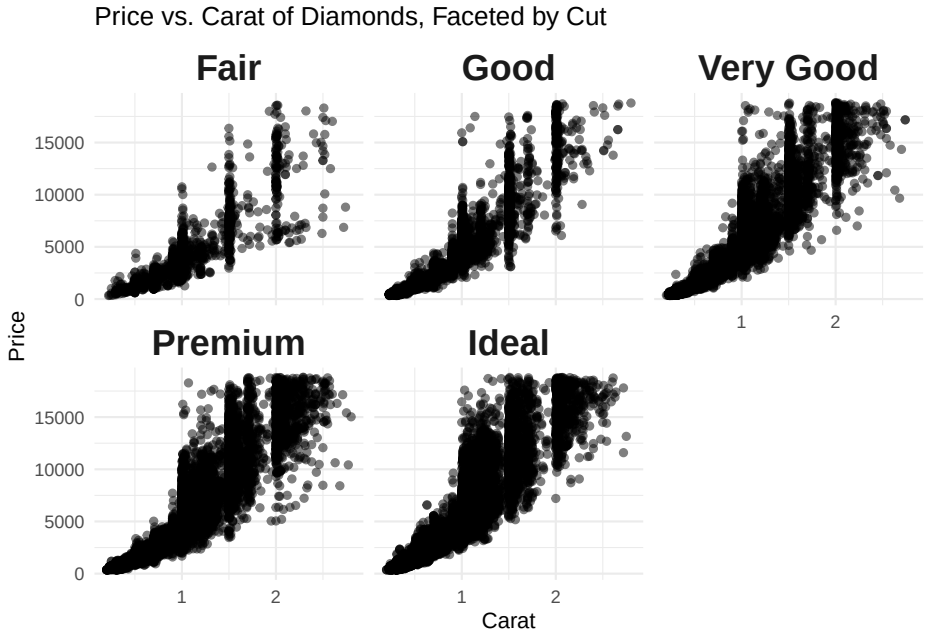
```
iris_scatter_plot <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Specie
  geom_point(size = 3) +
  labs(title = "Sepal Length vs. Sepal Width",
       x = "Sepal Length",
       y = "Sepal Width",
       color = "Species") +
  theme_minimal() +
  theme(legend.title = element_text(size = 18),
        legend.text = element_text(size = 14),
        legend.key.size = unit(1.5, "cm"))

iris_scatter_plot
```



5.3.6 Adjusting the size of facet labels

You can control the size of facet labels using the theme() function along with element_text().

```
diamonds_facet_plot <- ggplot(data = diamonds %>% filter(carat < 3), aes(x = carat, y = price)) +
geom_point(alpha = 0.5) +
facet_wrap(~cut) +
labs(title = "Price vs. Carat of Diamonds, Faceted by Cut",
x = "Carat",
y = "Price") +
theme_minimal() +
theme(strip.text = element_text(size = 18, face = "bold"))

diamonds_facet_plot
```



5.3.7 Adjusting the size of axis ticks

You can modify the size of axis ticks using the `theme()` function along with `element_line()`.

```
scatter_plot_custom_ticks <- scatter_plot +
  theme(axis.ticks = element_line(size = 1.5),
        axis.ticks.length = unit(0.3, "cm"))

scatter_plot_custom_ticks
```

5.3.8 Adding text labels to points

Use the geom_text() or geom_label() functions to add text labels to points.

```
mtcars$car_name <- rownames(mtcars)

scatter_plot_labels <- ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(color = gear)) +
  geom_text(aes(label = car_name), check_overlap = TRUE, vjust = 1.5) +
  labs(title = "Scatter plot of MPG vs Weight with Car Labels",
       x = "Weight",
```

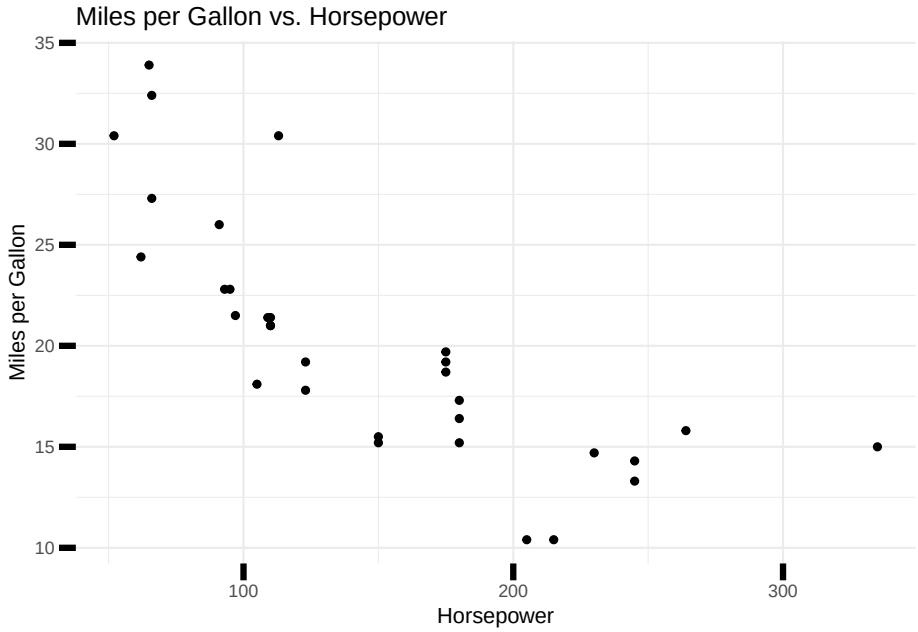
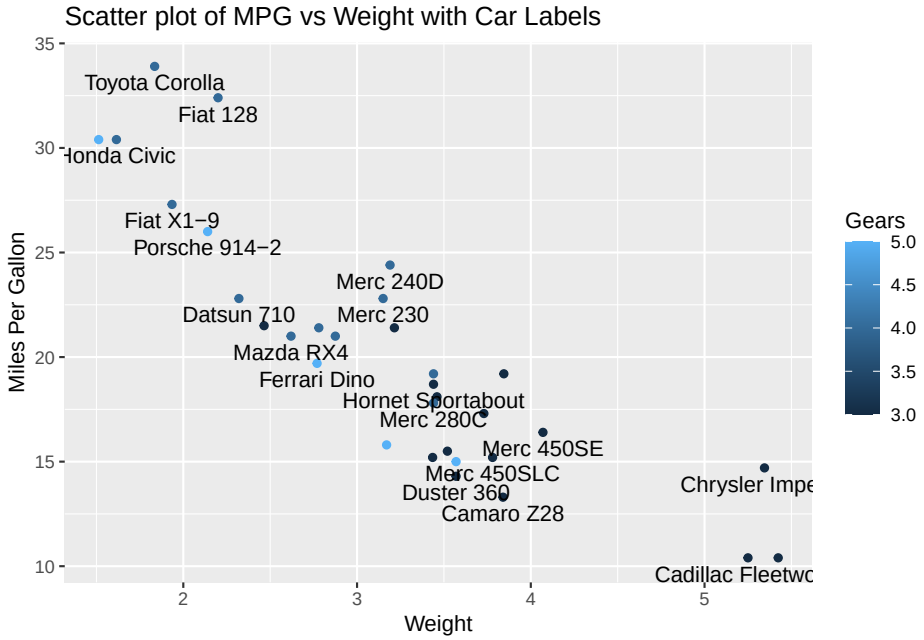


Figure 5.3: Figure 6: Scatterplot of mpg vs. hp with customized axis tick size.

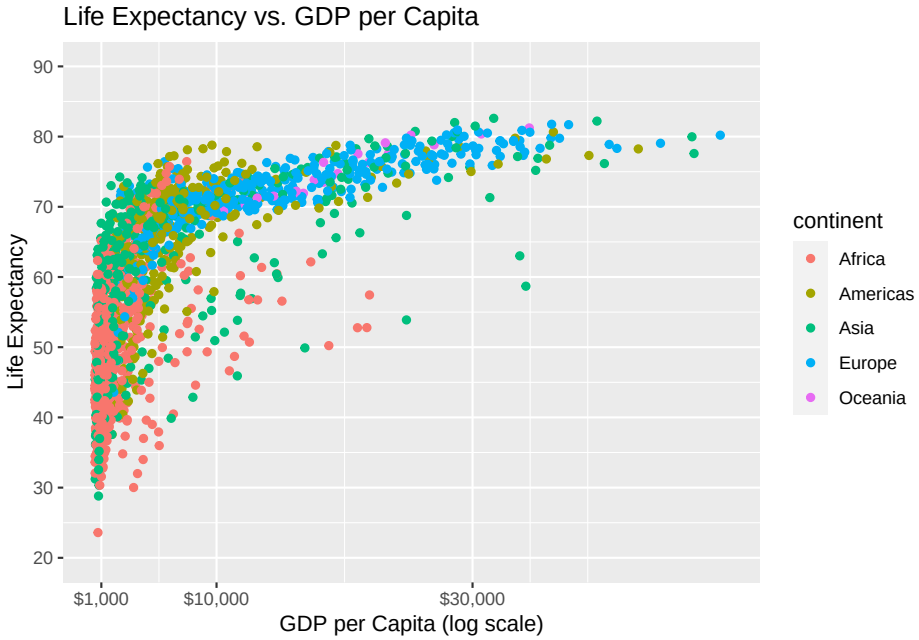
```
       y = "Miles Per Gallon",
       color = "Gears")

scatter_plot_labels
```



5.3.9 Modifying axis limits and scales

Use scale_x_continuous() and scale_y_continuous() to modify axis limits and scales.

```
data("gapminder", package = "gapminder")

ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, color = continent)) +
  geom_point() +
  scale_x_log10() +
  scale_x_continuous(limits = c(500, 50000), breaks = c(1000, 10000, 30000), labels = s
  scale_y_continuous(limits = c(20, 90), breaks = seq(20, 90, 10)) +
  labs(title = "Life Expectancy vs. GDP per Capita",
       x = "GDP per Capita (log scale)",
       y = "Life Expectancy")
```


Life Expectancy vs. GDP per Capita
Life Expectancy
90
80
70
60
50
40
30
20
$1,000
$10,000
$30,000
GDP per Capita (log scale)
continent
Africa
Americas
Asia
Europe
Oceania

Chapter 6

Table Formatting

In this worksheet, we will explore various options for outputting and formatting tables in R using the RMarkdown environment.

6.0.1 Basic Table Formatting with `kable`

The `kable()` function from the knitr package provides a simple way to output tables in RMarkdown.

```
library(knitr)
kable(mtcars[1:5, 1:5], caption = "A basic table using kable")
```

We will also use the `Gapminder` dataset for our examples. This dataset contains information about life expectancy, GDP per capita, and population size for various countries and years.Here's an example of how to display the first 10 rows of the `Gapminder` dataset.

```
data("gapminder", package = "gapminder")
knitr::kable(head(gapminder, 10), caption = "Table 1: First 10 rows of the Gapminder dataset.")
```

Table 6.1: A basic table using kable

	mpg	cyl	disp	hp	drat
Mazda RX4	21.0	6	160	110	3.90
Mazda RX4 Wag	21.0	6	160	110	3.90
Datsun 710	22.8	4	108	93	3.85
Hornet 4 Drive	21.4	6	258	110	3.08
Hornet Sportabout	18.7	8	360	175	3.15

Table 6.2: Table 1: First 10 rows of the Gapminder dataset.

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134
Afghanistan	Asia	1982	39.854	12881816	978.0114
Afghanistan	Asia	1987	40.822	13867957	852.3959
Afghanistan	Asia	1992	41.674	16317921	649.3414
Afghanistan	Asia	1997	41.763	22227415	635.3414

Table 6.3: A formatted table with kableExtra

	mpg	cyl	disp	hp	drat
Mazda RX4	21.0	6	160	110	3.90
Mazda RX4 Wag	21.0	6	160	110	3.90
Datsun 710	22.8	4	108	93	3.85
Hornet 4 Drive	21.4	6	258	110	3.08
Hornet Sportabout	18.7	8	360	175	3.15

6.0.2 Formatting Tables with `kableExtra`

To further customize the table appearance, we can use the `kableExtra` package.

```
#install.packages("kableExtra")
library(kableExtra)
kable(mtcars[1:5, 1:5], caption = "A formatted table with kableExtra") %>%
  kable_styling("striped", full_width = F)
```

6.0.3 Customizing column formats

Use the `column_spec()` function from the `kableExtra` package to customize the appearance of individual columns.

```
gapminder %>%
  head(10) %>%
  knitr::kable(caption = "Table 3: First 10 rows of the Gapminder dataset with custom c
  kableExtra::kable_styling("striped", full_width = F) %>%
  kableExtra::column_spec(2, bold = TRUE, color = "red") %>%
  kableExtra::column_spec(4, monospace = TRUE)
```

Table 6.4: Table 3: First 10 rows of the Gapminder dataset with custom column formatting.

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	**Asia**	1952	28.801	8425333	779.4453
Afghanistan	**Asia**	1957	30.332	9240934	820.8530
Afghanistan	**Asia**	1962	31.997	10267083	853.1007
Afghanistan	**Asia**	1967	34.020	11537966	836.1971
Afghanistan	**Asia**	1972	36.088	13079460	739.9811
Afghanistan	**Asia**	1977	38.438	14880372	786.1134
Afghanistan	**Asia**	1982	39.854	12881816	978.0114
Afghanistan	**Asia**	1987	40.822	13867957	852.3959
Afghanistan	**Asia**	1992	41.674	16317921	649.3414
Afghanistan	**Asia**	1997	41.763	22227415	635.3414

6.0.4 Formatting Tables with flextable

Another option for table formatting is the flextable package.

```
#install.packages("flextable")
library(flextable)
ft <- flextable(mtcars[1:5, 1:5])
ft <- set_caption(ft, caption = "A table using flextable")
ft
```

Table 6.5: A table using flextable

mpg	cyl	disp	hp	drat
21.0	6	160	110	3.90
21.0	6	160	110	3.90
22.8	4	108	93	3.85
21.4	6	258	110	3.08
18.7	8	360	175	3.15

6.0.5 Formatting Tables with gt

The gt package provides another way to create formatted tables in R.

```
#install.packages("gt")
library(gt)
gt(mtcars[1:5, 1:5]) %>%
  tab_header(title = "A table using gt")
```

A table using gt

mpg	cyl	disp	hp	drat
21.0	6	160	110	3.90
21.0	6	160	110	3.90
22.8	4	108	93	3.85
21.4	6	258	110	3.08
18.7	8	360	175	3.15

6.0.6 Hiding R commands and R output

As mentioned in the graph formatting handout, adding the chunk option echo=FALSE will display output (like graphs) produced by a chunk but not show the commands used in the chunk. You can stop both R commands and output from being displayed in a document by adding the chunk option include=FALSE.

As you work through a report analysis, you may initially want to see all of your R results as you are writing your report. But after you've summarized results in paragraphs or in tables, you can then use the include=FALSE argument to hide your R commands and output in your final document. If you ever need to rerun or reevaluate your R work for a report, you can easily recreate and edit your analysis since the R chunks used in your original report are still in your R Markdown .Rmd file.

6.0.7 Summary statistics with pander

We can use the pander package to create summary tables.

```
#install.packages("pander")
library(pander)
pander(summary(mtcars$mpg), caption = "Summary statistics for miles per gallon")
```

Table 6.7: Summary statistics for miles per gallon

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.4	15.43	19.2	20.09	22.8	33.9

6.0.8 t-test results with pander

Let's perform a t-test comparing the miles per gallon (mpg) for cars with 4 and 6 cylinders.

```
t_test_result <- t.test(mpg ~ as.factor(cyl), data = mtcars, subset = cyl %in% c(4, 6))
pander(t_test_result, caption = "Comparing MPG for 4 and 6 cylinder cars")
```

Table 6.8: Comparing MPG for 4 and 6 cylinder cars (continued below)

Test statistic	df	P value	Alternative hypothesis
4.719	12.96	0.0004048 * * *	two.sided

mean in group 4	mean in group 6
26.66	19.74

6.0.9 Chi-square test results with pander

Now let's perform a chi-square test to check for an association between the number of cylinders and the type of transmission (automatic or manual).

```
my_table <- table(mtcars$cyl, mtcars$am)
chisq_test_result <- chisq.test(my_table)
pander(chisq_test_result, caption = "Chi-square test for cylinders and transmission type")
```

Table 6.10: Chi-square test for cylinders and transmission type

Test statistic	df	P value
8.741	2	0.01265 *

6.0.10 Linear regression results with pander

Finally, let's fit a linear regression model of miles per gallon (mpg) as a function of weight (wt) and display the results.

```
lm_result <- lm(mpg ~ wt, data = mtcars)
pander(lm_result, caption = "Linear regression of MPG on weight")
```

Table 6.11: Linear regression of MPG on weight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.29	1.878	19.86	8.242e-19
wt	-5.344	0.5591	-9.559	1.294e-10

Chapter 7

Report Guidelines

The data analysis report that you will be writing for this class will ask two things of you:

1. use appropriate statistical methods to answer research questions and
2. clearly and concisely communicate the meaning of your statistics and graphs to a **reader** who has a basic knowledge of statistics.

Your report should be organized, well-written with proper use of grammar, and contain sound reasoning and correct interpretations of statistical evidence. Also include *at least one* graphical display of your data on the body of your report. Be sure to hide the code used to produce it!

1. Your lab reports should be organized into the following sections:

    - **Introduction:** describe the data and your research questions
    - **Results:** describe your statistical analysis and interpret your graphs and numbers
    - **Discussion:** summarize your findings and answer your research questions, describe any limitations of your analysis
    - **Technical Appendix:** Staple all relevant R commands and output to the end of your written report. Only including commands without output is not enough. You must appropriately comment your code (telling me what each section of code is doing), and edit your code and output (no typos or errors allowed).

2. Type your report in Word/Google doc (or similar software) and use R (not Excel, Statkey or other software) for your analysis. You can also use R Markdown to write up your reports but you need to take care to only include labeled and numbered graphical output from R in the main

document. Commands and numerical output should be placed in the Technical Appendix. If you are interested in using Markdown talk to me for hints on its use for reports.

3. Carefully decide **appropriate graphs and numbers** to include. There is no need to show the same data in different forms (e.g. no need to show both a histogram and boxplot for the same variable). You also don't need to include all numbers given in R output if you only use a few in your analysis. You also don't need to "show" (or prove) skewness or outliers using numbers, just use your graphs to display skewness or outliers.

4. **Interpret** and give meaning to **all** graphs and numbers that you choose to include in your report. Do not include algebraic calculations or too much technical detail.

5. **Including Numbers: Never include R numerical output or commands**. Summarize needed output in a nicely formatted Word table or just integrate numbers into your writing. In you include tables, label them numerically (Table 1, Table 2, etc) and give each a title. Number in order of how they appear in the paper and refer to these tables by their number ("Table 1 displays summary statistics for income.").

6. **Including Graphs:** Resize all graphs appropriately so they fit nicely into your written report. Large graphs that take up most of a page with no, or very little, writing on the page impede the flow of the report and reduce its readability. Label all graphs numerically (Figure 1, etc.) as they occur in the paper, give each a title and refer to by number. See the stats lab manual chapter 1 if you need help copying plots into a Word/google doc.

7. **Do not explain every step taken in your study.** For example, there is no need to include a statement such as "I used R to create a histogram of income and observed that the distribution was right skewed". Instead just say "The distribution of income is skewed to the right (Figure 1)".

8. Avoid using weak phrases like "The average height of men is higher than the average for women." **Use numbers to bolster your explanation:** "The average height of men is three inches more than the average for women (68.5 vs. 65.5 inches)."

9. The **precision** of your data should dictate the precision of your statistics. In general, your statistics can have one to two more significant digits than your data. For example, if height is recorded to the nearest inch then the mean height should be reported as 65.5 (or 65.49) rather than the R value of 65.49268.

10. Sometimes a question posed in a study is **ambiguous** and there may be more than one way to correctly answer to the question. In grading your reports and paper, **I am most concerned with the logic of your**

conclusions and how you support your claim using data and statistical evidence.

Class Activity

Chapter 8

Class Activity 1

8.1 Your Turn 1

---

a. Run the following chunk. Comment on the output.

```
example_data = data.frame(ID = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
                          Greeting = c(rep("Hello", 5), rep("Goodbye",5)),
                          Male = rep(c(TRUE, FALSE), 5),
                          Weight = runif(n=10, 50, 300))
```

Click for answer

```
example_data
```

```
   ID Greeting  Male    Weight
1   1    Hello  TRUE 273.44819
2   2    Hello FALSE 261.67394
3   3    Hello  TRUE 224.79329
4   4    Hello FALSE 193.95138
5   5    Hello  TRUE 140.22001
6   6  Goodbye FALSE 199.97890
7   7  Goodbye  TRUE 121.66948
8   8  Goodbye FALSE  81.25836
9   9  Goodbye  TRUE 226.12194
10 10  Goodbye FALSE 183.53788
```

Answer: We see a data frame with four columns, where the first column is an `identifier` for the cases. We have information on the greeting types, whether male or not, and weight on these cases in the remaining columns.

b. What is the dimension of the dataset called 'example_data'?

Click for answer

```
dim(example_data)
[1] 10  4
nrow(example_data)
[1] 10
ncol(example_data)
[1] 4
```

Answer: There are 10 rows and 4 columns.

8.2 Your Turn 2

a. Read the dataset `EducationLiteracy` from the Lock5 second edition book.

Click for answer

```
# read in the data
library(readr)
education_lock5 <- read_csv("https://www.lock5stat.com/datasets2e/EducationLiteracy.csv
```

b. Print the header (i.e. first 6 cases by default) of the dataset in part a.

Click for answer

```
head(education_lock5)
```

```
# A tibble: 6 x 3
  Country             EducationExpenditure Literacy
  <chr>                              <dbl>    <dbl>
1 Afghanistan                          3.1     31.7
2 Albania                              3.2     96.8
```

```
3 Algeria                          4.3     NA
4 Andorra                          3.2     NA
5 Angola                           3.5     70.6
6 Antigua and Barbuda              2.6     99
```

c. What is the dimension of the dataset in a?

Click for answer

```
dim(education_lock5)
```

```
[1] 188   3
```

Answer: There are 188 rows and 3 columns.

d. What type of variables are `Country`, `EducationExpenditure`, and `Literacy`?

Click for answer

Answer: `Country` is a categorical variable. `EducationExpenditure` and `Literacy` are both quantitative variables.

e. If we would like to use education expenditure to predict the literacy rate of each countries, which variable is the explanatory variable and which one is the response?

Click for answer

Answer: The education expenditure is the explanatory variable, and the literacy rate is the response.

Chapter 9

Class Activity 2

9.1 Your Turn 1

9.1.1 Summary of article on It depends on how you ask!

Click for answer

Answer:

This study aimed to measure the effects of psychological biases on estimates of compliance with public health guidance regarding COVID-19. Results showed that compliance estimates were reduced when questions were framed negatively and anonymity was increased. Effect sizes were large, with compliance estimates diminishing by up to 17% points and 10% points, respectively. These findings suggest that standard tracking surveys pose questions in ways that lead to higher compliance estimates than alternative approaches.

9.2 Your Turn 2

9.2.1 Gettysburg random sample

Let's take a simple random sample (SRS) of Gettysburg words. The "population" is contained in the spreadsheet `GettysburgPopulationCounts.csv`. Carefully load this data into R:

```
pop <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Gettysberg
head(pop)
```

```
  position size  word
1        1    4  Four
2        2    5 score
3        3    3   and
4        4    5 seven
5        5    5 years
6        6    3  ago,
```

The `position` variable enumerates the list of words in the population (address).

(a). Sample

Run the following command to obtain a SRS of 10 words from the 268 that are in the population:

```
samp <- sample(1:268, size=10)
samp
```

```
 [1] 175 259 125 170 120 262 165  47 238 136
```

This tells you the position (row number) of your sampled words. What are your sampled positions? Why are your sampled positions different from other folks in class?

(b). Get words and lengths

We will *subset* the data set `pop` to obtain only the sampled rows listed in `samp`. We do this using **square bracket notation** 'dataset[row number, column number/name]. Run the following command to find your sampled words and sizes:

```
pop[samp,]
```

```
    position size      word
175      175    4      work
259      259    6   people,
125      125    3       who
170      170    9 dedicated
120      120    5     brave
262      262    6   people,
165      165    3       the
47        47    2        so
238      238    4     vain,
136      136    2        to
```

c. Compute your sample mean

The word lengths in part (b) are the data for your sample. You can compute your sample mean using a calculator, or using R. Let's try R (you will find it faster!). First save the quantitative variable `size` in a new variable called `mysize`:

```
mysize <- pop[samp, "size"]
mysize
```

```
 [1] 4 6 3 9 5 6 3 2 4 2
```

Then find the mean of these values:

```
mean(mysize)
```

```
[1] 4.4
```

How does this sample mean (from a truly random sample) compare to your sample mean from the non-random sample?

Click for answer

Answer: The true mean is 4.29. Your two means will likely vary. Since the many non-random samples generally overestimated the population mean length, it is possible (but not guaranteed) that *your* one non-random sample gave a mean length that is greater than the random sample's mean length.

9.2.2 Driving with a Pet on your Lap

Over 30,000 people participated in an online poll on `cnn.com` conducted in April 2012 asking: "Have you ever driven with a pet on your lap"? We see that 34% of the participants answered yes and 66% answered no.

a. Can you conclude that a random sample was used from the description given? Explain.

Click for answer

Answer: No you can't make this conclusion from the info given. In fact, an online poll at a website like `cnn.com` is almost always reporting results from a non-random sample. The people who respond are individuals who visit `cnn.com`, then see the online poll and decide to respond.

b. Explain why it is not appropriate to generalize these results to all drivers, or even to all drivers who visit cnn.com.

Click for answer

Answer: This is a volunteer sample, and volunteer samples are often biased and can't be generalized to *all drivers* (the population). It is likely that people who have driven with a pet on their lap are more likely to respond to the poll.

c. How might we select a sample of people that would give us results that we can generalize to a broader population?

Click for answer

Answer: A random sample of individuals from all U.S. drivers would need to be selected and given the poll question. (There are many ways to do this, the most common being a variation of random digit dialing where phone numbers are randomly selected from known area codes.)

d. Is the variable measured in this study quantitative or categorical?

Click for answer

Answer: Categorical (yes or no answer to the question).

Chapter 10

Class Activity 3

10.1 Case Study 1

Consider the following case study:

"Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed substantial improvement (reducing their level of depression) by the end of the study (Antonioli and Reveley, 2005)."

Observed data:

The researchers found that 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group.

(a). Identify the observational units in this study.

Click for answer

Answer: The observational units in this study are the 30 subjects.

(b). Classify each variable as categorical or quantitative.

Click for answer

Answer: The variables in this study can be classified as follows: Categorical: Treatment Group (Dolphin and Control),
Quantitative: Age, Depression Score (Beginning and End of Study)

(c). Which variable would you regard as explanatory and which as response?

Click for answer

Answer: The explanatory variable would be the Treatment Group and the response variable would be the Level of Depression.

(d). Is this an observational study or an experiment? Justify your answer.

Click for answer

Answer: This is an experiment because the researchers randomly assigned the subjects to the two treatment groups, and then observed the effect of the treatment (presence of dolphins) on the response variable (level of depression).

(e). Construct a two-way table based on the results of the experiment.

Click for answer

Two-way table:

Treatment	Improved	Not Improved	Total
Dolphin Group	10	5	15
Control Group	3	12	15
Total	13	17	30

10.2 Case Study 2

Consider the following case study:

"Researchers want to find out how a new diet affects weight gain among underweight subjects. This experiment only has two treatment conditions, the new diet and the standard diet. For this study, the researchers recruited 200 subjects which will be grouped into 100 pairs based on shared characteristics such as age, gender, weight, height, lifestyle, and so on. A 20-year-old female within the weight range of 90-110 pounds and the height range of 60-63 inches will be paired with another 20-year-old female that falls into the same weight and height categories. Once all 100 pairs are made, a subject from each pair will be randomly assigned into the treatment group (will be administered the new diet for 2 months) while the other subject from the pair will be assigned to the control group (will be assigned to follow the standard diet for two months).

At the end of the time time period of 2 months, researchers will measure the total weight gain for each subject."

Observed data:

The researchers found that 60 of 100 subjects in the new diet group showed substantial improvement, compared to 43 of 100 subjects in the standard diet group.

(a). Identify the observational units in this study.

Click for answer

Answer: The observational units in this study are the 200 subjects.

(b). Classify each variable as categorical or quantitative.

Click for answer

Answer: The variables are: age (quantitative), gender (categorical), weight (quantitative), height (quantitative), lifestyle (categorical), and total weight gain (quantitative).

(c). Which variable would you regard as explanatory and which as response?

Click for answer

Answer: The explanatory variable is the type of diet (new or standard) and the response variable is the total weight gain.

(d). Is this an observational study or an experiment? Justify your answer.

Click for answer

Answer: This is an experiment because the researchers are manipulating the explanatory variables (type of diet) to observe the effects on the response variables (total weight gain).

(e). If it is an experiment, is it randomized comparative experiment or a matched pairs experiment?

Click for answer

Answer: This is a matched pairs experiment because each subject is paired with another subject who has similar characteristics and one subject from each pair is randomly assigned to the treatment group and the other to the control group. More specifically, this is a $pretest - posttest$ matched pairs design.

(f). Construct a two-way table based on the results of the experiment.

Click for answer

Two-way table:

Outcome	New Diet	Standard Diet	Total
Improvement	60	43	103
No Improvement	40	57	97
Total	100	100	200

Chapter 11

Class Activity 4

11.1 Your Turn 1

11.1.1 Flowers v. Mississippi

The data set `APM_DougEvensCases.csv` contains data from 1517 potential black and white jurors for 66 cases that Doug Evans was primary prosecutor for between 1992 and 2017. These jurors were available for Doug Evans to strike using his "peremptory strikes" during the jury selection phase.

(a). Inspect data

Read in the data

```
jurors <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/APM_DougEvansCase
```

```
# dimension of dataset
dim(jurors)
```

```
[1] 1517    6
```

Look at the first **three rows** of the data set

```
jurors[c(1,2,3), ]
```

```
  trial__id  race        struck_state defendant_race
1         4 Black Not struck by State          White
2         4 Black     Struck by State          White
3         4 White Not struck by State          White
```

```
       same_race                     struck_by
1 different race Juror chosen to serve on jury
2 different race           Struck by the state
3      same race Juror chosen to serve on jury
```

To get the data from one variable, we use the command `dataset$variable`. For example, `jurors$struck_state` gives us the data values from the `struck_state` variable, which tells us if a juror was struck by the state from the jury pool. Here we can see the first 10 entries in this variable:

```
jurors$struck_state[1:10]
```

```
 [1] "Not struck by State" "Struck by State"
 [3] "Not struck by State" "Not struck by State"
 [5] "Struck by State"     "Not struck by State"
 [7] "Struck by State"     "Not struck by State"
 [9] "Not struck by State" "Not struck by State"
```

(b). Table of counts and proportions

The `summary` command used with a data frame gives summaries of each variable

```
summary(jurors)
```

```
  trial__id         race           struck_state
 Min.   :  4.0   Length:1517        Length:1517
 1st Qu.: 52.0   Class :character   Class :character
 Median : 82.0   Mode  :character   Mode  :character
 Mean   :112.6
 3rd Qu.:170.0
 Max.   :301.0
 defendant_race      same_race          struck_by
 Length:1517        Length:1517        Length:1517
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

The `table` command gives the distribution of counts for a single categorical variable. To obtain the count table for `struck_state` you need to

```
counts <- table(jurors$struck_state)
counts
```

```
Not struck by State     Struck by State
               1084                 433
```

We can add the `prop.table` command to turn these counts into proportions:

```
prop.table(counts)
```

```
Not struck by State     Struck by State
          0.7145682           0.2854318
```

- What proportion of eligible jurors were struck by the state from the jury pool?

Click for answer

Answer: about 28.5% of eligible jurors were struck by the state.

(c). Bar graph for one variable

We can create a data frame `count_data` containing the counts and their corresponding categories.

```
count_data <- data.frame(counts)
count_data
```

```
                 Var1 Freq
1 Not struck by State 1084
2     Struck by State  433
```

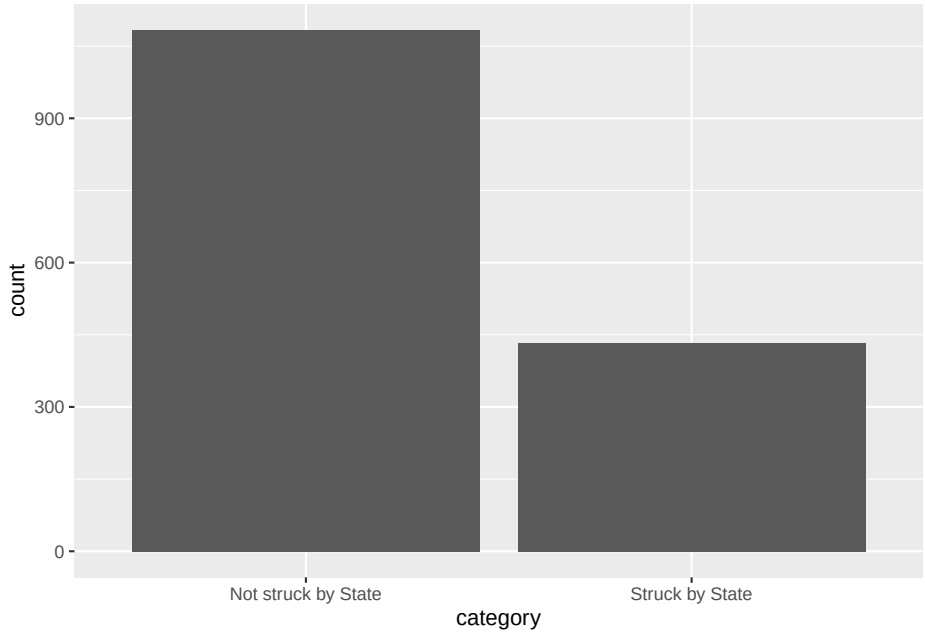
Then, we use `ggplot2` to create a bar plot with the categories on the x-axis and the counts on the y-axis. The column names of `count_data` are automatically assigned to be `Var1` and `Freq`. We can change the column names to `category` and `count`, for example, as:

```
colnames(count_data) = c("category", "count")
count_data
```

```
             category count
1 Not struck by State  1084
2     Struck by State   433
```

The `geom_bar(stat = "identity")` function is used to create the bars, and we set the y-axis label using `labs()`.

```
# Create a bar plot using ggplot2
library(ggplot2)  # load the package
ggplot(count_data, aes(x = category, y = count)) +
  geom_bar(stat = "identity") +
  labs(y = "count")
```



(d). Two-way tables

First 10 entries of `race` and `struck_state` variable is

```
jurors[(1:10),(2:3)]
```

```
    race        struck_state
1  Black Not struck by State
2  Black     Struck by State
3  White Not struck by State
4  White Not struck by State
5  Black     Struck by State
6  White Not struck by State
7  Black     Struck by State
8  White Not struck by State
9  White Not struck by State
10 White Not struck by State
```

The `table` command also gives two-way tables when two variables are included. Here is the two-way table for juror race and state struck status:

```
mytable <- table(jurors$race, jurors$struck_state)
mytable
```

```
       Not struck by State Struck by State
  Black                 225              310
  White                 859              123
```

- How many jurors were white and were not struck by the state?

Click for answer

answer: 859

(e). Conditional proportions: state strike status by juror race

The `prop.table` command gives conditional proportions for a two-way table. We plug our two-way table into `prop.table` with a `margin=1` to get proportions grouped by the **row** variable:

```
prop.table(mytable, margin = 1)
```

```
       Not struck by State Struck by State
  Black           0.4205607       0.5794393
  White           0.8747454       0.1252546
```

Of all eligible black jurors, about 57.9% were struck by the state.

- What proportion of eligible white jurors were struck by the state?

Click for answer

answer: about 12.5%

- Is there evidence of an association between juror race and state strikes?

Click for answer

answer: Yes, there is an association because the rate of state strikes varies greatly by juror race with about 60% of black jurors were struck compared to only 13% of white jurors

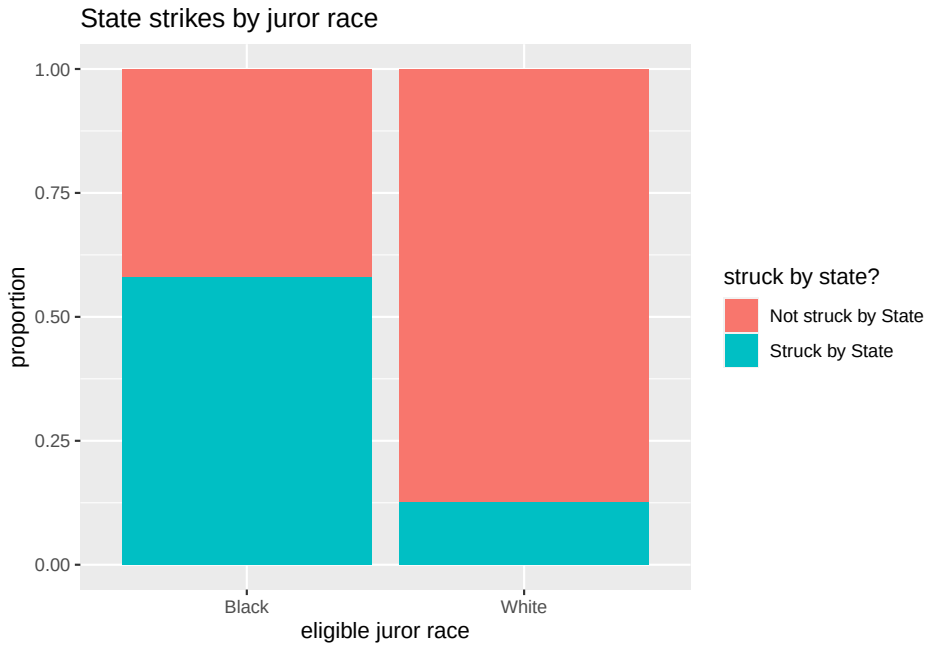
(f). Stacked bar graph for two variables

We can visualize the conditional distribution from part (e) with a stacked bar graph created using the `ggplot2` graphing package. First, load this package's functions with the `library` command:

```
library(ggplot2)
```

Now we can use the `geom_bar` command in this package. Here we get the conditional distribution of `struck_state` given `race`:

```
ggplot(jurors, aes(x = race, fill = struck_state)) +
  geom_bar(position = "fill") +
  labs(title = "State strikes by juror race", y = "proportion",
       x = "eligible juror race", fill = "struck by state?")
```



The basic syntax for this function is to let `ggplot` know your data set name (`jurors`), then specify the grouping or conditional variable on the x-axis (`race`) in the `aes` (aesthetic) argument. The `fill` variable is the response variable (`struck_state`). We add (`+`) the `geom_bar` geometry to get a bar graph with the `fill` position specified. Adding an informative label and title complete the graph.

(g). Conditional distribution of race grouped by strike status

We can "flip" our response and grouping variables easily (if we think it makes sense to do so). Here we specify the `margin=2` to get proportions grouped by the **column** variable:

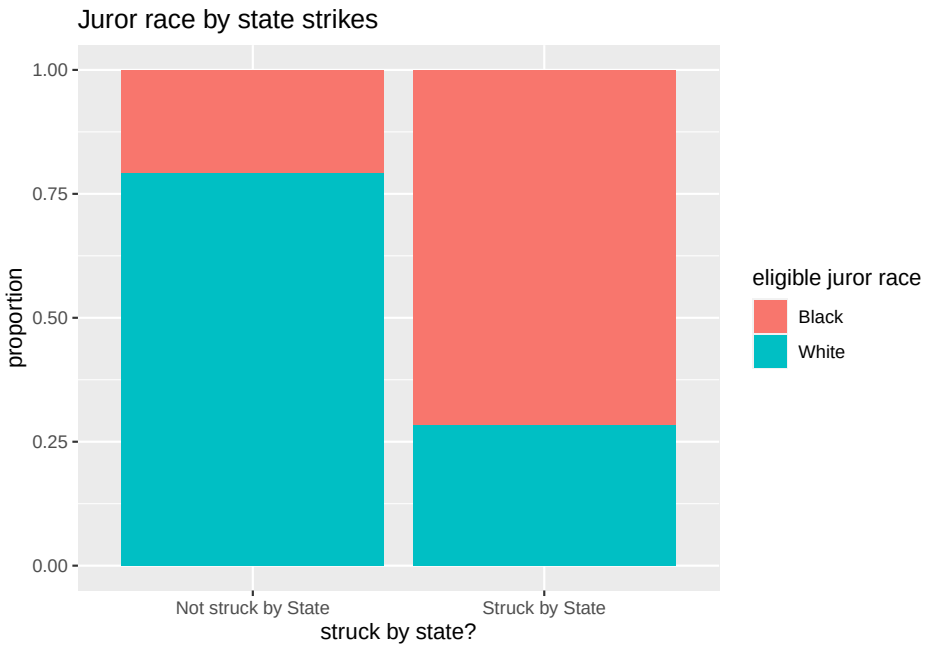
```
prop.table(mytable, margin = 2)
```

```
       Not struck by State Struck by State
  Black           0.2075646       0.7159353
  White           0.7924354       0.2840647
```

Notice that the proportions add to one **down** each column. Of all eligible jurors struck by the state, about 71.6% were black.

The stacked bar graph for this distribution is

```
ggplot(jurors, aes(x = struck_state, fill = race)) +
  geom_bar(position = "fill") +
  labs(title = "Juror race by state strikes", y = "proportion",
       fill = "eligible juror race", x = "struck by state?")
```



- What proportion of eligible jurors who were not struck by the state were black? were white?

Click for answer

Answer: Of all jurors not struck by the state, about 20.8% were black

---

11.2 Your Turn 2

11.2.1 Graduate programs acceptance and sex

How are grad school program acceptance rates associated with sex? We will look at a classic data set from Berkeley grad school applications from 1973 (*Science*, 1975). The data cases are applicants to four graduate programs at Berkeley during 1973. The variable `result` tells us if the applicant was accepted to the graduate program, `sex` tells us the sex of the applicant (male or female), and `program` tells us program type (programs 1,2,3 or 4).

```
grad <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Berkeley
```

```
# dimension of the dataset
dim(grad)
```

```
[1] 3014    3
```

```
# first 6 rows
head(grad)
```

```
   program  sex result
1 program1 male accept
2 program1 male accept
3 program1 male accept
4 program1 male accept
5 program1 male accept
6 program1 male accept
```

(a). Table of counts and proportions

```
prop.table(table(grad$result))
```

```
   accept    reject
0.4260119 0.5739881
```

- What proportion of applicants were accepted?

Click for answer

Answer: About 43% (1284/3014) of applicants were accepted.

(b). Two-way tables

The `table` command also gives two-way tables when two variables are included. Here is the two-way table for result and sex:

```
table(grad$sex, grad$result)
```

```
        accept reject
  female    262    587
  male     1022   1143
```

- How many applicants involved females who were accepted?

Click for answer

Answer: : 262 applicants involved females who were accepted.

(c). Conditional proportions: acceptance given sex

The `prop.table` command gives conditional proportions for a two-way table. First let's save the two-way table in an object named `mytable`:

```
mytable <- table(grad$sex, grad$result)
```

Then use `prop.table` to get the distribution of result conditioned (grouped) on applicant's sex:

```
prop.table(mytable, 1)
```

```
           accept    reject
  female 0.3085984 0.6914016
  male   0.4720554 0.5279446
```

The value of `1` in this command tell's R that you want *row* proportions (the denominator of the proportion is each row total).

- What proportion of female were accepted?

Click for answer

Answer: about 31% (262/(262+587))

- What proportion of males were accepted?

Click for answer

Answer: about 47% (1022/(1022+1143))

(d). Bar graph for one variable

We can create a data frame `count_data1` containing the counts and their corresponding categories.

```
counts1 <- table(grad$result)
count_data1 <- data.frame(counts1)
```

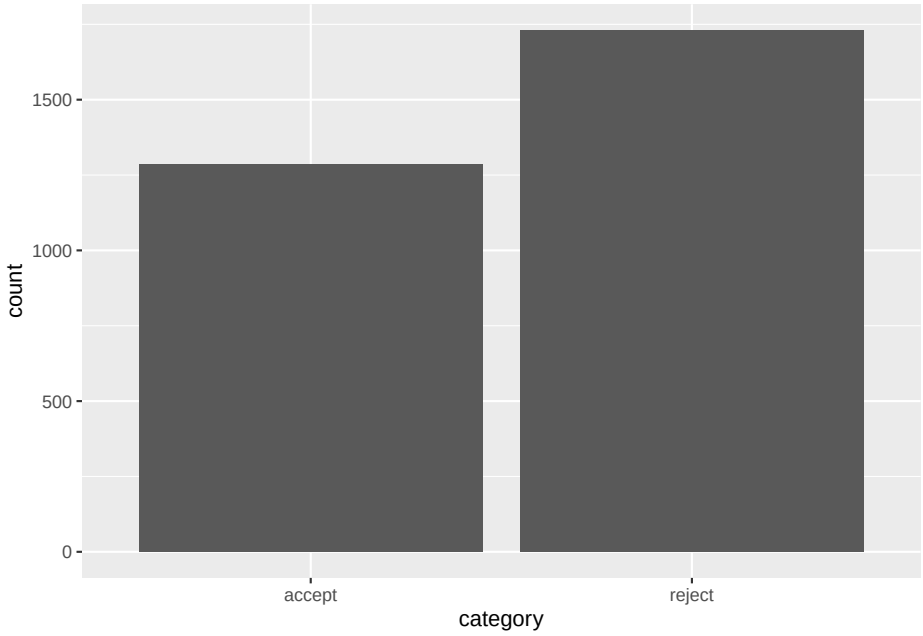
Then, we use `ggplot2` to create a bar plot with the categories on the x-axis and the counts on the y-axis. The column names of `count_data` are automatically assigned to be `Var1` and `Freq`. We can change the column names to `category` and `count`, for example, as:

```
colnames(count_data1) = c("category", "count")
count_data1
```

```
  category count
1   accept  1284
2   reject  1730
```

The `geom_bar(stat = "identity")` function is used to create the bars, and we set the y-axis label using `labs()`.

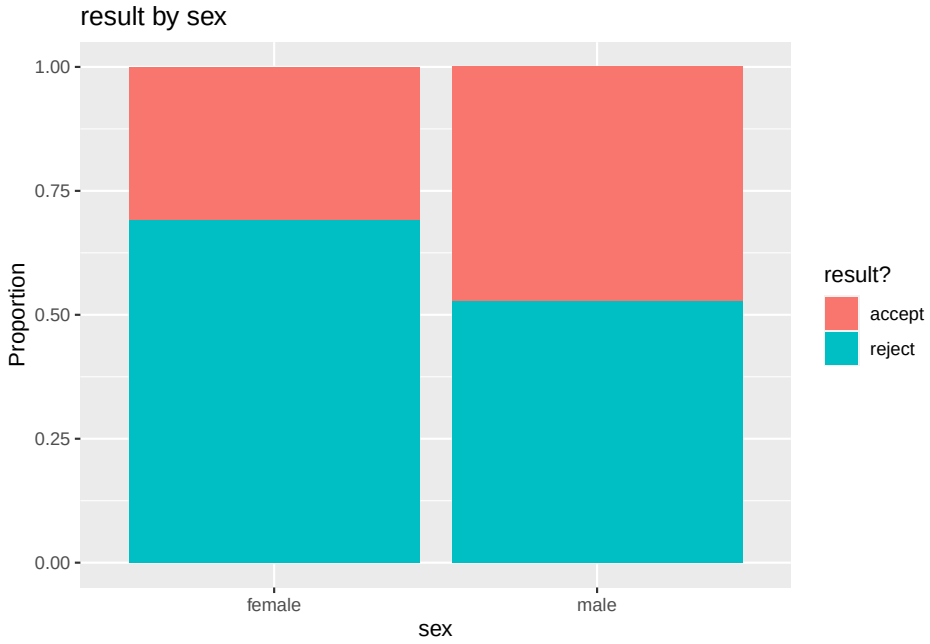
```
# Create a bar plot using ggplot2
library(ggplot2)  # load the package
ggplot(count_data1, aes(x = category, y = count)) +
  geom_bar(stat = "identity") +
  labs(y = "count")
```



(e). Stacked bar graph for two variables

Now we can use the `geom_bar` command in this package. Here we get the conditional distribution of `result` given `sex`:

```
library(ggplot2) # don't need if you already entered it for example 1
ggplot(grad, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y="Proportion", title = "result by sex", fill = "result?", x = "sex")
```



The basic syntax for this function is to let `ggplot` know your data set name (`grad`), then specify the grouping or conditional variable on the x-axis (`sex`) in the `aes` (aesthetic) argument. The `fill` variable is the response variable (`result`). We add (`+`) the `geom_bar` geometry to get a bar graph with the `fill` position specified. Adding an informative label and title complete the graph.

- Verify that this graph is plotting the conditional proportions from part (c)

(f). Subsetting by program type

Finally, we will repeat the previous analysis of result and sex, but this time we will divide (or subset) the data set by program type. To do this we need to know how the values of `program` are coded:

```
table(grad$program)
```

```
program1 program2 program3 program4
     933      585      782      714
```

Here we use the `filter` command available from the `dplyr` package to get only the applicants to program 1:

```
library(dplyr)
grad.p1 <- filter(grad, program == "program1")  # gets rows where program equal program1
head(grad.p1)
```

```
   program  sex result
1 program1 male accept
2 program1 male accept
3 program1 male accept
4 program1 male accept
5 program1 male accept
6 program1 male accept
```

```
dim(grad.p1)
```

```
[1] 933   3
```

Verify that the number of rows in the subsetted program 1 data set matches the number of program 1 applicants shown in the `table` of counts above.

- Repeat the `filter` command to get a data set for program 2 and call the new data set `grad.p2`. Verify that the number of rows in this dataset matches the number of program 2 applicants in the original data set.

```
# enter R code for (f) here
grad.p2 <- filter(grad, program == "program2") # gets rows where program equal program1
head(grad.p2)
```

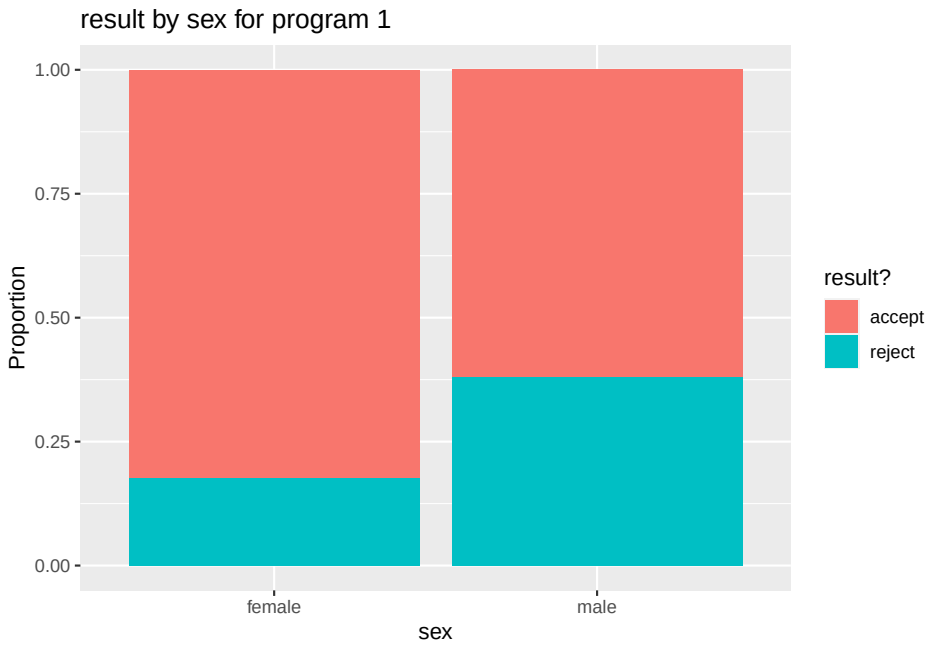
```
   program  sex result
1 program2 male accept
2 program2 male accept
3 program2 male accept
4 program2 male accept
5 program2 male accept
6 program2 male accept
```

(g). Result by sex for program 1.

- Show the distribution of result conditioned on applicant's sex for the program 1 data set. Get both a table of conditional proportions (or percentages) and a stacked bar graph.

Click for answer

```
# enter R code for (g) here
ggplot(grad.p1, aes(x = sex, fill = result)) +
 geom_bar(position = "fill") +
 labs(y="Proportion", title = "result by sex for program 1",
 fill = "result?", x = "sex")
```




```
prop.table(table(grad.p1$sex, grad.p1$result),1)
```

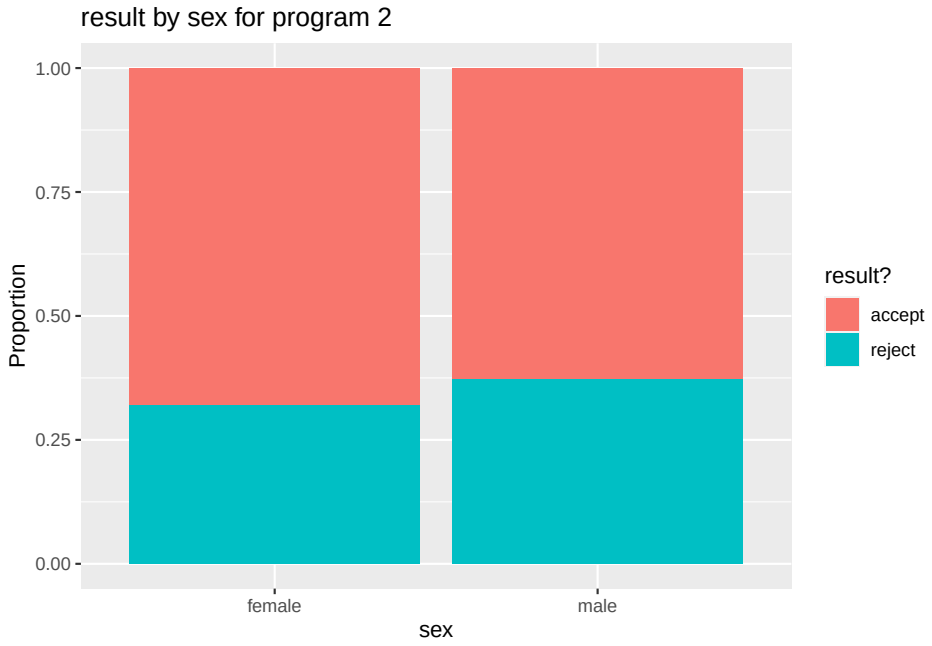
```
          accept    reject
female 0.8240741 0.1759259
male   0.6193939 0.3806061
```

(h). Result by sex for program 2.

- Repeat part (g) but this time use the program 2 data set. Compare the two bar graphs for (g) and (h) and explain how they show that females have a higher acceptance rate after accounting for program type (1 or 2).

Click for answer

```
# enter R code for (h) here
ggplot(grad.p2, aes(x = sex, fill = result)) +
 geom_bar(position = "fill") +
 labs(y="Proportion", title = "result by sex for program 2",
 fill = "result?", x = "sex")
```




```
prop.table(table(grad.p2$sex, grad.p2$result),1)
```

```
          accept    reject
female 0.6800000 0.3200000
male   0.6285714 0.3714286
```

Answer: For both programs 1 and 2, we see that female applicants have a slightly higher rate of acceptance than male applicants. After accounting for program type, we now see that black defendants have a higher rate of death penalty than white defendants. Without accounting for program type, the opposite was true (see parts (c) and (e)).

Why? the confounding affect of program type which is associated with both result and sex:

Click for answer

- females prefer to apply to programs 3 and 4 while males prefer programs 1 and 2 (more than 3 and 4).
  - 44% of females applied to program 3 and 40% to program 4
  - 38% of males applied to program 1 and 26% to program 2

```
prop.table(table(grad$sex, grad$program), 1)
```

```
          program1   program2   program3   program4
female 0.12720848 0.02944641 0.44169611 0.40164900
male   0.38106236 0.25866051 0.18799076 0.17228637
```

-Programs 3 and 4 were much harder to get into than programs 1 and 2 - 64% of applicants to program 1 were accepted and 63% of applicants to program 2 were accepted - 6% of applicants to program 4 were accepted and 34% of applicants to program 3 were accepted

```
prop.table(table(grad$program, grad$result), 1)
```

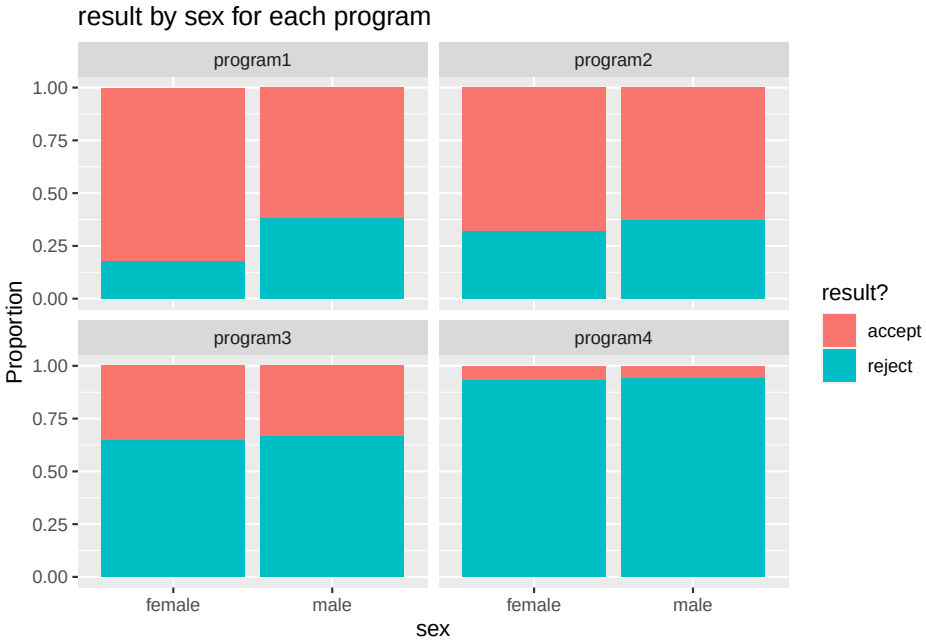
```
             accept     reject
program1 0.64308682 0.35691318
program2 0.63076923 0.36923077
program3 0.34398977 0.65601023
program4 0.06442577 0.93557423
```

So since the majority of females applied to the toughest programs (as measured by acceptance rates), there overall rate of acceptance was lower for females compared to males. But when we break down these rates by program type, we see that females have higher acceptance rates than males (see the visual in part (i)).

(i). A bar graph with three variables

If we simply want to graph the relationship between result and sex for each type of program, we can avoid subsetting the data by using the `facet_wrap` command in `ggplot2`. It is one simple addition to the stacked bar graph in part (e):

```
ggplot(grad, aes(x = sex, fill = result)) +
  geom_bar(position = "fill") +
  labs(y="Proportion",
       title = "result by sex for each program",
       fill = "result?",
       x = "sex") +
  facet_wrap(~program)
```



- Verify that this command creates side-by-side stacked bar graphs that match your graphs in parts (g) and (h) for programs 1 and 2.

Click for answer

Answer: The graphs match.

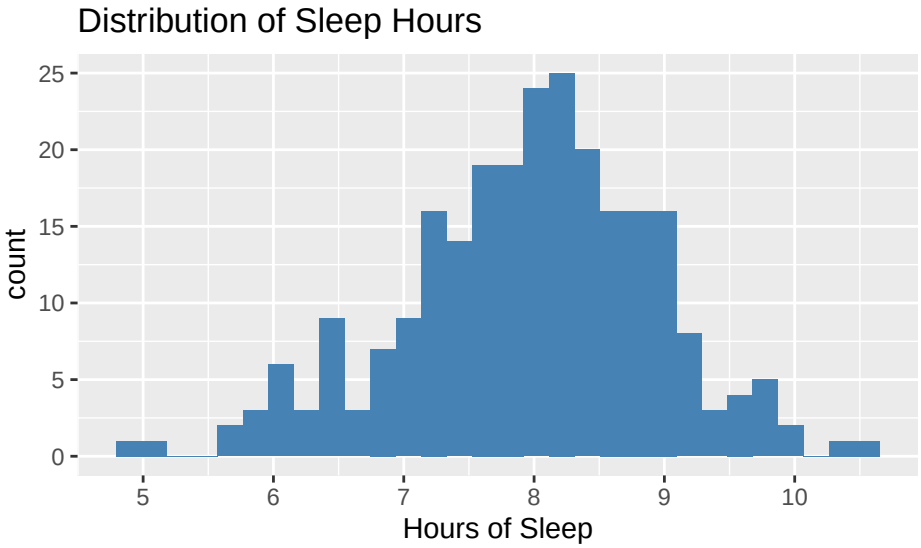
Chapter 12

Class Activity 5

12.1 Example 1: Sleep

This histogram shows the distribution of hours or sleep per night for a large sample of students.

```
library(ggplot2)
sleep <- read.csv("http://math.carleton.edu/Stats215/Textbook/SleepStudy.csv")
ggplot(sleep, aes(x=AverageSleep)) +
  geom_histogram(fill="steelblue", bins = 30) +
  labs(title = "Distribution of Sleep Hours", x = "Hours of Sleep")
```



12.1.1 (a) Estimate the average hours of sleep per night.

Click for answer

Answer: The mean is around 8 hours

12.1.2 (b) Use the 95% rule to estimate the standard deviation for this data.

Click for answer

Answer: Most of the data is between about 6 and 10, with a mean around 8 (due to the roughly symmetric distribution). So two standard deviations is about 2 hours of sleep, making one standard deviation about 1 hours of sleep.

Let's check the rule! Here are the actual mean and SD:

```
mean(sleep$AverageSleep)
```

```
[1] 7.965929
```

```
sd(sleep$AverageSleep)
```

```
[1] 0.9648396
```

12.2 Example 2: Z-scores for Test Scores

The ACT test has a population mean of 21 and standard deviation of 5. The SAT has a population mean of 1500 and a standard deviation of 325. You earned 28 on the ACT and 2100 on the SAT.

12.2.1 (a) Which test did you do better on?

Click for answer

Answer:

- ACT: The z-score for the score of 28 is z = (28 - 21)/5 = 1.4.
- SAT: The z-score for the score of 2100 is z = (2100 - 1500)/325 = 1.85.
- The SAT score is 1.85 standard deviations above average while the ACT score is only 1.4 standard deviations above. You did better on the SAT.

```
z_ACT <- (28 - 21) / 5
z_SAT <- (2100 - 1500) / 325
z_ACT
```

```
[1] 1.4
```

```
z_SAT
```

```
[1] 1.846154
```

12.2.2 (b) For each test, find the interval that is likely to contain about 95% of all test scores.

Click for answer

Answer:

- ACT: Two standard deviations is $2(5) = 10$. About 95% of ACT scores are between 21 - 10 = 11 and 21 + 10 = 31. This claim assumes that ACT scores follow a bell-shaped distribution.
- SAT: Two standard deviations is $2(325) = 650$. About 95% of SAT scores are between 1500 - 650 = 850 and 1500 + 650 = 2150. This claim assumes that SAT scores follow a bell-shaped distribution.

```
ACT_lower <- 21 - 2 * 5
ACT_upper <- 21 + 2 * 5
SAT_lower <- 1500 - 2 * 325
SAT_upper <- 1500 + 2 * 325
c(ACT_lower, ACT_upper)
```

```
[1] 11 31
```

```
c(SAT_lower, SAT_upper)
```

```
[1]  850 2150
```

12.3 Example 3: 5 number summaries

For the given vector of observations indicate whether the resulting data appear to be symmetric, skewed to the right, or skewed to the left.

(2, 10, 15, 20, 69, 34, 23, 2, 45)

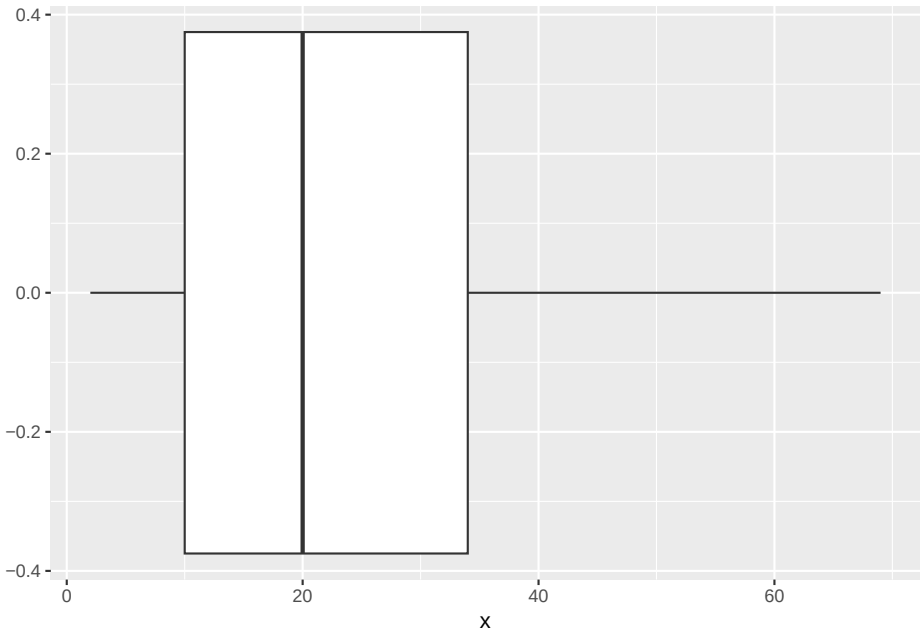
```
my_vector <- c(2, 10, 15, 20, 69, 34, 23, 2, 45)
summary(my_vector)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   10.00   20.00   24.44   34.00   69.00
```

Click for answer

Answer: Skewed right. It has a longer right tail than left since max -Q3 » Q1 - min

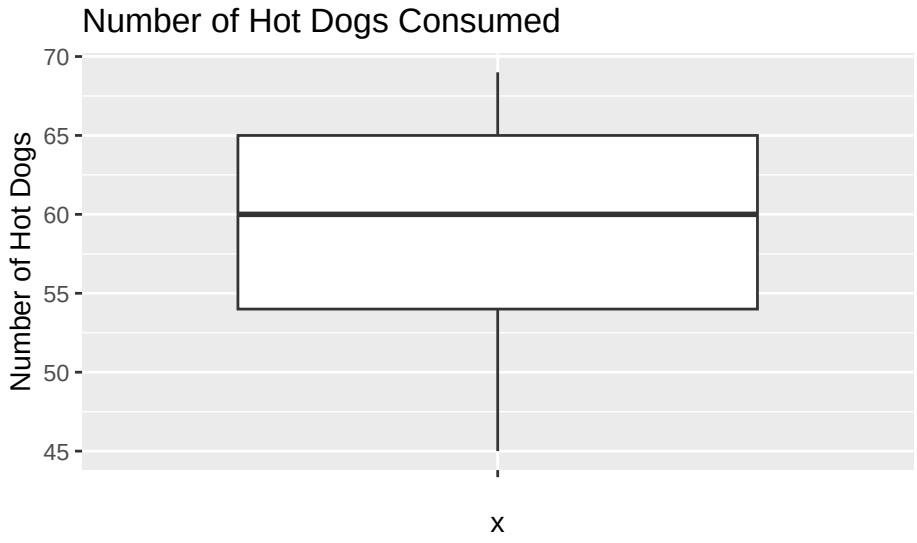
```
ggplot(data.frame(x=my_vector), aes(x)) + geom_boxplot()
```



12.4 Example 4: Hot dog

This boxplot shows the number of hot dogs eaten by the winners of Nathan's Famous hot dog eating contests from 2002-2011.

```
hotdogs <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/HotDogs.csv")
ggplot(hotdogs, aes(x = "", y = HotDogs)) +
  geom_boxplot() +
  labs(title = "Number of Hot Dogs Consumed", y = "Number of Hot Dogs")
```



12.4.1 (a) Use the boxplot to estimate the 5 number summary and IQR for this data. Verify that there are no outliers in this data.

Click for answer

Answer:

```
hotdog_q1 <- quantile(hotdogs$HotDogs, 0.25)
hotdog_q3 <- quantile(hotdogs$HotDogs, 0.75)
hotdog_iqr <- IQR(hotdogs$HotDogs)
lower_fence <- hotdog_q1 - 1.5 * hotdog_iqr
upper_fence <- hotdog_q3 + 1.5 * hotdog_iqr

library(dplyr)
outliers <- filter(hotdogs, HotDogs < lower_fence | HotDogs > upper_fence)
outliers
```

```
[1] Year    HotDogs
<0 rows> (or 0-length row.names)
```

12.5 Example 5: Hollywood Movies World Gross

Let's visit the `WorldGross` analysis from the Hollywood movies data set:

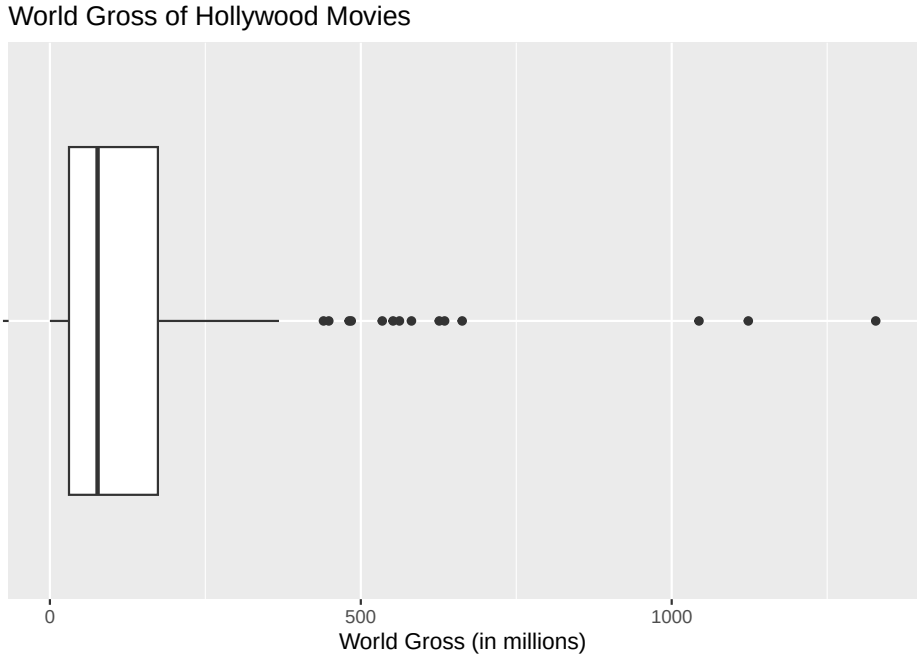
```
movies <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Hollyw
```

12.5.1 (a) Draw a boxplot of `WorldGross`.

Click for answer

Answer:

```
ggplot(movies, aes(x = WorldGross, y = "")) +
  geom_boxplot() +
  labs(title = "World Gross of Hollywood Movies", x = "World Gross (in millions)", y ="
```



How many movies are identified as outliers for world gross?

12.5.2 (b) Calculating boxplot values

Use the boxplot outlier rule to find the "fence" (cutoff) between an outlier and non-outlier for `WorldGross`. Then determine the value (of `WorldGross`) that

the upper "whisker" (non-outlier) extends to.

Click for answer

Answer:

```
library(tidyr)
movies_no_na <- drop_na(movies)
q1_world_gross <- quantile(movies_no_na$WorldGross, 0.25)
q3_world_gross <- quantile(movies_no_na$WorldGross, 0.75)
iqr_world_gross <- IQR(movies_no_na$WorldGross)
lower_fence_world_gross <- q1_world_gross - 1.5 * iqr_world_gross
upper_fence_world_gross <- q3_world_gross + 1.5 * iqr_world_gross

outliers <- filter(movies_no_na, WorldGross < lower_fence_world_gross | WorldGross > upper_fence_
outliers
```

```
                                      Movie
1   Harry Potter and the Deathly Hallows Part 2
2                          The Hangover Part II
3                       Twilight: Breaking Dawn
4                 Transformers: Dark of the Moon
5                                           Rio
6                 Rise of the Planet of the Apes
7                                    The Smurfs
8                                 Kung Fu Panda 2
9  Pirates of the Caribbean:\nOn Stranger Tides
10                           Mission Impossible
11                            Sherlock Holmes 2
12                                          Thor
13                                       Cars 2
                LeadStudio RottenTomatoes AudienceScore
1              Warner Bros             96            92
2       Legendary Pictures             35            58
3              Independent             26            68
4       DreamWorks Pictures             35            67
5          20th Century Fox             71            73
6          20th Century Fox             83            87
7  Sony Pictures Animation             23            50
8      DreamWorks Animation             82            80
9                  Disney             34            61
10               Paramount             93            86
11             Warner Bros             60            79
12                 Disney             77            80
13                  Pixar             38            56
               Story     Genre TheatersOpenWeek
```

```
1            Rivalry   Fantasy             4375
2             Comedy    Comedy             3615
3               Love   Romance             4061
4              Quest    Action             4088
5              Quest Animation             3826
6            Revenge    Action             3648
7  Fish Out Of Water Animation             3395
8            Rivalry Animation             3925
9              Quest    Action             4155
10           Pursuit    Action             3448
11           Pursuit    Action             3703
12     Monster Force    Action             3955
13 Fish Out Of Water Animation             4115
   BOAverageOpenWeek DomesticGross ForeignGross WorldGross
1              38672        381.01       947.10   1328.111
2              23775        254.46       327.00    581.464
3              34012        260.80       374.00    634.800
4              23937        352.39       770.81   1123.195
5              10252        143.62       341.02    484.634
6              15024        176.70       304.52    481.226
7              10489        142.61       419.54    562.158
8              12142        165.25       497.78    663.024
9              21697        241.07       802.80   1043.871
10              8672        197.80       336.70    534.500
11             10704        179.04       261.00    440.040
12             16618        181.03       267.48    448.512
13             16072        191.45       360.40    551.850
   Budget Profitability OpeningWeekend
1     125     10.624888         169.19
2      80      7.268300          85.95
3     110      5.770909         138.12
4     195      5.759974          97.85
5      90      5.384822          39.23
6      93      5.174473          54.81
7     110      5.110527          35.61
8     150      4.420160          47.66
9     250      4.175484          90.15
10    145      3.686207          29.55
11    125      3.520320          39.63
12    150      2.990080          65.72
13    200      2.759250          66.14
```

(c) Create a new dataset called `movies_no_outliers` that contains only the rows from `movies_no_na` where the `WorldGross` values are within the range defined by the lower and upper fences.

Click for answer

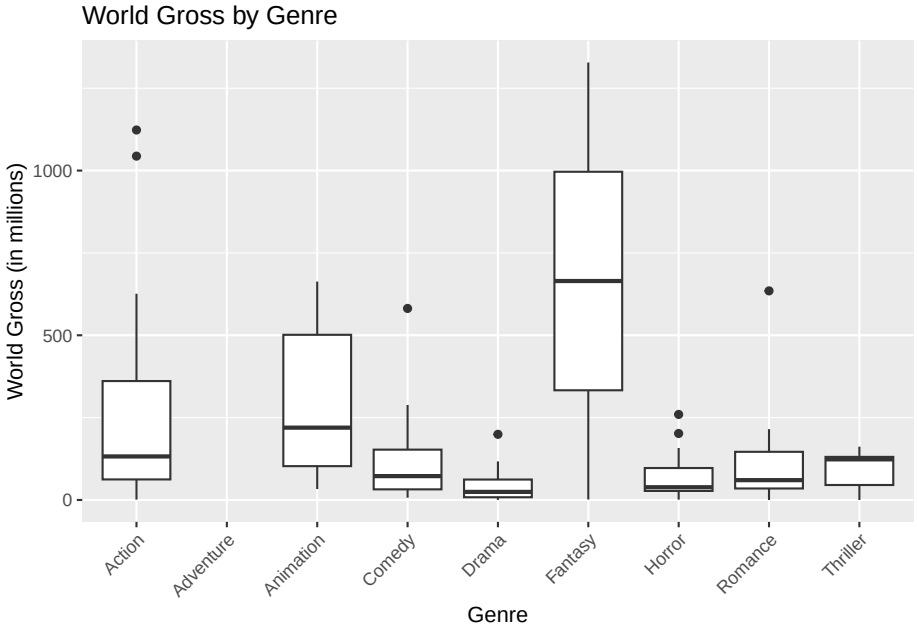
Answer:

```
library(dplyr)
movies_no_outliers <- filter(movies_no_na, WorldGross >= lower_fence_world_gross & WorldGross <=
```

12.5.3 (d) Side-by-side boxplot

We can compare boxplots of `WorldGross` across `Genre` categories:

```
ggplot(movies, aes(x = Genre, y = WorldGross)) +
  geom_boxplot() +
  labs(title = "World Gross by Genre", x = "Genre", y = "World Gross (in millions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- What does this type of graph illustrate well about the relationship between `WorldGross` and `Genre`?

Click for answer

Answer: Does a good job comparing median values and extremes

- What does this type of graph not illustrate well about the relationship between `WorldGross` and `Genre`?

Click for answer

Answer: It doesn't illustrate sample sizes well, e.g. the fantasy genre only has 2 movies in it

Chapter 13

Class Activity 6

13.1 Your Turn 1

13.1.1 Beer Example

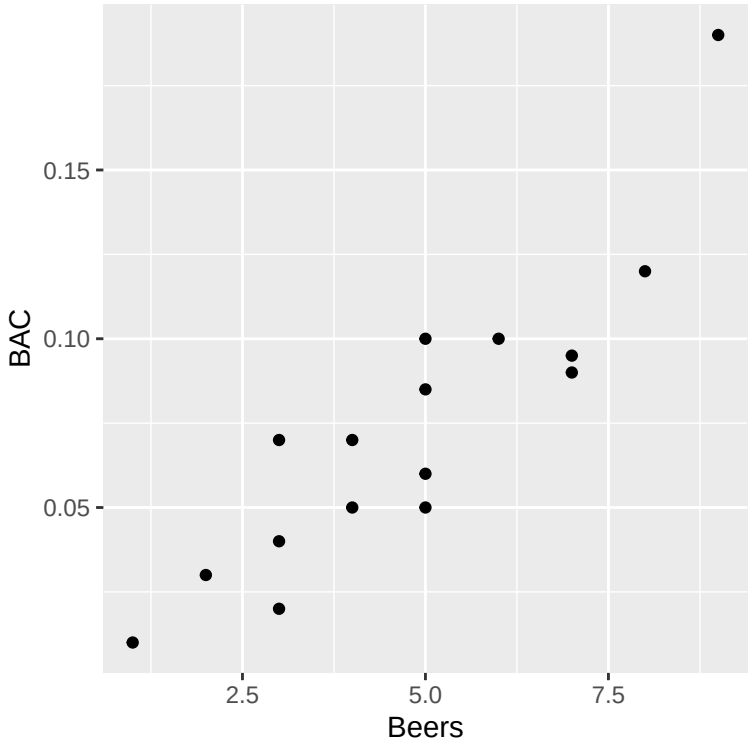
A study of 16 Ohio State University students looked at the relationship between the number of beers a student consumes and their blood alcohol content (BAC) 30 minutes after their last beer. The regression information from R to predict BAC from number of beers consumed is given below.

```
library(readr)
bac <- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/BAC.csv")
```

13.1.2 (a) Always start with a visual!!!!

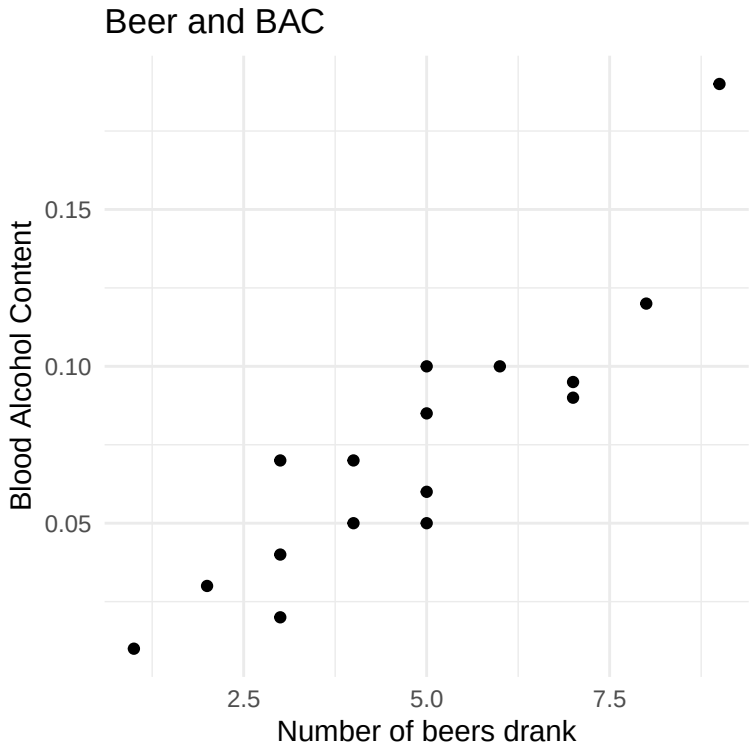
Plot the response (BAC) on the y-axis and the explanatory ("predictor") on the x-axis.

```
ggplot(data = bac, aes(x = Beers, y = BAC)) + geom_point()
```



You can modify this basic graph by adding a title and axes labels.

```
ggplot(data = bac, aes(x = Beers, y = BAC)) +
     geom_point(shape = 19) +
     labs(title = "Beer and BAC",
          x = "Number of beers drank",
          y = "Blood Alcohol Content") +
     theme_minimal()
```



- Is there a relationship?
  - direction?
  - strength?
  - form?

13.1.3 (b) Computing correlation

Since the *form* of the relationship is linear, we can use **correlation** to measure its strength:

```
cor(bac$BAC, bac$Beers)
```

```
[1] 0.8943381
```

If there are any missing values (`NA`'s) on either of the variables involved in the correlation calculation, use `use = "complete.obs"` as an extra argument to the `cor` function.

```
cor(bac$BAC, bac$Beers, use = "complete.obs")
```

```
[1] 0.8943381
```

13.1.4 (c) Fitting a regression line

We use the `lm(y ~ x, data=mydata)` function to fit a **l**inear (regression) **m**odel for a response y given an explanatory variable x. This command creates a **linear model object** that needs to be assigned a name, here we call it `bac.lm`. You can get the slope and intercept by typing out the object name:

```
bac.lm <- lm(BAC ~ Beers, data=bac)
bac.lm
```

```

Call:
lm(formula = BAC ~ Beers, data = bac)

Coefficients:
(Intercept)        Beers
   -0.01270      0.01796
```

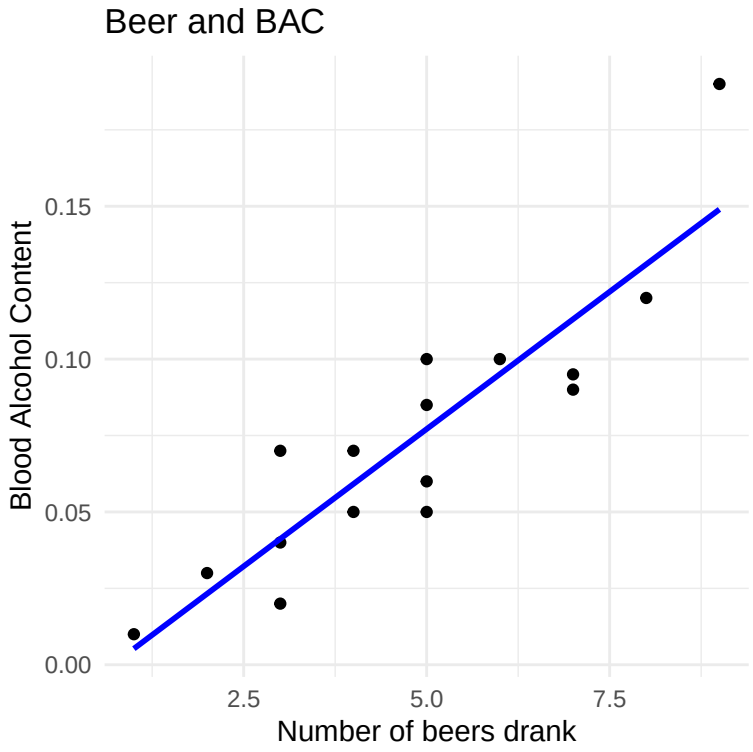
- After running the `lm` command above in your R console, check the **Environment** tab to see that the object `bac.lm` is now one of the objects stored in R's memory (for this session of Rstudio).
- Write down the fitted regression equation to predict BAC from number of beers.

Click for answer

Answer: $\hat{y} = ...$

- You can add this regression line to your scatterplot from part (a) by creating the plot and using the `abline` command:

```
# Customized scatter plot with regression line
ggplot(data = bac, aes(x = Beers, y = BAC)) +
     geom_point(shape = 19) +
     geom_smooth(method = "lm", se = FALSE, color = "blue") +
     labs(title = "Beer and BAC",
          x = "Number of beers drank",
          y = "Blood Alcohol Content") +
     theme_minimal()
```



13.1.5 (d) Interpret the slope in context.

Click for answer

Answer: Drinking one more beer is associated with a 0.0180 unit increase in predicted BAC.

13.1.6 (e) Interpret the intercept in context, if it makes sense to do so.

Click for answer

Answer: The intercept is -0.0127. A student who drinks 0 beers would be predicted to have a negative blood alcohol content. This is not possible so the intercept does not make sense in this context, but the intercept is included in the model to get the best fit line for the data collected.

13.1.7 (f) If your friend at Ohio State drank 2 beers, what would you predict their BAC to be?

Click for answer

Answer: The predicted BAC is

$$\widehat{BAC} = -0.0127 + 0.0180(2) = 0.0233.$$

```
y.hat <- -0.0127 + 0.0180*(2)
y.hat
```

```
[1] 0.0233
```

13.1.8 (g) Find the residual for the student in the dataset who drank 2 beers and had a BAC of 0.03.

Click for answer

Answer: The residual is

$$BAC - \widehat{BAC} = .03 - .0233 = 0.0067$$

```
0.03 - (-0.0127 + 0.0180*(2))
```

```
[1] 0.0067
```

13.1.9 (h) Getting residuals in R

We can use the `resid` command to get the residuals for each case in the data set:

```
# Residuals
residuals <- data.frame(Beers = bac$Beers, Residuals = resid(bac.lm))
residuals
```

```
   Beers    Residuals
1      5  0.022881795
2      2  0.006773080
3      9  0.041026747
4      8 -0.011009491
5      3 -0.001190682
```

```
6      7 -0.018045729
7      3  0.028809318
8      5 -0.017118205
9      3 -0.021190682
10     5 -0.027118205
11     4  0.010845557
12     6  0.004918033
13     5  0.007881795
14     7 -0.023045729
15     1  0.004736842
16     4 -0.009154443
```

13.1.10 (i) Getting R^2 value

You can use the `summary` command on an `lm` object to get a more detailed print out of your linear model, along with the R^2 value for your model:

```
summary(bac.lm)
```

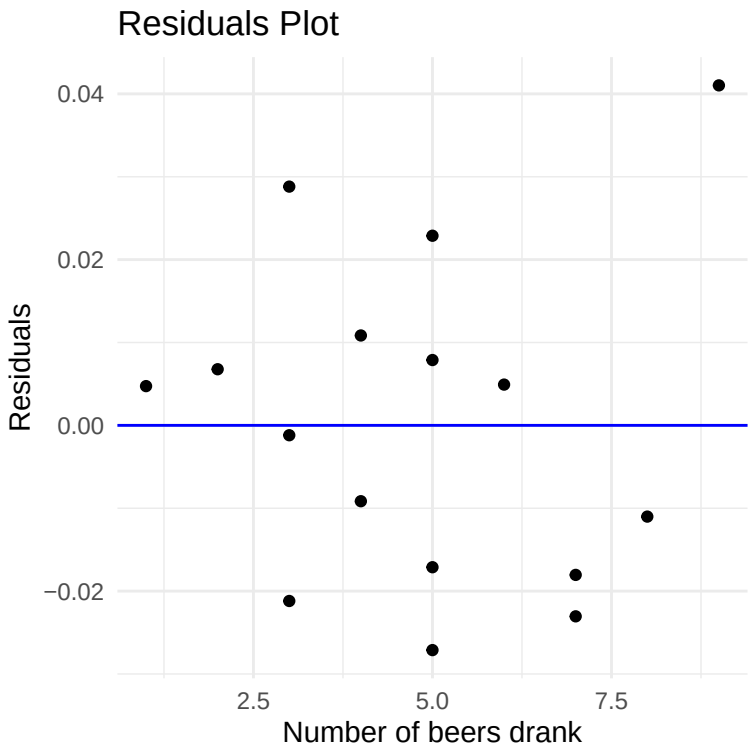
```
Call:
lm(formula = BAC ~ Beers, data = bac)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

13.1.11 (j) Making a residuals plot

The regression of `BAC` on `Beers` has a residuals plot that plots the model's residuals on the y-axis and the explanatory ("predictor") on the x-axis. We add

a horizontal reference line (the detrended regression line) with the `geom_hline()` command:

```
# code for residual plot
ggplot(data = residuals, aes(x = Beers, y = Residuals)) +
           geom_point(shape = 19) +
           geom_hline(yintercept = 0, color = "blue") +
           labs(title = "Residuals Plot",
                x = "Number of beers drank",
                y = "Residuals") +
           theme_minimal()
```



Interpret: There is one case of 9 beers with a large residual (much higher BAC than predicted), but since there is no clear pattern (trend) in this plot it looks like our regression model adequately describes the relationship between number of beers and BAC.

- Is the magnitude of the scatter around the horizontal 0-line in the residuals plot greater than, less than, or the same as the magnitude of the scatter around the regression line in the scatterplot?

Click for answer

Answer: The same! The residuals plot is only a "detrended" scatterplot, meaning the vertical distances between a point and the regression line on the scatterplot or a point and the 0-line on the residuals plot are exactly the same. The residual plot looks more scattered because the trend is removed and the scale of the y-axis compressed.

13.1.12 (k) Identifying points

We can use the functions `filter` and `row_number` from the `dplyr` package to find the index of Beers equal to 9.

```
# Use `which` to find the index of Beers equal to 9
index <- which(bac$Beers == 9)
index
```

```
[1] 3
```

Click for answer

Answer: Row 3.

What is the row number of the case with the most negative residual? We could eyeball the graph to see that the most negative residual is less than -0.02:

```
# Find residuals less than -0.02
resid_less_than_neg_002 <- resid(bac.lm) < -0.02
resid(bac.lm)[resid_less_than_neg_002]
```

```
          9          10          14
-0.02119068 -0.02711821 -0.02304573
```

But this identifies 3 cases. We also can see that the lowest residual drank 5 beers. We can add this statement to the original one using the "and" sign `&`:

```
# which case had resid less than -0.02 AND drank 5 beers
resid(bac.lm)[which(resid(bac.lm) < -0.02 & bac$Beers == 5)]
```

```
         10
-0.02711821
```

13.1.13 (l) Checking outlier influence

Will the regression line slope increase, decrease or stay the same if we remove case 3, the 9 beer case, from our model?

Check your answer by adding `subset = -3` to the `lm` command (this removes row 3):

```
# define a different linear model with row 3 removed
bac.lm2 <- lm(BAC ~ Beers, data=bac, subset = -3)
```

```
# Compare the two models
summary(bac.lm2)
```

```

Call:
lm(formula = BAC ~ Beers, data = bac, subset = -3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.023685 -0.010068 -0.003685  0.011985  0.027208

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.481e-05  1.088e-02   0.002    0.998
Beers       1.455e-02  2.216e-03   6.568  1.8e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01624 on 13 degrees of freedom
Multiple R-squared:  0.7684,    Adjusted R-squared:  0.7506
F-statistic: 43.14 on 1 and 13 DF,  p-value: 1.802e-05
```

```
summary(bac.lm)
```

```

Call:
lm(formula = BAC ~ Beers, data = bac)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

- After removing case 3, how has the slope changed? Explain the why the change occurred.

Click for answer

Answer: The slope drops from 0.0180 to 0.0146. Explanation given above.

- After removing case 3, how has the R^2 changed? Explain the why the change occurred.

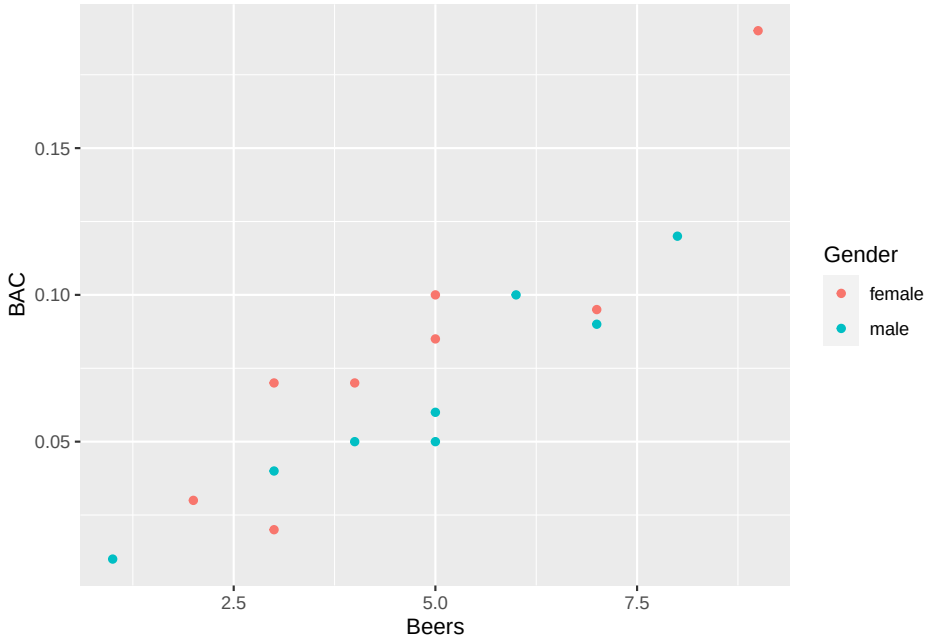
Click for answer

Answer: The R^2 descreases from 79.9% to 76.8%. This small decrease happens because case 3 actually enhances the overall linear trend and removing it results is a slight decrease to correlation and R^2.

13.1.14 (m) Adding a categorical variable to your plot

We can create a scatterplot with plotting symbols color coded by a categorical grouping variable using `ggplot2` package. We use the `geom_point()` plot geometry to get a scatterplot with the `x`, `y`, and `color` aesthetics specified. Here we look at the `BAC` vs. `Beers` plot with `Gender` added:

```
ggplot(bac, aes(x=Beers, y=BAC, color=Gender)) + geom_point()
```



- Are the associations similiar? (form, strength, direction)

Click for answer

Answer: Both females and males have similar strong, positive linear associations.

13.1.15 (n) Regression lines by groups

A quick way to get the male and female regression line formulas for part (c) is to add a `subset` argument to the `lm` command:

```
# Fit linear regression model for female
bac.lm.female <- lm(BAC ~ Beers, data = bac, subset = Gender == "female")
bac.lm.female
```

```
Call:
lm(formula = BAC ~ Beers, data = bac, subset = Gender == "female")

Coefficients:
(Intercept)        Beers
   -0.01567      0.02067
```

```
# Fit linear regression model for male
bac.lm.male <- lm(BAC ~ Beers, data = bac, subset = Gender == "male")
bac.lm.male
```

```
Call:
lm(formula = BAC ~ Beers, data = bac, subset = Gender == "male")

Coefficients:
(Intercept)        Beers
  -0.009785     0.015341
```

- What is the regression line for females? for males?

Click for answer

Answer: For females: $\widehat{BAC} = -0.016 + 0.021(BAC)$ and for males: $\widehat{BAC} = -0.01 + 0.015(BAC)$

- Which gender has the largest slope? What does this suggest about the relationship between number of beers and BAC for this gender?

Click for answer

Answer: The slope for females is slightly higher. This shows that the effect of one more beer on predicted BAC in females is larger than males (a 0.021 increase vs. a 0.015 increase).

Another way to obtain regression models by `Gender` is to split the data set in a female and male data set, then run your `lm` on these two data sets. The benefit of this method is you can then create a residuals plot for your model much easier than the quicker method above:

```
# Filter data for female
bac_female <- filter(bac, Gender == "female")

# Fit linear regression model for female
bac_lm_female <- lm(BAC ~ Beers, data = bac_female)
bac_lm_female
```

```
Call:
lm(formula = BAC ~ Beers, data = bac_female)

Coefficients:
(Intercept)        Beers
   -0.01567      0.02067
```

```r
# Filter data for male
bac_male <- filter(bac, Gender == "male")

# Fit linear regression model for male
bac_lm_male <- lm(BAC ~ Beers, data = bac_male)
bac_lm_male
```

```
Call:
lm(formula = BAC ~ Beers, data = bac_male)

Coefficients:
(Intercept)        Beers
  -0.009785     0.015341
```

Chapter 14

Class Activity 7

14.1 Your Turn 1

14.2 Example 1: Using Search Engines on the Internet

A 2012 survey of a random sample of 2253 US adults found that 1,329 of them reported using a search engine (such as Google) every day to find information on the Internet.

14.2.1 a). Find the relevant proportion and give the correct notation with it.

Click for answer

Answer: $\hat{p} = 1329/2253$

```
p.hat <- 1329/2253
p.hat
```

```
[1] 0.5898802
```

14.2.2 b). Is your answer to part (a) a parameter or a statistic?

Click for answer

Answer: Statistic

14.2.3 c). Give notation for and define the population parameter that we estimate using the result of part (a).

Click for answer

Answer: p = the proportion of all US adults that would report that they use an Internet search engine every day

14.3 Example 2: Bootstrapping mean

Let's use R to perform bootstrapping and visualize the distribution of the sample mean. We need to load the purrr library to do more effective simulation. Create a vector X containing the data points:

```
X <- c(20, 24, 19, 23, 22, 16)
```

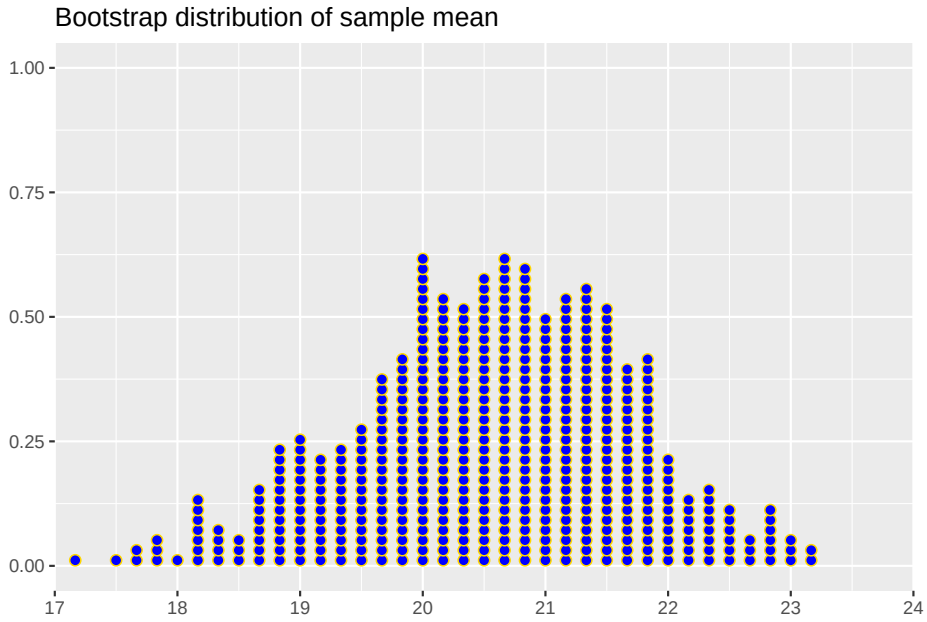
14.3.1 a. Generate 500 bootstrapped samples of the data, calculate the mean for each sample, and store the results in a tibble:

```
bootstrapped_means <- tibble(
  iteration = 1:500,
  mean = map_dbl(iteration,
                 ~mean(sample(X, replace = TRUE)))
)
```

14.3.2 b. Create a dot plot to visualize the distribution of the bootstrapped sample means:

```
ggplot(bootstrapped_means, aes(x = mean)) +
  geom_dotplot(dotsize = 0.7,
               stackratio = 0.9,
               binwidth = .13,
               color = "gold",
               fill = "blue") +
  ggtitle("") + xlab("") + ylab("") +
  scale_x_continuous(limits = c(17, 24),
                     expand = c(0, 0),
```

```
                    breaks = seq(17, 24, 1)) +
  labs(title = "Bootstrap distribution of sample mean")
```



Question: What does each dot represent?

Click for answer

Answer: One sample mean from the bootstrapped sample.

Question: What is the shape of your sampling distribution?

Click for answer

Answer: Roughly symmetric.

Question: Where is your distribution centered?

Click for answer

Answer: About 20.5

Question: The distribution should be centered at the original sample mean. Verify it. Do we know the population mean? If not, what does it tell us about the center of this distribution.

Click for answer

Answer: It is close to the original sample mean. We do not know the population mean, so the bootstrap distribution will carry the bias of the original sample mean.

```
# r-code
mean(bootstrapped_means$mean)
```

```
[1] 20.54333
```

```
mean(X)
```

```
[1] 20.66667
```

Question: What is the standard deviation of this distribution? (Hint: use the 95% rule.)

Click for answer

Answer: About 1.25, it looks like most of the bootstrapped sample means are between 18 to 23 so 2 standard deviations is about 2.5. This makes the SD about 1.25.

Question: The standard deviation of sampling distribution has a separate name. It is called the **Standard Error**. Verify the standard deviation of this distribution using R-code.

Click for answer

Answer: It's close.

```
# r-code
sd(bootstrapped_means$mean)
```

```
[1] 1.120948
```

14.4 Example 3: Simulation of a Sample Proportion

According to a PEW survey, 66% of U.S. adult citizens casted a ballot in the 2020 election. Suppose we take a random sample of $n = 100$ eligible U.S. voters and computed the sample proportion who voted.

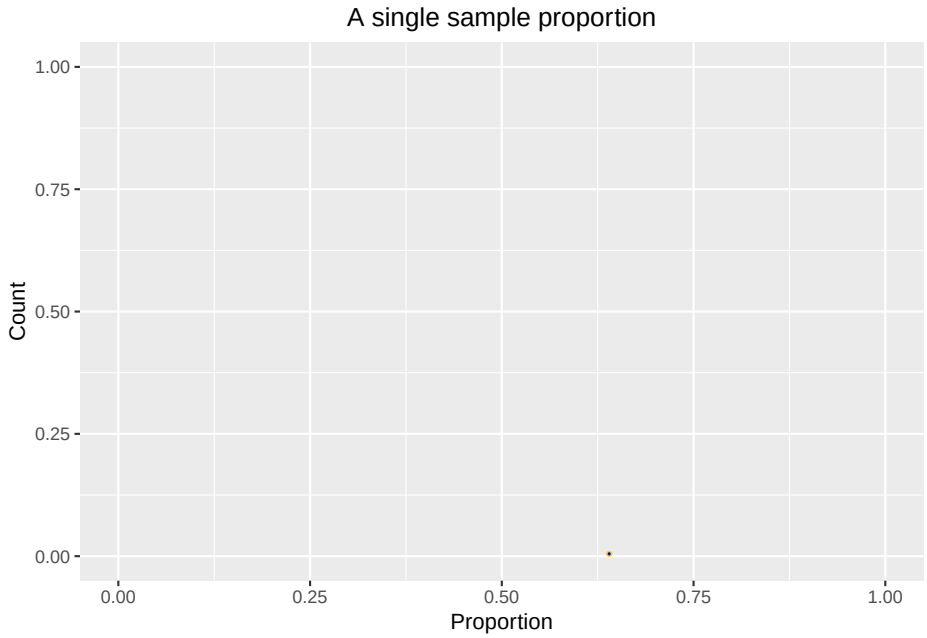
```
# Call the library
library(ggplot2)
```

```
# Define parameters
pop.prop <- .66 # Population proportion
n.size <- 100  # sample size
```

14.4.1 (a) Generate a random sample of size $n = 100$ and plot its sample proportion.

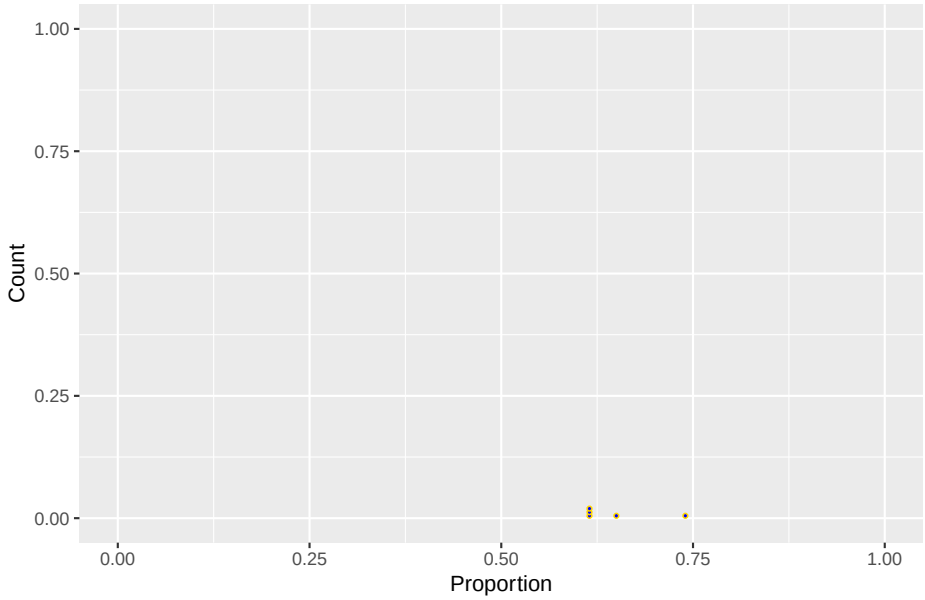
```
# Generate 1 sample
sample1 <- rbinom(n = 1, size = n.size, p = pop.prop) # R simulates the samples
sample.prop1 <- sample1/n.size #  Proportion = No. of Success / Sample Size
```

```
# define a data frame
mydata <- data.frame(x = sample.prop1)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = x)) +
   geom_dotplot(dotsize = 0.25,
                stackratio = 0.75,
                binwidth = .025,
                color = "gold",
                fill = "blue") +
  ggtitle("A single sample proportion") +  xlab("Proportion") + ylab("Count")+
  scale_x_continuous(limits = c(0, 1))+
  theme(plot.title = element_text(hjust = 0.5))
```



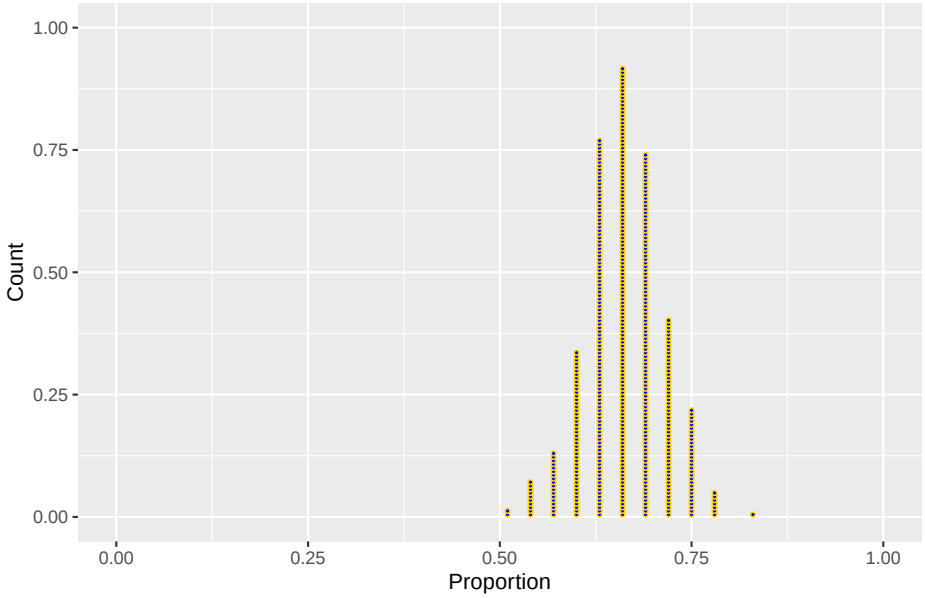
14.4.2 (b) Generate 5 random samples of size $n = 100$ and plot the sample proportions.

```
# generate 5 random samples of size 100
sample5 <- rbinom(n = 5, size = n.size, p = pop.prop)
sample.prop5 <- sample5/n.size

data <- data.frame(x = sample.prop5)

ggplot(data, aes(x = x)) +
   geom_dotplot(dotsize = 0.25,
                stackratio = 0.75,
                binwidth = .025,
                color = "gold",
                fill = "blue") +
  ggtitle("") +  xlab("Proportion") + ylab("Count")+
  scale_x_continuous(limits = c(0, 1))+
  theme(plot.title = element_text(hjust = 0.5))
```



14.4.3 (c) Generate 500 random samples of size $n = 100$ and plot the sample proportions.

```
# Generate 500 samples
set.seed(143)
sample500 <- rbinom(n = 500, size = n.size, p = pop.prop)
sample.prop500 <- sample500/n.size

data <- data.frame(x = sample.prop500)

ggplot(data, aes(x = x)) +
  geom_dotplot(dotsize = 0.25,
               stackratio = 0.75,
               binwidth = .025,
               color = "gold",
               fill = "blue") +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0, 1))+
  theme(plot.title = element_text(hjust = 0.5))
```



Question: What does each dot represent?

Click for answer

Answer: One sample proportion from a sample of n=100 eligible voters.

Question: What is the shape of your sampling distribution?

Click for answer

Answer: Roughly symmetric.

Question: Where is your distribution centered?

Click for answer

Answer: About 0.66, which is the population proportion.

Question: The distribution should be centered at the population proportion. Verify that the distribution is centered around the population proportion, $p = 0.66$.

Click for answer

Answer:

```
# r-code
mean(sample.prop500)
```

```
[1] 0.66298
```

Question: What is the standard deviation of this distribution? (Hint: use the 95% rule.)

Click for answer

Answer: About 0.05, it looks like most sample proportions are between 0.55 to 0.75 so 2 standard deviations is about 0.10. This makes the SD about 0.05.

Question: The standard deviation of sampling distribution has a separate name. It is called the **Standard Error**. Verify the standard deviation of this distribution using R-code.

Click for answer

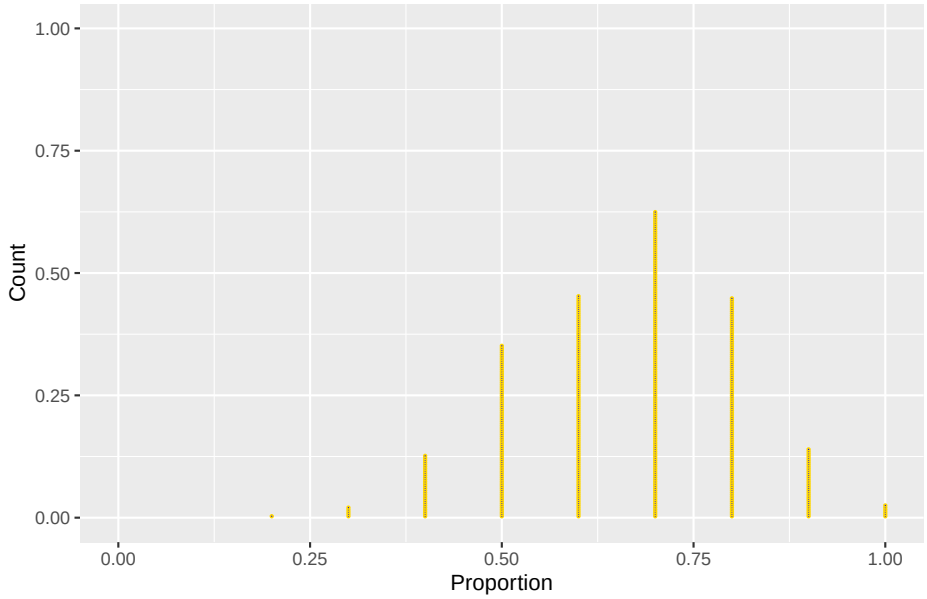
Answer:

```
# r-code
sd(sample.prop500)
```

```
[1] 0.0494673
```

(d) Repeat part(c) with sample size 20 instead of 100. Generate 500 samples.

```
n.size <- 10
pop.prop <- .66 # Population proportion

sample500_size10 <- rbinom(n = 500, size = n.size, p = pop.prop)
sample.prop500_size10 <- sample500_size10/n.size

data_size10 <- data.frame(x = sample.prop500_size10)

ggplot(data_size10, aes(x = x)) +
 geom_dotplot(dotsize = 0.25,
               stackratio = 0.75,
               binwidth = .015,
               color = "gold",
               fill = "blue") +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0, 1))+
  theme(plot.title = element_text(hjust = 0.5))
```



Question: How has the sampling distribution changed? (Shape? Center? Variability?)

Click for answer

Answer: The shape is slightly left skewed, still centered at 0.66 but with more variability that before (SD of about 0.10). This distribution is more discrete looking because there are just a few sample proportions possible with n=20 (e.g. 20/20, 19/20, 18/20, etc).

```
mean(sample.prop500_size10)
```

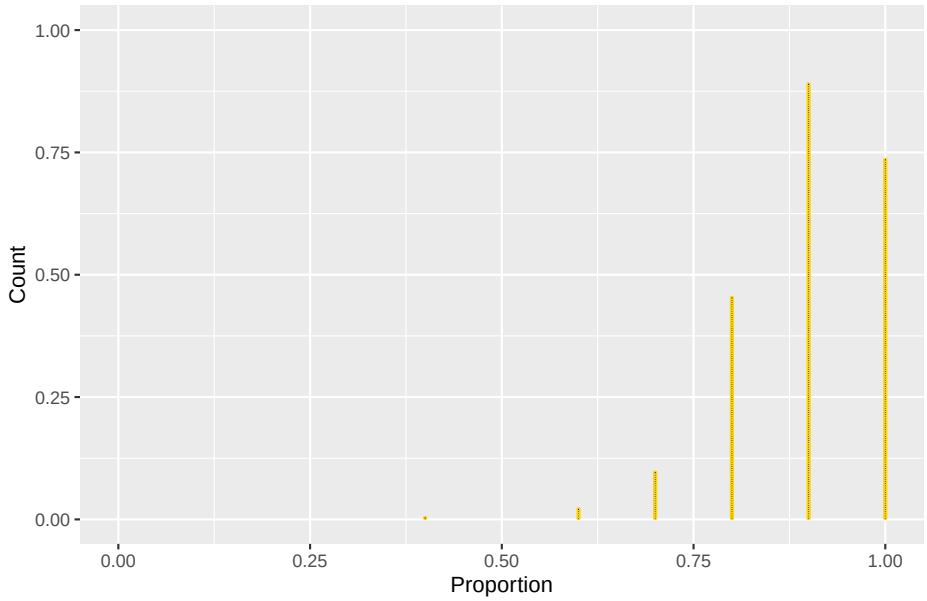
```
[1] 0.6618
```

```
sd(sample.prop500_size10)
```

```
[1] 0.1415657
```

(d) Now suppose the population proportion is $p = 0.90$ instead of $p = 0.66$ in part (e). Keep n.size=10.

```
pop.prop.large <- 0.90
n.size <- 10

sample500_size10_large_p <- rbinom(n = 500, size = n.size, p = pop.prop.large)
sample.prop500_size10_large_p <- sample500_size10_large_p/n.size

data_size10_large_p <- data.frame(x = sample.prop500_size10_large_p)

ggplot(data_size10_large_p, aes(x = x)) +
 geom_dotplot(dotsize = 0.25,
               stackratio = 0.75,
               binwidth = .015,
               color = "gold",
               fill = "blue") +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0, 1))+
  theme(plot.title = element_text(hjust = 0.5))
```



Question: How has the sampling distribution changed? (Shape? Center? Variability?)

Click for answer

Answer: The shape is much more left skewed than when p=0.66. Center is around 0.90 and SD is around 0.07. Note that increasing the population proportion closer to 1 results in a decrease in the SD because most samples give proportion near 1.

```
mean(sample.prop500_size10_large_p)
```

```
[1] 0.9
```

```
sd(sample.prop500_size10_large_p)
```

```
[1] 0.09261293
```

Chapter 15

Class Activity 8

15.1 Example 1: Textbook Prices

Prices of a random sample of 10 textbooks (rounded to the nearest dollar) are shown:

$132 $87 $185 $52 $23 $147 $125 $93 $85 $72

15.1.1 (a). What is the sample mean? Verify using r-code.

Click for answer

Answer: The sample mean is $\bar{x} = 100.1$

```
prices <- c(132,87, 185, 52, 23, 147, 125, 93, 85, 72)
mean(prices)
```

```
[1] 100.1
```

15.1.2 (b). Describe carefully how we could use cards to create one bootstrap statistic from this sample. Be specific.

Click for answer

Answer: We use 10 cards and write the 10 sample values on the cards. We then mix them up and draw one and record the value on it and put it back. Mix

them up again, draw another, record the value, and put it back. Do this 10 times to get a "with replacement" sample of size 10. Then compute the sample mean of this bootstrap sample.

15.1.3 (c). We can easily instruct R to do this with a simple code as follows:

```
resample <- sample(prices, replace = TRUE)
resample
```

```
 [1] 125 125 147  72  52 147  85  93  23 125
```

15.1.4 (d). Where will be bootstrap distribution be centered? What shape do we expect it to have?

Click for answer

Answer: It will be centered approximately at the sample mean of 100.1 and we expect it to be roughly bellshaped (it may be a bit skewed since the sample size of 10 is smallish).

15.2 Example 2: Statkey Atlanta Commute Distance

Go to the website at Lock5Statkey. Under the "Bootstrap Confidence Intervals" column, select the "CI for Single Mean, Median, St.Dev". Change the data set to Atlanta Commute (Distance). This data set gives a random sample of 500 worker commute distances (miles) for metropolitan Atlanta

15.2.1 (a). Use the "Original Sample" pane to determine the shape of these 500 commuter distances, along with their mean and standard deviation. Write down these stats using correct notation.

Click for answer

Answer: The sample mean is $\bar{x} = 18.16$ and the sample standard deviation is $s = 13.798$.

15.2.2 (b). Click "Generate 1 Sample" to create one bootstrap sample from this data. Explain how this sample was generated. Use the "Bootstrap Sample" pane to find the bootstrap statistic that was computed from this sample. What value is this bootstrap statistic? Repeat this a couple times.

Click for answer

Answer: The bootstrap sample was obtained by resampling from the 500 observed commute distances with replacement. Basically we randomly select 500 distances from the data (with replacement).

The value of the bootstrap mean will vary.

15.2.3 (c). Now click the "Generate 1000 Samples" to get 1000 bootstrap sample means. Is the bootstrap distribution centered at the population or sample mean commute distance?

Click for answer

Answer: The bootstrap distribution is always centered around the statistic that is being bootstrapped. Here it will be centered around the sample mean commute distance of about 18.16 miles. The population mean commute distance is unknown!

15.2.4 (d). What is the bootstrap SE for the sample mean?

Click for answer

Answer: The standard error from the bootstrap distribution is about 0.628.

15.2.5 (e). Compute a 95% confidence interval for the average commute distance in metropolitan Atlanta.

Click for answer

Answer: The sample mean is $\bar{x} = 18.16$ and the standard error from the bootstrap distribution is about 0.618 so we compute the 95% confidence interval using $18.16 \pm 2(0.628)$, giving an interval of 16.90 to 19.42 miles.

15.2.6 (f). Interpret your answer to (e) in context.

Click for answer

Answer: We are 95% confident that the average commuting distance in metropolitan Atlanta is between 16.90 and 19.42 miles.

15.3 Example 3: Statkey Global Warming

What percentage of Americans believe in global warming? A survey on 2,251 randomly selected individuals conducted in October 2010 found that 1,328 answered Yes to the question "Is there solid evidence of global warming?" To compute a bootstrap confidence interval for the proportion of all Americans who believe in global warming, go to the website at Lock5Statkey. Under the "Bootstrap Confidence Intervals" column, select the "CI for Single Proportion".

15.3.1 (a). Enter the data for this survey by clicking the "Edit Data" button. Enter 2251 as the sample size and 1328 as the count. What is the sample proportion of people who believe in global warming? Use correct notation!

Click for answer

Answer: The sample proportion is $\hat{p} = 0.59$.

15.3.2 (b). Generate 1 bootstrap sample. Explain how this sample was generated. Use the "Bootstrap Sample" pane to find the bootstrap statistic that was computed from this sample. What value is this bootstrap statistic? Repeat this a couple times.

Click for answer

Answer: The bootstrap sample was obtained by resampling the observed answers ("yes" and "no") to the global warming question with replacement. Answers will vary for the bootstrap statistic (proportion)

15.3.3 (c). Generate 1000 samples to get 1000 bootstrap sample proportions. Is the bootstrap distribution centered at the population or sample proportion? Describe the shape and center of this bootstrap distribution

Click for answer

Answer: The shape is symmetric around a center value of about 0.59, which is the sample proportion not the population proportion (which is unknown).

15.3.4 (d). Compute a 95% confidence interval for the proportion of Americans who believe in global warming

Click for answer

Answer: The sample proportion is $\hat{p} = 0.59$ and the standard error from the bootstrap distribution is 0.010 so we compute the 95% confidence interval using $0.590 \pm 2(0.010)$, giving an interval of 0.57 to 0.61.

15.3.5 (e). Interpret your interval from part (d).

Click for answer

Answer: We are 95% confident that the proportion of Americans who believe there is solid evidence of global warming is between 0.57 and 0.61.

15.3.6 (f). Does this data support a claim that a majority of Americans believe there is solid evidence of global warming? Explain.

Click for answer

Answer: Yes, the data does support this claim since we are confident that at least 50% of Americans believe in global warming since the lower bound on the CI is 57%.

15.4 Example 4. Statkey Global Warming by Political Party

Does belief in global warming differ by political party? When the question "Is there solid evidence of global warming?" was asked, the sample proportion answering "yes" was 79% among Democrats and 38% among Republicans. To compute a bootstrap confidence interval for the difference in the proportion of Democrats and Republicans who believe in global warming, go to the website at Lock5Statkey. Under the "Bootstrap Confidence Intervals" column, select the "CI for Difference in Proportions".

15.4.1 (a). Enter the data for this survey by clicking the "Edit Data" button. One big assumption we will make is that the sample sizes for both groups (Dems and Reps) were each 1000. Enter the Democrat data into the "Group 1" boxes (count of 790 and size of 1000) and the Republican data into the "Group 2" boxes (count of 380 and size of 1000). Verify that the sample proportions for the two groups are 79% and 38%. What is the difference in the two sample proportions? Use correct notation.

Click for answer

Answer: The sample difference in proportions is $\hat{p}_{Dem} - \hat{p}_{Rep} = 0.79 - 0.38 = 0.41$

15.4.2 (b). Generate 1 bootstrap sample. Explain how this sample was generated (give this some thought now that you have two samples of data). Use the "Bootstrap Sample" pane to find the bootstrap statistic that was computed from this sample. What value is this bootstrap statistic? Repeat this a couple times.

Click for answer

Answer: One bootstrap sample was obtained from the group 1 sample (resampling the observed "believe/not believe" responses with replacement) and a separate bootstrap sample was obtained from the group 2 sample. The difference in the bootstrap proportions for each group was computed for the bootstrap difference statistic.

For individual bootstrap samples: answers will vary.

15.4.3 (c). Generate 1000 samples to get 1000 bootstrap sample proportion differences. Describe the shape and center of this bootstrap distribution

Click for answer

Answer: The shape is symmetric around a center value of about 0.41 (the sample difference in proportions).

15.4.4 (d). Compute a 95% confidence interval for the difference between the proportion of Democrats and Republicans who believe in global warming.

Click for answer

Answer: The sample difference in proportions is $\hat{p}_{Dem} - \hat{p}_{Rep} = 0.79 - 0.38 = 0.41$, the standard error from the bootstrap distribution is 0.020 so we compute the 95% confidence interval using $0.41 \pm 2(0.020)$ giving an interval of 0.37 to 0.45.

15.4.5 (e). Interpret your interval from part (d) in context and without using the word difference!! (i.e. give a directional claim that uses words like "more" or "less")

Click for answer

Answer: We are 95% confident that the percent of Democrats who believe there is solid evidence of global warming is between 37 and 45 percentage points higher than the percent of Republicans who believe this.

15.4.6 (f). To compute this interval, we assumed that 1000 people were sampled from each subpopulation (Dems and Reps). Suppose this sample size was just 500 people for each group. Would your 95% confidence interval be wider or shorter than the one computed in part (d)? Explain.

Click for answer

Answer: With fewer people in each group, we will get a larger bootstrap SE and hence a larger margin of error for the CI. Remember that the SE of a sampling distribution gets smaller as the sample size increases, the same behavior is seen in a bootstrap distribution.

15.5 Example 5: Credit Loan Data

The data set `CreditData.csv` contains records for 1000 loans that either defaulted (BadLoan) or did not default (GoodLoan). There are 300 loans that defaulted and 700 that did not. Let's consider that the 300 loans that defaulted are random sample of loans that default and the 700 non-defaulting loans are a random sample of loans that don't default.

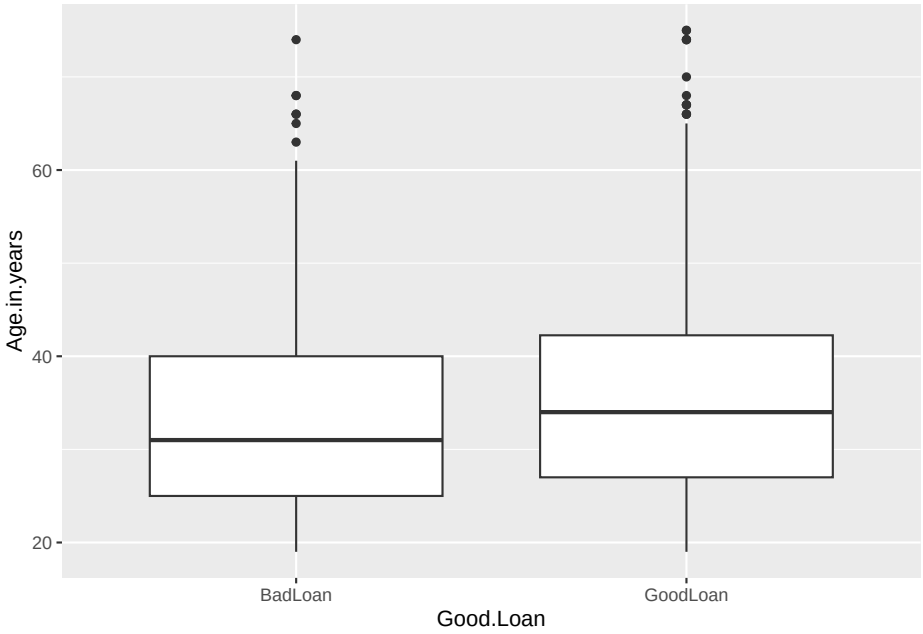
```
credit <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Creditl
table(credit$Good.Loan)
```

```
 BadLoan GoodLoan
     300      700
```

15.5.1 (a) Visualize age vs. default

The variable `Age.in.years` gives the age of the person who received the loan. Construct a side-by-side boxplot of age by `Good.Loan` and compute the sample means for each group.

```
# Boxplot using ggplot2
ggplot(credit, aes(x = Good.Loan, y = Age.in.years)) +
  geom_boxplot()
```




```
# Mean age for each Good.Loan category using dplyr
credit %>%
  group_by(Good.Loan) %>%
  summarize(mean_age = mean(Age.in.years))
```

```
# A tibble: 2 x 2
  Good.Loan mean_age
  <chr>        <dbl>
1 BadLoan       34.0
2 GoodLoan      36.2
```

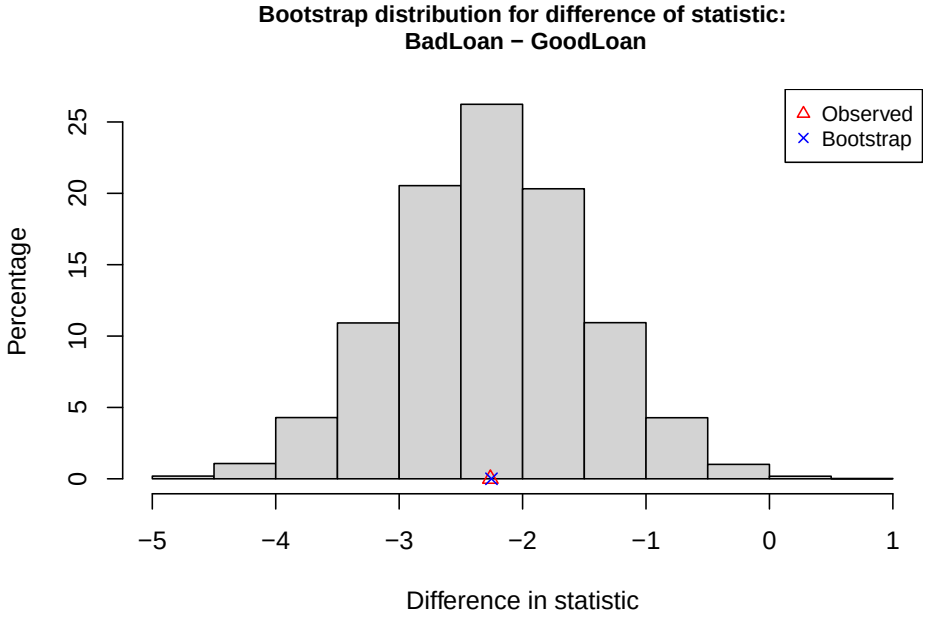
- What are the mean ages in each group?
  Click for answer
  Answer: 34.0 years for the bad loan group and 36.2 years for the good loan group.
- Describe the distribution of ages in each group. Are there any outliers that could be overly influential on the value(s) of the sample mean(s)?
  Click for answer
  Answer: Both age distributions are somewhat right skewed with a few outliers identified by the boxplot rule. But there aren't any extremely unusual cases.

15.5.1.1 (b) Bootstrap CI for a difference in means

The `boot(y ~ x, data=)` command generates 10000 bootstrap samples for the true difference in means of `y` for each of the two groups in `x`. The command is contained in the `CarletonStats` package. Here we use it to compute the bootstrap distribution for the difference in mean ages of the two default groups:

```
library(CarletonStats)
boot(Age.in.years ~ Good.Loan, data=credit)
```

```
    ** Bootstrap interval for difference of statistic

 Observed difference of statistic:  BadLoan - GoodLoan =  -2.26095
 Mean of bootstrap distribution: -2.25219
 Standard error of bootstrap distribution: 0.77504

 Bootstrap percentile interval
     2.5%      97.5%
-3.7842976 -0.7204167

        *--------------*
```



- Give the difference in sample mean ages reported by the output. Use correct notation.

Click for answer
Answer: The average age of people with a bad loan is about 2.3 years less than the average age of people with a good loan.

- Give the 95% confidence interval for the difference in mean ages using the percentile method
Click for answer
Answer: The percentile interval is -3.8 to -0.7 years.
- Compute the 95% confidence interval for the difference in mean ages using the bootstrap SE. Is it similar to the CI from the percentile method?

Click for answer

Answer: The CI using the SE is -3.8 to -0.7. The intervals are very similar.

$$-2.26095 \pm 2(0.77852) = (-3.81799, -0.70391)$$

```
-2.26095 - 2*(0.77852)
```

```
[1] -3.81799
```

```
-2.26095 + 2*(0.77852)
```

```
[1] -0.70391
```

15.5.1.2 (c) Interpret

Interpret your percentile interval in context using a directional statement. Does this interval suggest that the mean ages differ in the population of all good and bad loan holders?

Click for answer

Answer: We are 95% confident that the mean age of people who default on a loan for this population is about 0.7 to 3.8 years less than the mean age of people who do not default. This interval does support the notation that there is a difference in mean ages of these two groups in the population. It suggests that the average age of people who default is less than the average age of those who don't.

15.6 Example 6 : Credit data continued

The variable `Telephone` tells us if the individual has a phone number on their loan file. Let's look at the proportion of individuals who have a phone number for each type of loan (default or not).

15.6.0.1 (a) Data clean up

The entries in the `Telephone` column are either `none` or `yes, registered under the customers name`.

```
table(credit$Telephone)
```

```
                                    none
                                     596
yes, registered under the customers name
                                     404
```

```
# Modify the Telephone variable levels using dplyr and forcats
credit <- credit %>%
  mutate(Telephone = recode(Telephone,
                            "none" = "no",
                            "yes, registered under the customers name" = "yes"))
# Convert the Telephone variable to a factor
credit$Telephone <- as.factor(credit$Telephone)
# Display the levels of the modified Telephone variable
levels(credit$Telephone)
```

```
[1] "no"  "yes"
```

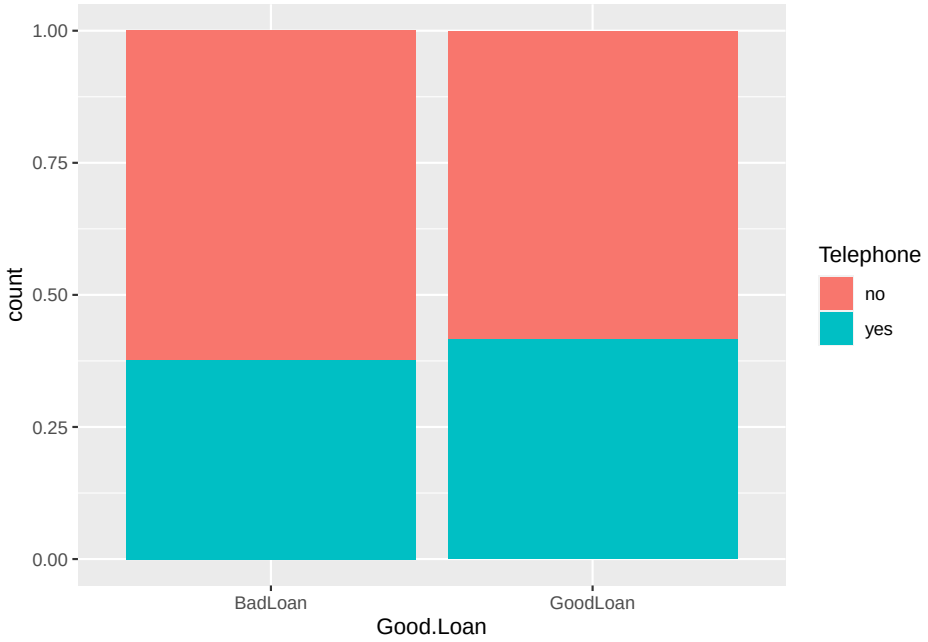
15.6.0.2 (b) Phone rate by default type

Here we get the distribution of phone numbers (yes or no) by default type (good vs bad loan):

```
prop.table(table(credit$Good.Loan, credit$Telephone),1)
```

```
                no       yes
  BadLoan  0.6233333 0.3766667
  GoodLoan 0.5842857 0.4157143
```

```
library(ggplot2)
ggplot(credit, aes(x=Good.Loan, fill=Telephone)) + geom_bar(position="fill")
```



- What proportion of bad loans have a phone number on the account?
  Click for answer
  Answer: About 37.7% of bad loans have a phone number.
- What proportion of good loans have a phone number on the account?
  Click for answer
  Answer: About 41.6% of good loans have a phone number.
- What is the sample difference in the proportion of good loans and bad loans that have a phone number? Use correct notation for this number.
  Click for answer
  Answer: Here we get $\hat{p}_{good} - \hat{p}_{bad} = 0.4157143 - 0.3766667 = 0.0390476$.

```
0.4157143 - 0.3766667
```

```
[1] 0.0390476
```

15.6.0.3 (c) Using the `boot` command with a categorical response

In order to get the bootstrap distribution for the sample difference in proportions, we need to recode the "response" variable `Telephone` to have a 1 indicating a "yes" response and 0 indicating a "no" response. This is done with an `ifelse` command:

```
credit$Telephone_binary<- ifelse(credit$Telephone == "yes", 1, 0)
head(credit[,c("Telephone", "Telephone_binary")])
```

```
  Telephone Telephone_binary
1       yes                1
2        no                0
3        no                0
4        no                0
5        no                0
6       yes                1
```

which reads "if `Telephone` equals `yes` than assign a `1`, else assign a 0". These 0's and 1's are assigned to a variable called `Telephone_binary` that is now in your data frame (checked this with the `View(credit)` command).

Check your work to make sure `Telephone_binary` records what you want it to record

```
table(credit$Telephone)
```

```
 no yes
596 404
```

```
table(credit$Telephone_binary)
```

```
  0   1
596 404
```

The mean of the 0/1 coded variable computes the proportion of "yes" responses:

```
mean(credit$Telephone_binary)
```

```
[1] 0.404
```

```
404/1000  # proportion of yes
```

```
[1] 0.404
```

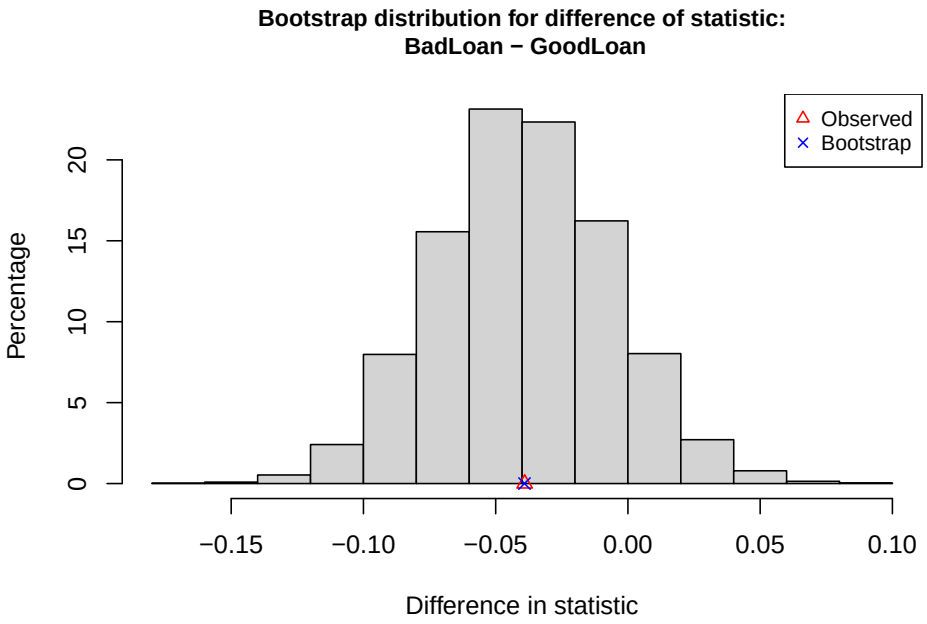
Note: All examples in your **Lab Manual** already have this 0/1 recoding done in the lab manual data sets. But I thought you might want to learn how to do this recoding in case you plan to use this command with other, non-lab manual data sets!

15.6.0.4 (d) 95% confidence interval for the difference in phone

We can now use the 0/1 version of telephone in the `boot` command (like example 1) to compute a 95% bootstrap confidence interval for the difference in the population proportion of good loans and bad loans that have a phone number.

```
boot(Telephone_binary ~ Good.Loan, data=credit)
```

```
    ** Bootstrap interval for difference of statistic

 Observed difference of statistic:  BadLoan - GoodLoan =  -0.03905
 Mean of bootstrap distribution: -0.03911
 Standard error of bootstrap distribution: 0.03303

 Bootstrap percentile interval
      2.5%       97.5%
-0.10238095  0.02666667

         *--------------*
```



Even though the language used in the output says "statistic" we are computing a difference in "proportions"!!

- Give the 95% confidence interval for the difference in the population proportion of bad loans and good loans that have a phone number using the

percentile method
Click for answer
Answer: The percentile interval for Bad − Good is -0.105 to 0.028.

- Give the 95% confidence interval for the difference in the population proportion of bad loans and good loans that have a phone number using the bootstrap SE. Is it similar to the CI from the percentile method?
Click for answer
Answer: The SE method gives an interval for Bad − Good of -0.107 to 0.028 which is very similar to the percentile interval.

```
-0.03905 - 2* 0.03373
```

```
[1] -0.10651
```

```
-0.03905 + 2* 0.
```

```
[1] -0.03905
```

15.6.0.5 (e) Interpret

Interpret your percentile interval in context using a directional statement. Does this interval suggest that there is a difference in the percentage of bad loan holders who provided a phone number comapared to the percetage of good loan holders who gave a number? Explain.

Click for answer

Answer: We are 95% confident that the percentage of good loan accounts with a phone number is anywhere from 10.7 percentage points higher than to 2.8 percentage points less than the percentage of bad loans with a phone number.

Chapter 16

Class Activity 9

16.0.1 Example 1: A Muslim president?

A survey of 1,527 American adults conducted in June 2015 stated that 60% would vote for a qualified Muslim presidential candidate. The survey goes on to say that "… the margin of sampling error is +/- 3 percentage points at the 95% confidence level."

(a). What is the relevant sample statistic? Give appropriate notation and the value of the statistic.

Click for answer

Answer: $\hat{p} = 0.60$

(b). What population parameter are we estimating with this sample statistic?

Click for answer

Answer: p = the proportion of all American adults who vote for a qualified Muslim candidate

(c). Use the margin of error to give a confidence interval for the population parameter and interpret this interval in context.

Click for answer

Answer: $0.60 \pm .03$ gives an interval from 0.57 to 0.63 I am 95% confident that the proportion of American adults who would vote for a Muslim presidential candidate is between 57% and 63%.

(d). Is it reasonable to say that a majority of American adults would vote for a qualified presidential candidate? (Is 0.50 a plausible value?)

Click for answer

Answer: The proportion of all Americans who would vote for a Muslim presidential candidate is likely between 57% and 63%, so we could say that majority (>50%) would vote for a Muslim candidate.

(e). Explain what "95% confidence" mean for this example. (Don't just repeat your answer to 1c.)

Click for answer

Answer: About 95% of all samples of 1527 American adults will give us a sample proportion who would vote for a Muslim presidential candidate that is within 3% of the population proportion who would vote for a Muslim presidential candidate.

16.0.2 Example 2: Biomass in Tropical Forests

Using a random sample of 4079 inventory plots, scientists found a sample average of 11,600 tons of carbon per square kilometer with a standard error of 1000 tons. Give a 95% confidence interval for the mean amount of carbon per square kilometer in tropical forests. Clearly interpret the meaning of this confidence interval.

Click for answer

Answer: $11,600 \pm 2(1000)$ gives an interval from 9,600 to 13,600. We are 95% sure that the mean amount of carbon per square kilometer in all tropical forests is between 9,600 and 13,600 tons.

16.0.3 Example 3: Change in gun ownership?

A 2016 study described in The Guardian found that a random sample of US adults in 1994 found a female rate of gun ownership of 9%. A similar random sample in 2015 found the rate of female gun ownership rose to 12%. Though not given in the article, let's assume that the SE for the difference in these two sample proportions is 2%.

(a). Use correct notation to describe our parameter of interest: the difference in the proportion of female gun owners in 1994 and 2015.

Click for answer

Answer: $\hat{p}_{1994} - \hat{p}_{2015}$

(b). Use the data collected to estimate your parameter in (a) and use correct notation for this statistic.

Click for answer

Answer: $\hat{p}_{1994} - \hat{p}_{2015}$ = 0.09 - 0.12 = -0.03

(c). Compute a 95% confidence interval for the parameter in a.

Click for answer

Answer: $(0.09 - 0.12) \pm (.02) = -0.03 \pm 0.04 = -0.07$ to $0.01, or -7\%$ to 1%

(d). Interpret your interval in c and explain why this support the authors claim that "the increase [in female gun ownership] was not meaningful."

Click for answer

Answer: We are 95% confident that the female gun owner rate in 1994 could be 7 percentage point lower to 1 percentage point higher than the rate in 2015. We can say that the observed increase from 1994 to 2015 of 3% is not "statistically significant" because it is within the margin of error (4%) for this study.

16.0.4 Example 4: Interpreting a Confidence Interval

Using a sample of 24 deliveries described in "Diary of a Pizza Girl" on the Slice website, we find a 95% confidence interval for the mean tip given for a pizza delivery to be \$2.18 to \$3.90. Which of the following is a correct interpretation of this interval? Indicate all that are correct interpretations.

(a). I am 95% sure that all pizza delivery tips will be between \$2.18 and \$3.90.

Click for answer

Answer: Incorrect. The interval is about the mean, not individual tips..

(b). 95% of all pizza delivery tips will be between \$2.18 and \$3.90.

Click for answer

Answer: I am 95% sure that the mean pizza delivery tip for this sample will be between \$2.18 and \$3.90.

(c). I am 95% sure that the mean tip for all pizza deliveries in this area will be between \$2.18 and \$3.90.

Click for answer

Answer: Correct!

(d). I am 95% sure that the confidence interval for the mean pizza delivery tip will be between \$2.18 and \$3.90.

Click for answer

Answer: Incorrect. The confidence is in where the population mean is, not where the interval itself is.

Chapter 17

Class Activity 10

17.0.1 Example 1: Extrasensory Perception (ESP)

In an ESP test, one person writes down one of the letters A, B, C, D, or E and tries to telepathically communicate the choice to a partner. The partner then tries to guess what letter was selected.

(a). Repeat this a couple of times, then switch roles with your partner. How often did you guess correctly?

Click for answer

Answer: Answers will vary!

(b). If there is no ESP and people are just randomly guessing from among the five choices, what proportion of guesses would we expect to be correct? If no ESP, we expect p = ...

Click for answer

Answer: $p = 0.2$ (since there are five choices and they are randomly guessing)

(c). Which sample proportion correct would provide the greatest evidence that people have ESP: (If we assume the sample size is the same in every case.)

Click for answer

Answer: $\hat{p} = 3/4$ since this means more correct.

(d). Write down the null and alternative hypotheses for testing whether people have ESP:

Click for answer

Answer:

$$H_0 : p = 0.2$$

$$H_a : p > 0.2$$

where p is the proportion correct for all people's guesses. Since we are looking for evidence that the proportion is significantly above 0.2 (random guesses), the alternate hypothesis is larger than.

17.0.2 Example 2

In an experiment, students were given words to memorize, then were randomly assigned to either take a 90 minute nap or take a caffeine pill. A couple hours later, they were tested on their recall ability. We wish to test to see if the sample provides evidence that there is a difference in mean number of words people can recall depending on whether they take a nap or have some caffeine.

(a). What is the explanatory variable? Is it categorical or quantitative?

Click for answer

Answer: Explanatory = nap or caffeine (categorical)

(b). What is the response variable? Is it categorical or quantitative?

Click for answer

Answer: Response = number of words recalled (quantitative)

(c). What is the parameter of interest for this experiment? Use correct notation.

Click for answer

Answer: Quantitative = mean responses, where μ_1 and μ_2 are the mean words recalled in the two different conditions

(d). What are the null and alternative hypotheses for this test? Use correct notation.

Click for answer

Answer:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

The alternate hypothesis is not equals to since we are looking for evidence that the means are different (We do not know which one is larger!)

17.0.3 Example 1 Revisited:

(a). If the results of a test for ESP are statistically significant, what does that mean in terms of ESP?

Click for answer

Answer:

It means we can conclude that $p > 0.2$ and that the sample results were so strong that we can conclude that ESP does exist and get more right than would be expected by random chance.

(b). If the results are not statistically significant, what does that mean in terms of ESP?

Click for answer

Answer:

The sample results are inconclusive. People may or may not have ESP. Sample results could be just random chance.

Chapter 18

Class Activity 11

Midterm !!

Chapter 19

Class Activity 12

19.1 Example 1: Sleep or Caffeine for Memory

In an experiment, 24 students were given words to memorize, then were randomly assigned to take a 90 minute nap or take a caffeine pill (12 in each group). They were then tested on their recall ability. We test to see if the sample provides evidence that there is a difference in mean number of words people can recall depending on whether they take a nap or have some caffeine. The hypotheses are:

$$H_0 : \mu_S - \mu_C = 0 \quad H_A : \mu_S - \mu_C \neq 0$$

The sample mean difference is $\bar{x}_S - \bar{x}_C = 3$. We want to know if this difference in sample means is statistically significant.

19.1.0.1 (a) Explain how to generate a randomization distribution for $\bar{x}_S - \bar{x}_C$ that is consistent with $H_0 : \mu_S - \mu_C = 0$.

Click for answer

Answer: We could randomly reassign the treatment to the study participants since, under the null, their recall abilities would be the same under either treatment. For each reassignment, we recomputed the sample mean difference and plot it in the dotplot shown below

19.1.0.2 (b) Navigate to the Statkey website.

Select the **Test for Difference in Means** option under **Randomization Hypothesis Tests**. Change the data set from **Leniency and Smiles** to **Sleep**

Caffeine Words. Note that the original sample data has a sample mean difference of 3 words.

- **Generate 1 Sample** from this null randomization distribution. What is the difference in the average word recall of the two groups in this sample? Repeat this a couple of times.
- **Generate 1000 Samples** a few of times (get at least 3000 resamples). How unusual is getting a difference in means of 3 or more words?

19.1.0.3 (c) Compute the randomization p-value

Select the **Two-Tail** button at the top of the plot. Change the positive x-axis value to the observed difference of 3.0. The p-value is 2 times the proportion of resamples that have a difference of 3 or above. What is the p-value?

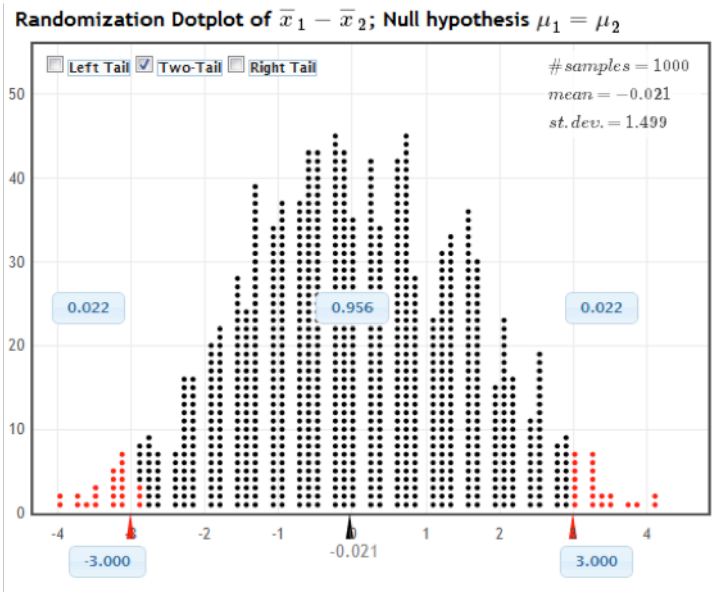


Figure 19.1: Example 1

Click for answer

Answer: We see in the image that the proportion in the tail beyond the sample statistic of 3.0 is 0.022. Because this is a two-tail test, we have to account for both tails, so the p-value is $2(0.022) = 0.044$.

19.1.0.4 (d) Interpret + Conclusion

Interpret the p-value. Does the p-value support the alternative hypothesis (do you think difference of means of 3 is statistically significant) or is it inconclusive? Explain.

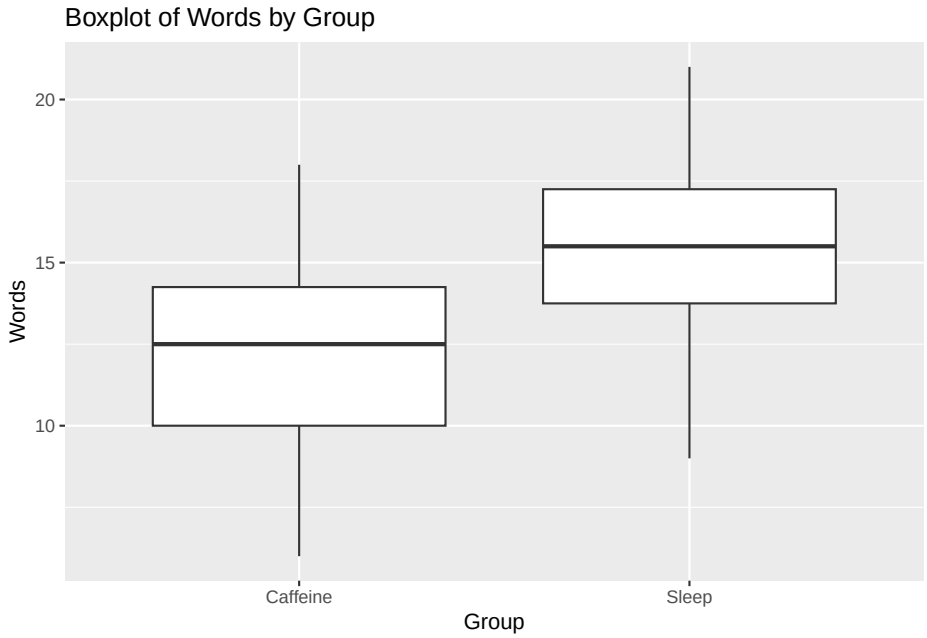
Click for answer

Answer: We would see a difference of at least 3 words recalled, on average, in about 4.4% of all possible samples if the influence of sleep and caffeine on recall was the same The results show some evidence of statistical significance, meaning that the caffeine and sleep may have some difference effects on word recall ability.

19.1.0.5 (e) Redo in Rstudio

First get the data from the Lock website and check important summary stats:

```
library(readr)
wordData <- read_csv("http://math.carleton.edu/Stats215/Textbook/SleepCaffeine.csv")
# Create a boxplot using ggplot2
ggplot(wordData, aes(x = Group, y = Words)) +
  geom_boxplot() +
  labs(title = "Boxplot of Words by Group")
```

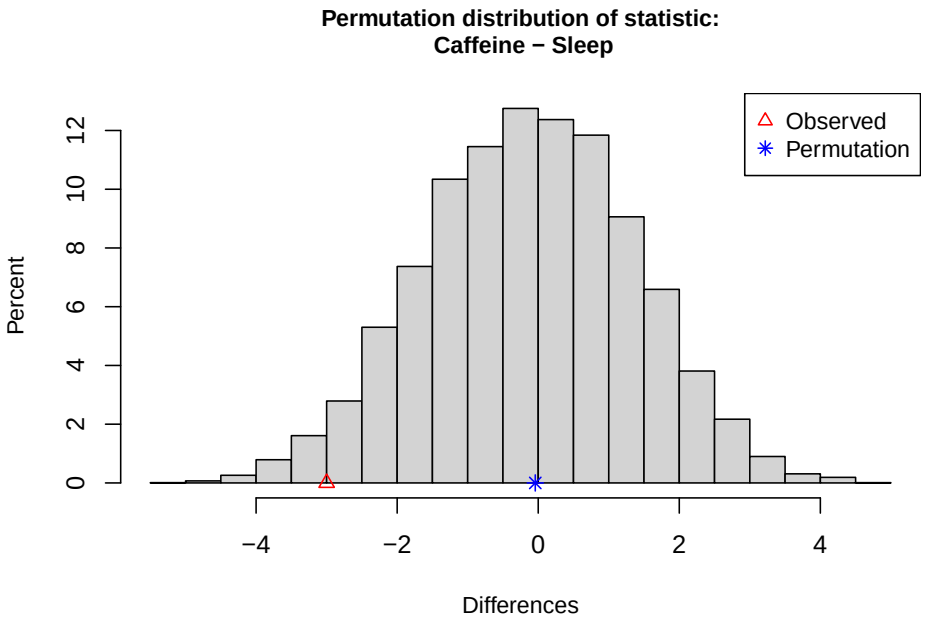


```
# Summary statistics using dplyr
wordData %>%
  group_by(Group) %>%
  summarise_all(list(min, q1 = ~quantile(., 0.25), median, mean, q3 = ~quantile(., 0.75
```

```
# A tibble: 2 x 7
  Group      fn1    q1   fn2   fn3    q3   fn4
  <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Caffeine     6  10    12.5  12.2  14.2    18
2 Sleep        9  13.8  15.5  15.2  17.2    21
```

Then load the `CarletonStats` package and run the `permTest(y ~ x, data=)` command where `y` is your quantitative (or 0/1 coded) response and `x` defines the two groups you are comparing.

```
library(CarletonStats)
permTest(Words ~ Group, data=wordData)
```




```
   ** Permutation test **

 Permutation test with alternative: two.sided
 Observed statistic
  Caffeine :  12.25      Sleep :  15.25
```

```
Observed difference: -3

Mean of permutation distribution: -0.04195
Standard error of permutation distribution: 1.48239
P-value:  0.055

   *-------------*
```

- Why is the observed difference reported as -3?

Click for answer

Answer: The difference is computed alphabetically: Caffeine minus Sleep so the difference in now -3 instead of +3.

- What is the p-value? Is it the same as the Statkey p-value? The same as your neighbors p-value? Why not?

Click for answer

Answer: The p-value is around 5%. Any difference between Statkey, neighbors or different runs of the `permTest` command stem from the fact that different resamples are obtained each time a randomization distribution is generated. There may be some small (inconsequential) difference in p-values due to this.

19.2 Example 2: Resident vs Non-resident Tuition

The lab manual data set `Tuition2006` is a random sample of state colleges and universities in the U.S. We want to know if the average tuition charged to non-residents is higher than residents for all state colleges and universities:

$$H_0 : \mu_{Non-res} - \mu_{Res} = 0 \quad H_A : \mu_{Non-res} - \mu_{Res} > 0$$

19.2.0.1 (a) Paired Data

Read in the data. Note that each case (school) has a response value for the resident and non-resident tuition variables. This makes this a paired data example. Contrast this with the word recall example in which each case (student) only had one response (word recall) and treatment (caffeine/sleep).

```
library(readr)
tuition <- read_csv("http://math.carleton.edu/Stats215/RLabManual/Tuition2006.csv")
head(tuition)
```

```
# A tibble: 6 x 5
   ...1 Institution          Res NonRes  Diff
  <dbl> <chr>              <dbl>  <dbl> <dbl>
1     1 Univ of Akron (OH)  4200   8800 -4600
2     2 Athens State (AL)   1900   3600 -1700
3     3 Ball State (IN)     3400   8600 -5200
4     4 Bloomsburg U (PA)   3200   7000 -3800
5     5 UC Irvine (CA)      3400  12700 -9300
6     6 Central State (OH)  2600   5700 -3100
```

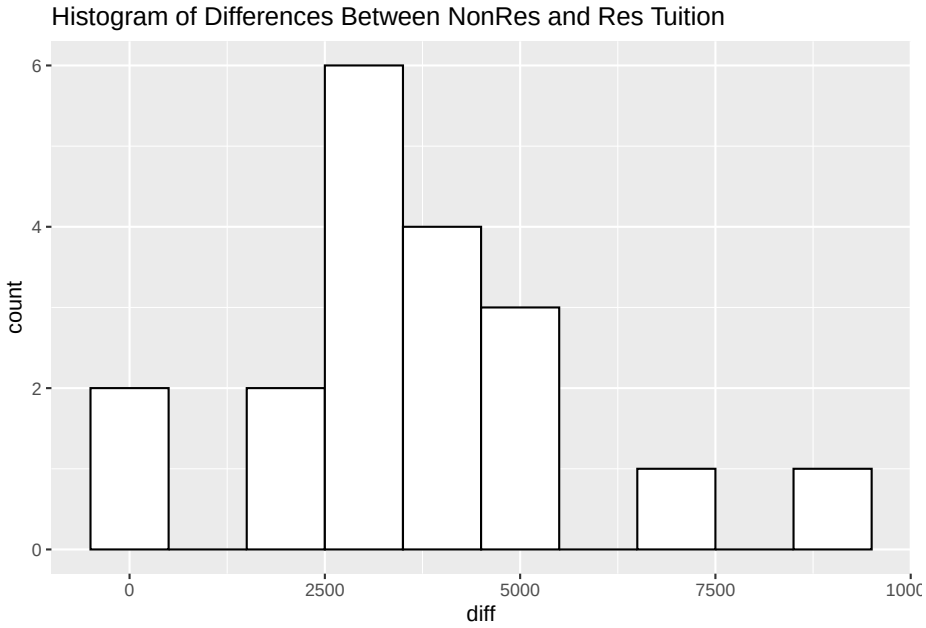
19.2.0.2 (b) Permutation test for paired data

Let's compute the difference of non-resident and resident tuitions (NR minus R):

```
diff <- tuition$NonRes - tuition$Res
summary(diff)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    200    2650    3100    3584    4500    9300
```

```
# Histogram of differences using ggplot2
tuition %>%
  mutate(diff = NonRes - Res) %>%
  ggplot(aes(x = diff)) +
  geom_histogram(binwidth = 1000, color = "black", fill = "white") +
  labs(title = "Histogram of Differences Between NonRes and Res Tuition")
```



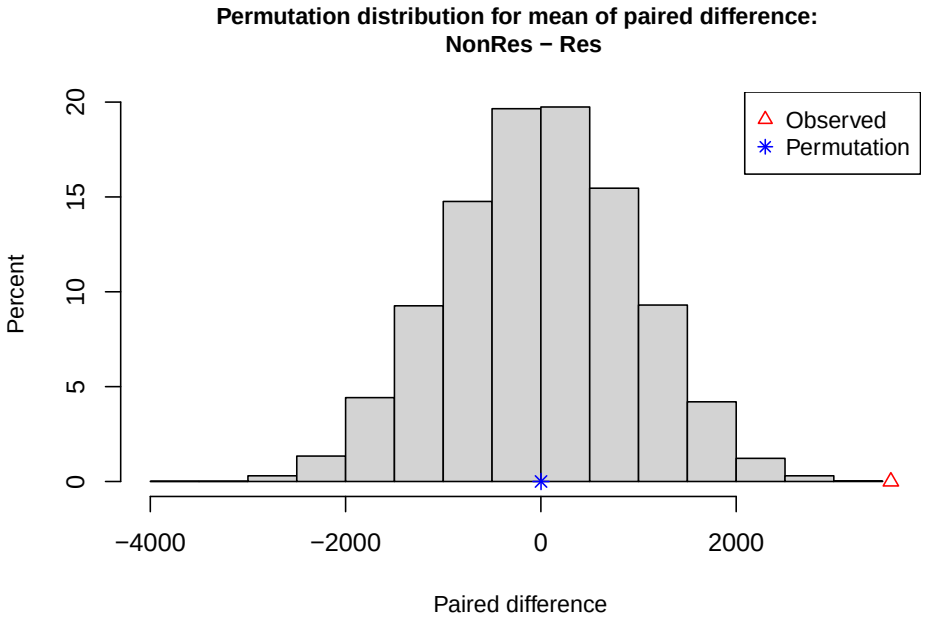
- What is the average difference in tuition costs?

Click for answer

Answer: The observed mean difference is $3584

- Is this observed mean difference statistically significant? To test use the command `permTestPaired`:

```
permTestPaired(NonRes ~ Res,data = tuition, alt = "greater")
```




```
   ** Permutation test for mean of paired difference **

Permutation test with alternative: greater
Observed mean
 NonRes :  6405.263     Res :  2821.053
Observed difference  NonRes - Res : 3584.211

Mean of permutation distribution: 1.37961
Standard error of permutation distribution: 951.5183
P-value:  1e-04

   *-------------*
```

The `alt` of `greater` was used because the function `permTestPaired(A ~ B)` computes paired differences as "A" minus "B".

- What is the p-value for this test?

Click for answer

Answer: Less than 0.0001

- Is this observed mean difference statistically significant?

Click for answer

Answer: Yes, an observed mean difference of at least \$3584 would rarely occur just by chance which provides us strong evidence that the mean tuition amount of non-residents is higher than residents in the population of state colleges and universities (in 2006).

Chapter 20

Class Activity 13

20.1 Example 1: Effect of Exercise on Heart Rate

In a study, 30 participants were randomly assigned to engage in either aerobic exercise or resistance training (15 in each group). Their resting heart rate was measured before and after a 6-week exercise program. We want to test if there is a difference in the mean decrease in resting heart rate between aerobic exercise and resistance training. The hypotheses are:

$$H_0 : \mu_A - \mu_R = 0 \qquad H_A : \mu_A - \mu_R \neq 0$$

The sample mean difference is $\bar{x}_A - \bar{x}_R = 6.2$. We want to know if this difference in sample means is statistically significant.

20.1.1 (a) Randomization Distribution

Describe how you could generate a randomization distribution for $\bar{x}_A - \bar{x}_R$ that is consistent with $H_0 : \mu_A - \mu_R = 0$.

Click for answer

Answer: To generate a randomization distribution for the sample mean difference, we would randomly reassign the treatment (aerobic exercise or resistance training) to the study participants. Under the null hypothesis, the mean decrease in resting heart rate would be the same under either treatment. For each reassignment, we would compute the sample mean difference and plot it in the dotplot.

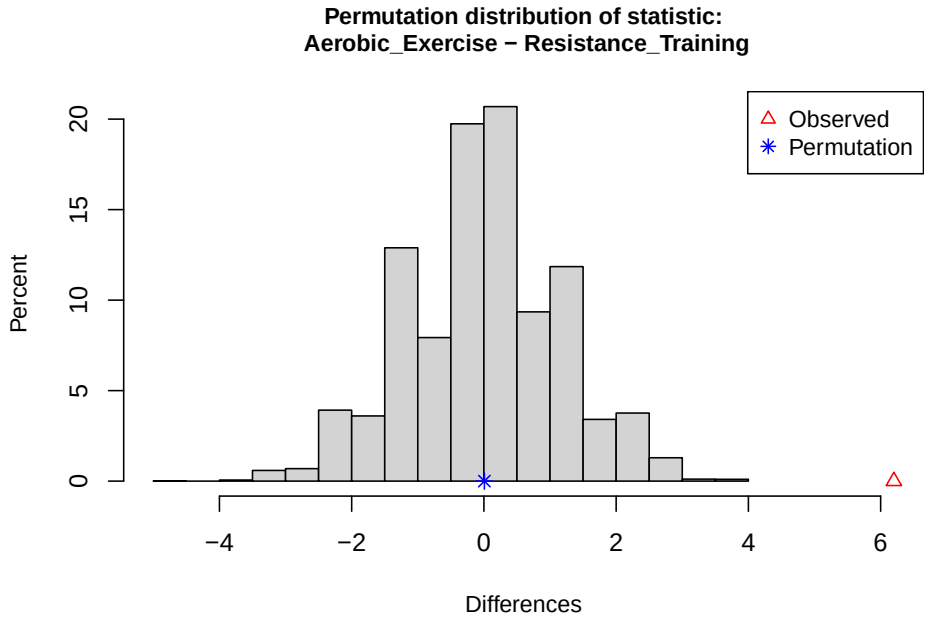
20.1.2 (b) Calculating the p-value

Using a statistical software, generate a randomization distribution for the difference in means and calculate the p-value for the two-tailed test. Is the observed difference of 6.2 bpm statistically significant?

Click for answer

Answer: We can use the `permTest` function from the `CarletonStats` package to generate a randomization distribution for the difference in means and calculate the p-value for the two-tailed test. We can compare the p-value of $2e^{-4}$ to a chosen significance level (e.g., 0.05) to determine if the observed difference of 6.2 bpm is statistically significant. Since the p-value is less than the significance level, the observed difference is statistically significant.

```
library(CarletonStats)
library(readr)
exercise <- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/exer
permTest(Decrease_in_Resting_Heart_Rate~Group, data= exercise)
```




```
   ** Permutation test **

Permutation test with alternative: two.sided
Observed statistic
 Aerobic_Exercise :  9.2    Resistance_Training :  3
```

```
Observed difference: 6.2

Mean of permutation distribution: 0.00713
Standard error of permutation distribution: 1.16749
P-value:  2e-04

   *-------------*
```

20.1.3 (c) Interpretation and Conclusion

Interpret the p-value and state your conclusion regarding the difference in mean decrease in resting heart rate between aerobic exercise and resistance training.

Click for answer

Answer: Since the p-value is less than the significance level, it means that the observed difference of 6.2 bpm is highly unlikely to occur just by random chance under the null hypothesis. In this case, we would reject the null hypothesis and conclude that there is a statistically significant difference in the mean decrease in resting heart rate between aerobic exercise and resistance training.

20.2 Example 2: Job Interview Success

A study investigated the success rate of job applicants who used a career coaching service (CCS) compared to those who didn't (NCCS). Out of 120 applicants, 60 used the CCS and 60 did not. We want to test if there is a difference in the proportion of successful applicants between CCS and NCCS groups. out of the 60 applicants who used the career coaching service (CCS), 42 were successful and 18 were unsuccessful. Out of the 60 applicants who did not use the career coaching service (NCCS), 30 were successful and 30 were unsuccessful. The hypotheses are:

$$H_0 : p_{CCS} - p_{NCCS} = 0 H_A : p_{CCS} - p_{NCCS} \neq 0$$

The sample proportion difference is $\hat{p}CCS - \hat{p}NCCS = 0.20$. We want to know if this difference in sample proportions is statistically significant.

(a) Randomization Distribution Describe how you could generate a randomization distribution for $\hat{p}_{CCS} - \hat{p}_{NCCS}$ that is consistent with $H_0 : p_{CCS} - p_{NCCS} = 0$.

Click for answer

Answer: To generate a randomization distribution for the sample proportion difference, we would randomly reassign the group (CCS or NCCS) to the job

applicants. Under the null hypothesis, the proportion of successful applicants would be the same for both groups. For each reassignment, we would compute the sample proportion difference and plot it in the dotplot.

(b) Calculating the p-value

Using a statistical software, generate a randomization distribution for the difference in proportions and calculate the p-value for the two-tailed test. Is the observed difference of 0.20 statistically significant?

Click for answer

Answer: Using *Statkey* to generate a randomization distribution for the difference in proportions and observing the p-value for the two-tailed test, we can compare the p-value to a chosen significance level (e.g., 0.05) to determine if the observed difference of 0.20 is statistically significant. If the p-value is less than the significance level, the observed difference is statistically significant; otherwise, it is not. The p-value based on this randomization distribution under null hypothesis is $2 \times 0.018 = 0.036$. So, the observed difference of 0.020 is statistically significant.

(c) Interpretation and Conclusion Interpret the p-value and state your conclusion regarding the difference in the proportion of successful job applicants between CCS and NCCS groups.

Click for answer

Answer: Since the p-value is less than the significance level, it means that the observed difference of 0.20 would occur with low chance under the null hypothesis. In this case, we would reject the null hypothesis and conclude that there is a statistically significant difference in the proportion of successful job applicants between CCS and NCCS groups.

20.3 Example 3: New Teaching Method Effectiveness

A school is testing a new teaching method for math. They randomly assign 40 students to either the new method (NM) or the traditional method (TM), 20 in each group. After 3 months, the students take a standardized math test. We want to test if there is a difference in the mean test scores between NM and TM groups. The hypotheses are:

$$H_0 : \mu_{NM} - \mu_{TM} = 0 \mathsf{H}_A : \mu_{NM} - \mu_{TM} \neq 0$$

The sample mean difference is $\bar{x}_{NM} - \bar{x}_{TM} = 8.9$. We want to know if this difference in sample means is statistically significant.

20.3.1 (a) Randomization Distribution

Describe how you could generate a randomization distribution for $\bar{x}_{NM} - \bar{x}_{TM}$ that is consistent with $H_0 : \mu_{NM} - \mu_{TM} = 0$.

Click for answer

Answer: To generate a randomization distribution for the sample proportion difference, we would randomly reassign the teaching method (new or traditional) to the classes. Under the null hypothesis, the proportion of students passing would be the same under either teaching method. For each reassignment, we would compute the sample proportion difference and plot it in the dotplot.

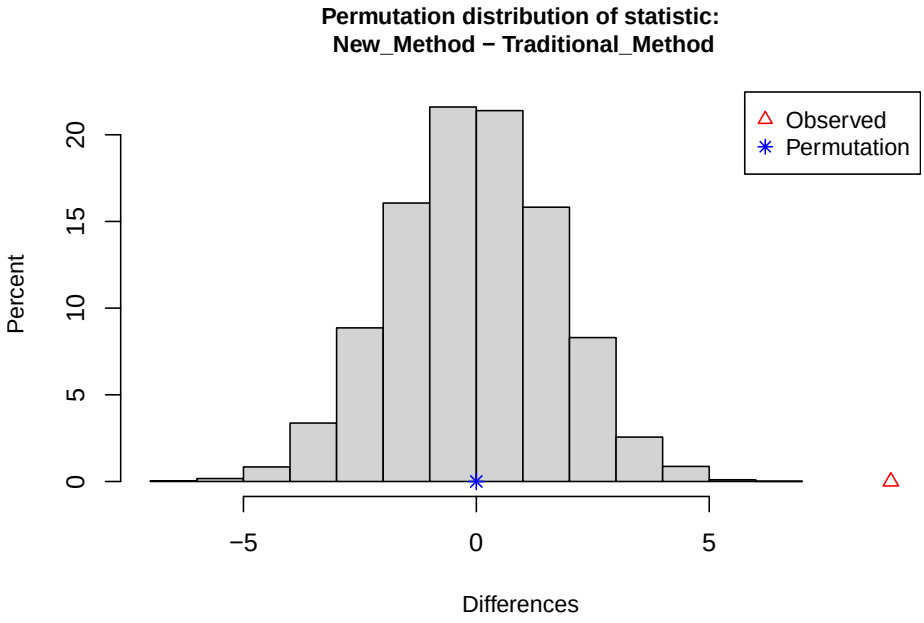
20.3.1.1 (b) Calculating the p-value

Using a statistical software, generate a randomization distribution for the difference in means and calculate the p-value for the two-tailed test. Is the observed difference of 8.9 points statistically significant?

Click for answer

Answer: The p-value is the proportion of resamples that have a difference of 8.9 or above. Depending on the generated randomization distribution, the p-value is 0.0002. This means is interpreted in terms of how likely it is to observe a difference of 8.9 or greater under the null hypothesis, which is 0.02%.

```
teaching <- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/teaching_method
permTest(Math_Test_Score~Group, data= teaching)
```




```
   ** Permutation test **

Permutation test with alternative: two.sided
Observed statistic
 New_Method :  91.05    Traditional_Method :  82.15
Observed difference: 8.9

Mean of permutation distribution: -0.00075
Standard error of permutation distribution: 1.72562
P-value:  2e-04

   *-------------*
```

20.3.1.2 (c) Interpretation and Conclusion

Interpret the p-value and state your conclusion regarding the difference in the mean test scores between the new teaching method and the traditional teaching method groups.

Click for answer

Answer: Since the p-value is less than the chosen significance level (e.g., 0.05), the results are statistically significant, and we would reject the null hypothesis in favor of the alternative. This would indicate that there is evidence suggesting the new teaching method is more effective than the traditional method.

20.4 Example 4: Type I and Type II Error Rates

Consider the previous example of a new teaching method compared to a traditional teaching method. Suppose that the school administration wants to minimize the chances of making Type I and Type II errors when deciding whether to adopt the new teaching method.

20.4.1 (a) Type I Error

Explain what a Type I error is in the context of this example, and describe the consequences of making such an error.

Click for answer

Answer: A Type I error occurs when we reject the null hypothesis when it's actually true. In this context, it means that we conclude the new teaching method is more effective than the traditional method when, in reality, there is no difference. The consequences of making a Type I error could include investing time and resources into a new teaching method that isn't actually more effective, leading to inefficient allocation of resources.

20.4.2 (b) Type II Error

Explain what a Type II error is in the context of this example, and describe the consequences of making such an error.

Click for answer

Answer: A Type II error occurs when we fail to reject the null hypothesis when it's actually false. In this context, it means that we conclude that there is no difference between the new teaching method and the traditional method when, in reality, the new method is more effective. The consequences of making a Type II error could include missing out on the opportunity to improve students' learning outcomes by not adopting the more effective teaching method.

20.4.3 (c) Adjusting the Significance Level

If the school administration believes that making a Type I error is much worse than making a Type II error, what adjustments could be made to the significance level to account for this? Explain your reasoning.

Click for answer

Answer: To decrease the chance of making a Type I error, the school administration could choose a smaller significance level, such as 0.01 instead of the typical 0.05. By using a smaller significance level, we require stronger evidence

(smaller p-value) to reject the null hypothesis, thus reducing the probability of making a Type I error.

20.4.4 (d) Factors Influencing Type II Error Rate

List the factors that influence the probability of making a Type II error and briefly describe how they affect the error rate in this context.

Click for answer

Answer: Factors that influence the probability of making a Type II error include:

Effect size: The true difference between the new teaching method and the traditional method. A larger effect size makes it easier to detect a difference, reducing the Type II error rate.

Sample size: The number of students involved in the study. A larger sample size increases the power of the test, making it more likely to detect a true difference and reducing the Type II error rate.

Variability: The amount of variation in the students' learning outcomes. Higher variability makes it more difficult to detect a true difference, increasing the Type II error rate.

Significance level: The chosen significance level (alpha) also affects the Type II error rate. A larger significance level (e.g., 0.10) decreases the Type II error rate but increases the Type I error rate.

20.4.5 (e) Balancing Error Rates

Discuss how the school administration could balance the Type I and Type II error rates when evaluating the effectiveness of the new teaching method. What factors should they consider?

Click for answer

Answer: Balancing the Type I and Type II error rates involves considering the consequences of each type of error and the desired level of confidence in the results. The administration should weigh the risks and benefits of adopting a new teaching method versus maintaining the traditional method. They should also consider factors such as the cost and feasibility of implementing the new method, as well as the potential impact on student learning outcomes. Ultimately, the administration should choose a significance level and sample size that balance the risks associated with Type I and Type II errors while taking into account practical constraints and priorities.

Chapter 21

Class Activity 14

21.1 Example 1: Gender stereotypes in children - study 4

The data for this example comes from study 4 described in this *Science* article: http://science.sciencemag.org/content/355/6323/389. This study involved asking children their interest level in a game that researcher described as for "children who are really, really smart." The higher the value of the variable `interest`, the more interested a child was in playing that game.

```
study4 <- read.csv("http://math.carleton.edu/kstclair/data/Stereo4.csv")
head(study4)
```

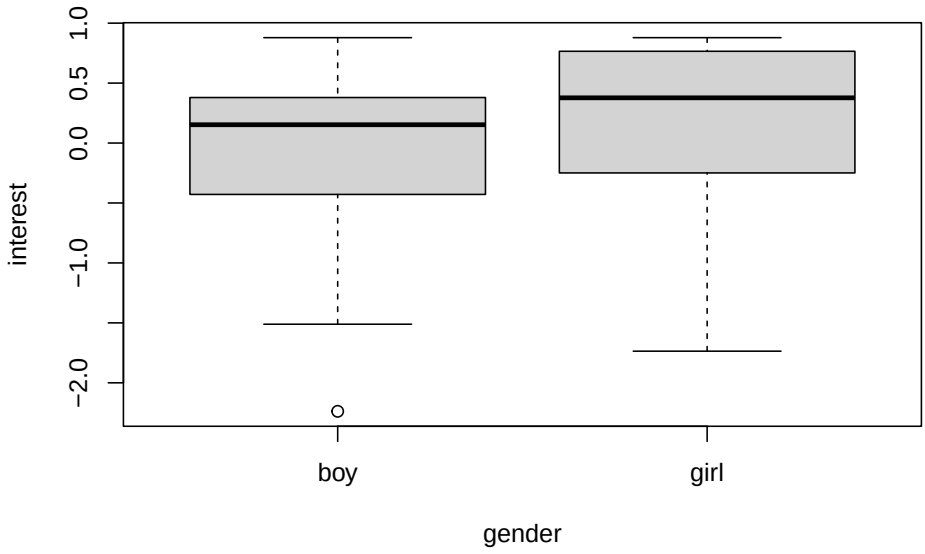
```
    study subj gender   age    interest race     race2
1 Study 4   65   girl age 6  0.37953534    5     white
2 Study 4   66   girl age 6 -0.78071539    5     white
3 Study 4   67   girl age 6 -0.47631654    5     white
4 Study 4   68   girl age 6 -0.07234632    5     white
5 Study 4   69    boy age 6 -0.70319450    6 non-white
6 Study 4   70   girl age 6  0.52467564    5     white
  eduave income        ses        age2
1     16  90000 -0.1543908 age 6 and 7
2     16 125000  0.2298424 age 6 and 7
3     18  25000 -0.3446883 age 6 and 7
4     17 125000  0.4914816 age 6 and 7
5     19 125000  1.0147600 age 6 and 7
6     12  65000 -1.4753998 age 6 and 7
```

21.1.0.1 (a) Interest in 5 year olds - test

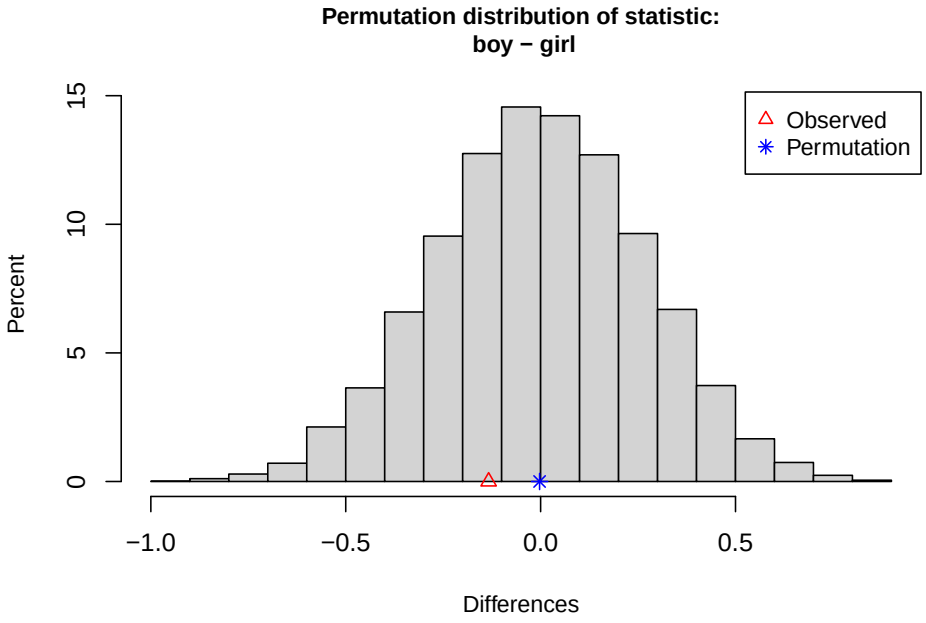
Recall the comparison of mean interest level in 5 year old boys and girls. Generate the randomization distribution for this test:

$$H_0 : \mu_{B5} - \mu_{G5} = 0 \quad H_0 : \mu_{B5} - \mu_{G5} \neq 0$$

```
library(dplyr)
study4age5 <- filter(study4, age2 == "age 5")
boxplot(interest ~ gender, data=study4age5)
```




```
library(CarletonStats)
permTest(interest ~ gender, data = study4age5)
```




```
   ** Permutation test **

Permutation test with alternative: two.sided
Observed statistic
 boy :  -0.10435    girl :  0.02906
Observed difference: -0.13341

Mean of permutation distribution: -0.00266
Standard error of permutation distribution: 0.26542
P-value:  0.622

   *-------------*
```

- What is the SE of this randomization distribution?

Click for answer

Answer: SE is about 0.26.

- What is the z-score for the observed difference in means using this distribution? Interpret the value.

Click for answer

Answer: The distribution has a center of 0 and SE of 0.26. The z-score is

$$z = \frac{-0.13341 - 0}{0.26051} = -0.51$$

This means the observed difference of -0.133 is about 0.51 SEs below the hypothesized difference of 0.

```
-0.13341/0.26051
```

```
[1] -0.5121109
```

- How large or small would the observed difference in sample means need to be to reject the null hypothesis using a 5% significance level.

Click for answer

Answer: Since the distribution is bell-shaped, we can use the fact that about 5% of sample differences are further than 2 SE's above/below the center difference of 0. Any sample difference this extreme will lead to a two-sided p-value that is less than the significance level of 5%. For this data, 2 SE's is a sample difference of 0.521 so any observed difference that is more extreme than 0.521 would lead to rejecting the null hypothesis of no difference.

```
2*0.26051
```

```
[1] 0.52102
```

21.1.0.2 (b) Interest in 5 year olds - CI

Consider the 95% (bootstrap) CI for the true difference in mean interest $\mu_{B5} - \mu_{G5}$.

- Will this interval contain the difference of 0?

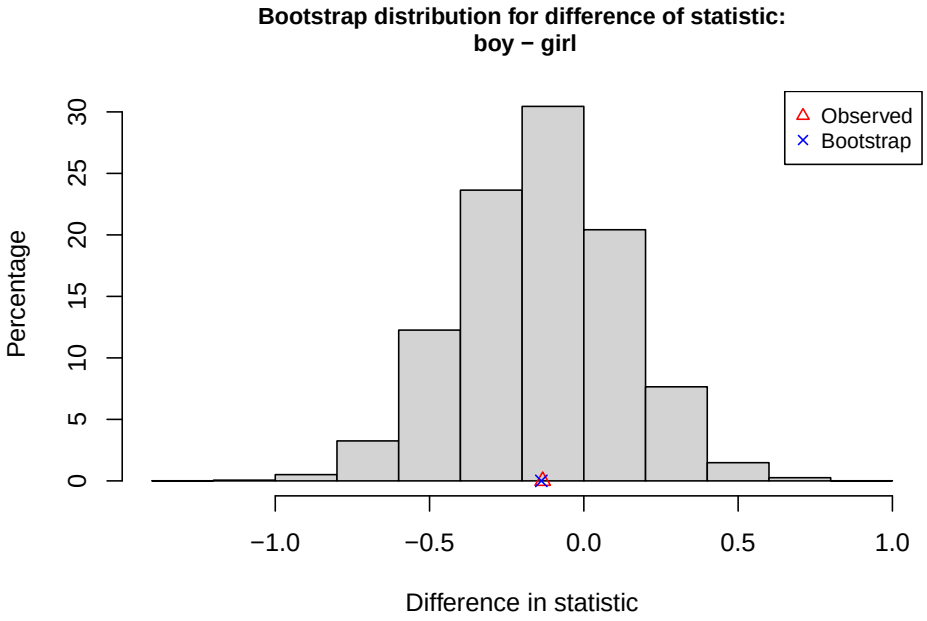
Click for answer

Answer: Yes, since we didn't reject the null difference of 0 using a 5% significance level (p-value = 0.617).

- Compute the bootstrap distribution. Does the CI capture 0?

```
set.seed(7)
boot(interest ~ gender, data = study4age5)
```

```
   ** Bootstrap interval for difference of statistic

 Observed difference of statistic:  boy - girl =  -0.13341
 Mean of bootstrap distribution: -0.13776
 Standard error of bootstrap distribution: 0.25939

 Bootstrap percentile interval
     2.5%      97.5%
-0.6459180  0.3652884

        *--------------*
```



Click for answer

Answer: Yes, the CI captures the difference of 0.

- What is the bootstrap SE? Is it similar to the randomization distribution SE?

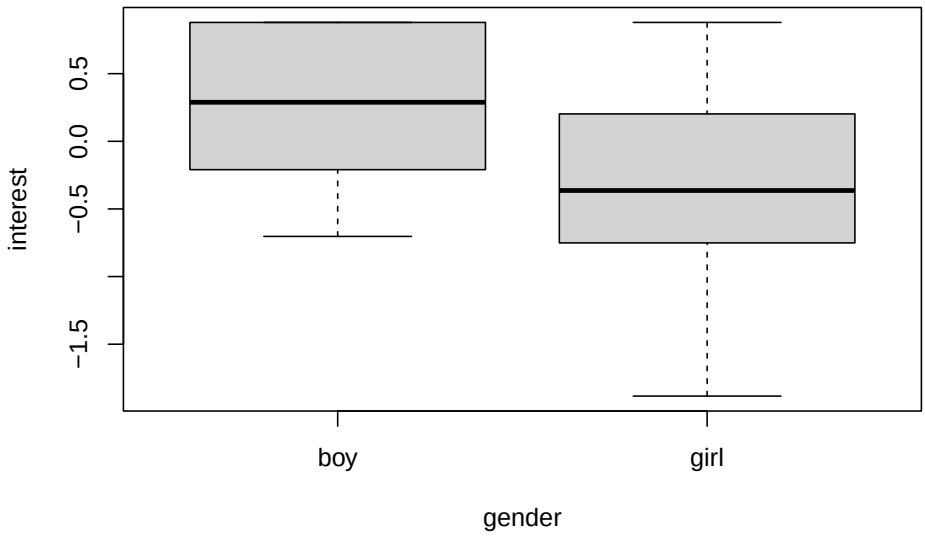
Click for answer

Answer: SE is about 0.26, which is very similar to the randomization distribution SE.

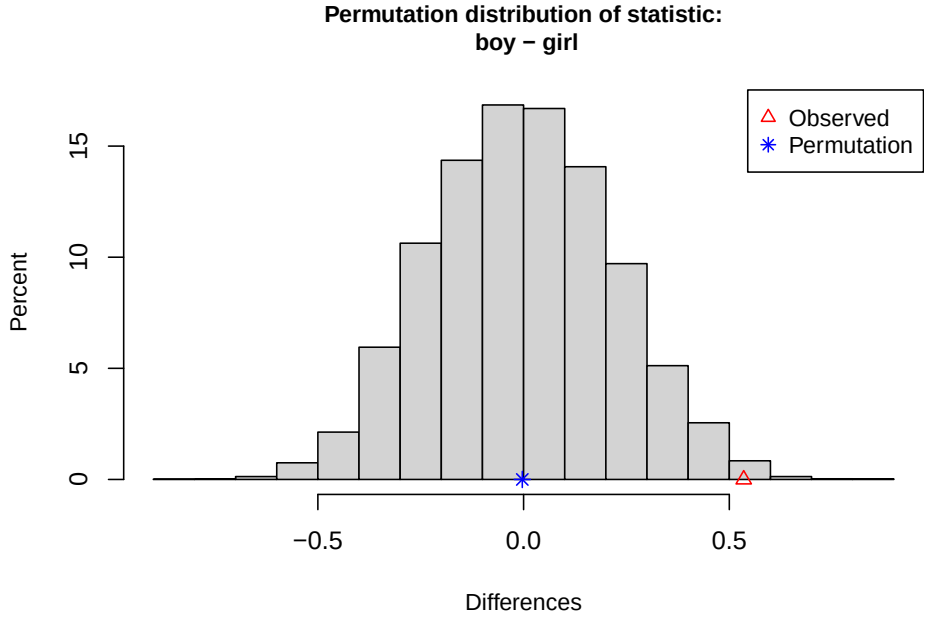
21.1.0.3 (c) Interest in 6 and 7 year olds - test

Redo part (a) for the age group `age 6 and 7`.

```
study4age67 <- filter(study4, age2 == "age 6 and 7")
boxplot(interest ~ gender, data=study4age67)
```




```
permTest(interest ~ gender, data = study4age67)
```



```
   ** Permutation test **

Permutation test with alternative: two.sided
Observed statistic
 boy :  0.21635     girl :  -0.31869
Observed difference: 0.53505

Mean of permutation distribution: -0.00312
Standard error of permutation distribution: 0.22035
P-value:  0.0114

   *-------------*
```

- What is the SE of this randomization distribution?

Click for answer

Answer: SE is about 0.225.

- What is the z-score for the observed difference in means using this distribution? Interpret the value.

Click for answer

Answer: The distribution has a center of 0 and SE of 0.225. The z-score is

$$z = \frac{0.53505 - 0}{0.22539} = 2.37$$

This means the observed difference of 0.535 is about 2.37 SEs above the hypothesized difference of 0.

```
0.53505/0.22539
```

```
[1] 2.373885
```

- How large or small would the observed difference in sample means need to be to reject the null hypothesis using a 5% significance level.

Click for answer

Answer: Since the distribution is bell-shaped, we can use the fact that about 5% of sample differences are further than 2 SE's above/below the center difference of 0. Any sample difference this extreme will lead to a two-sided p-value that is less than the significance level of 5%. For this data, 2 SE's is a sample difference of 0.451 so any observed difference that is more extreme than 0.451 would lead to rejecting the null hypothesis of no difference.

```
2*0.22539
```

```
[1] 0.45078
```

21.1.0.4 (d) Interest in 6 and 7 year olds - CI

Redo part (b) for 6 and 7 year olds.

- Will this interval contain the difference of 0?

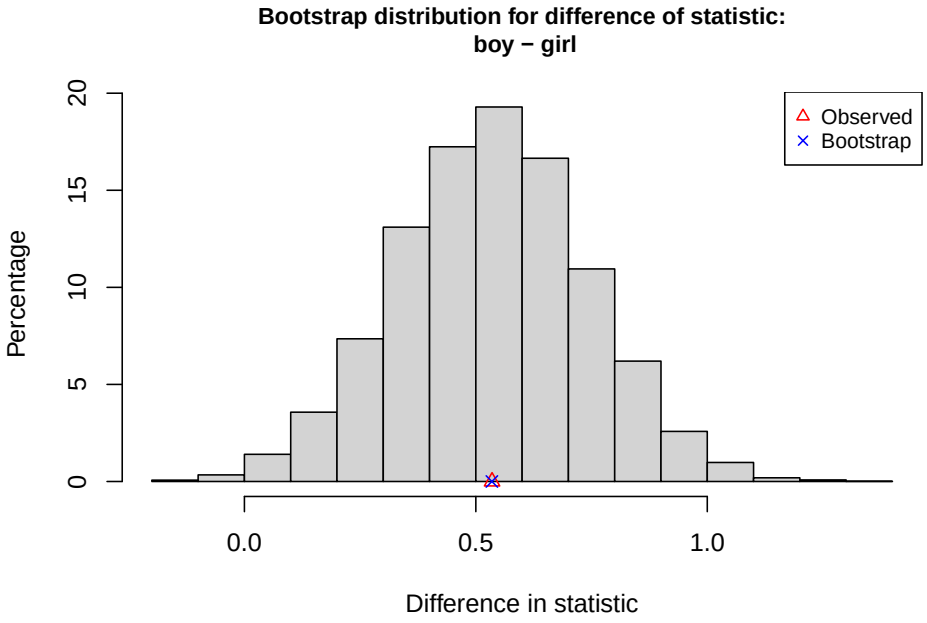
Click for answer

Answer: No, since we rejected the null difference of 0 using a 5% significance level (p-value = 0.015).

- Compute the bootstrap distribution. Does the CI capture 0?

```
boot(interest ~ gender, data = study4age67)
```

```
    ** Bootstrap interval for difference of statistic

 Observed difference of statistic:  boy - girl =  0.53505
 Mean of bootstrap distribution: 0.53468
 Standard error of bootstrap distribution: 0.20659

 Bootstrap percentile interval
     2.5%     97.5%
0.1259895 0.9348764

        *--------------*
```



Click for answer

Answer: No, the CI does not capture the difference of 0.

- What is the bootstrap SE? Is it similar to the randomization distribution SE?

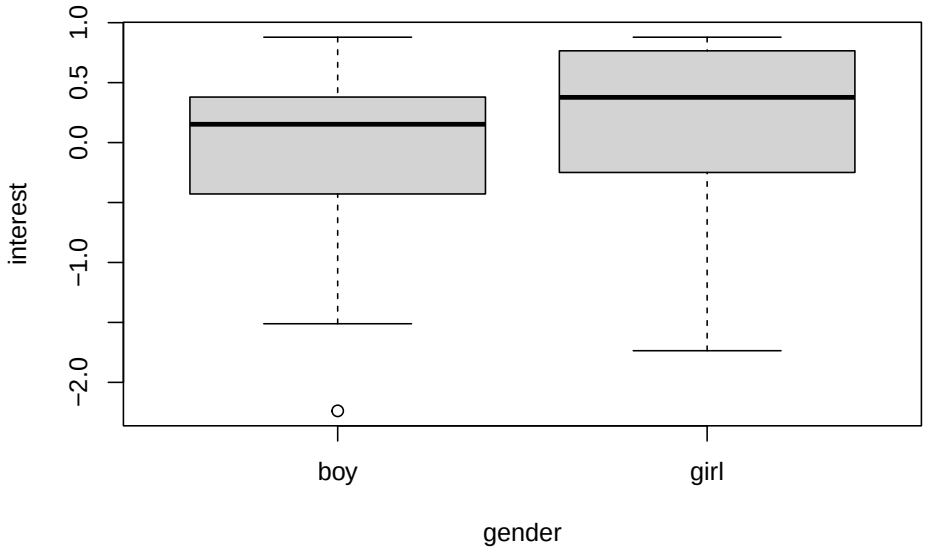
Click for answer

Answer: SE is about 0.21, which is very similar to the randomization distribution SE.

21.1.0.5 (e) Interest in 5 year olds

Redo the randomization test and bootstrap CI for 5 year olds, but this time omit the outlier boy case that has a very low interest level. Recall how to use the `which` command:

```
boxplot(interest ~ gender, data=study4age5)
```

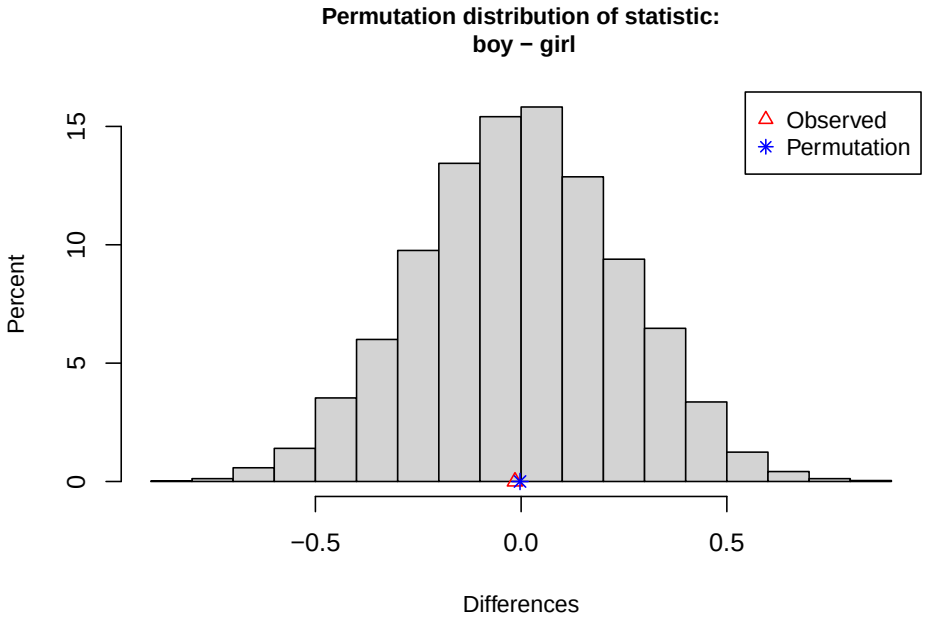



```
which(study4age5$interest < -2)
```

```
[1] 39
```

Then to omit this case, add the argument `subset = -39` to the `permTest` and `boot` commands used in (a) and (b).

```
set.seed(7)
permTest(interest ~ gender, data = study4age5, subset = -39)
```

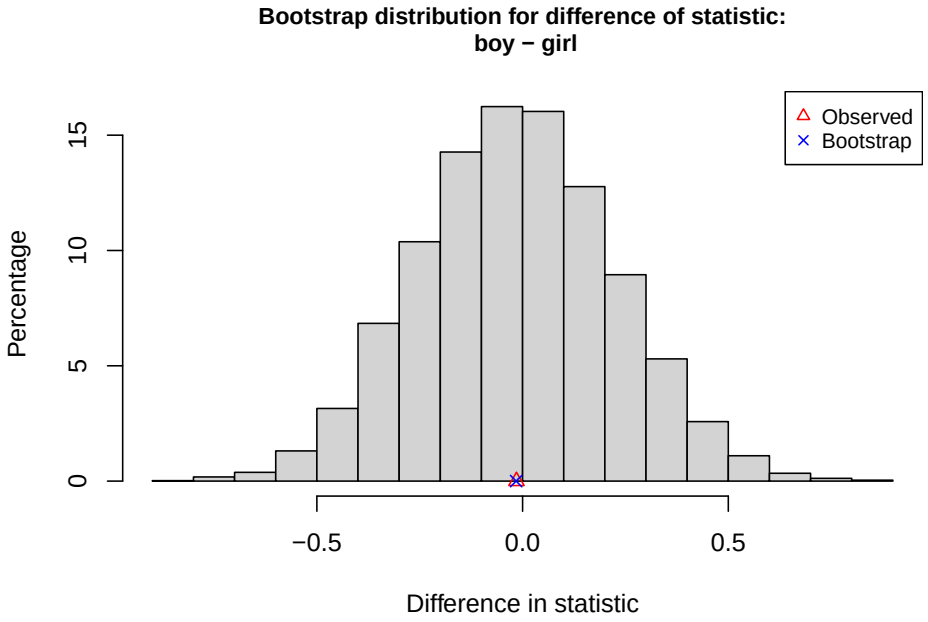



```
   ** Permutation test **

Permutation test with alternative: two.sided
Observed statistic
 boy :  0.01417     girl :  0.02906
Observed difference: -0.01488

Mean of permutation distribution: -0.00265
Standard error of permutation distribution: 0.2469
P-value:  0.957

   *-------------*
```

```
boot(interest ~ gender, data = study4age5, subset = -39)
```

```
   ** Bootstrap interval for difference of statistic

Observed difference of statistic:  boy - girl =  -0.01488
Mean of bootstrap distribution: -0.01565
Standard error of bootstrap distribution: 0.23826

Bootstrap percentile interval
     2.5%      97.5%
```

```
-0.4751690  0.4520149

          *--------------*
```



- Does the observed difference get closer or further from 0 with the case omitted? Explain why it changes.

Click for answer

Answer: The very low case pulls down the mean response for boys (with: $\bar{x}_{B5} = -0.10435$, without: $\bar{x}_{B5} = 0.01417$). Since the girl mean response doesn't change ($\bar{x}_{G5} = 0.02906$), omitting this case will make the *two means closer together* which makes their difference closer to 0 (with: $\bar{x}_{B5} - \bar{x}_{G5} = -0.13341$, without: $\bar{x}_{B5} - \bar{x}_{G5} = -0.01488$).

- Do the SEs of the distributions (bootstrap and randomization) get smaller or larger with the case omitted? Explain why these change.

Click for answer

Answer: The very low case creates larger variability in the sample mean for boys, which in turn makes the SE for the sample mean difference more variable (with: SE about 0.26, without: SE about 0.24).

- Compute the z-score for the observed difference in means using randomization distribution. Is this value futher or closer to a z-score of 0 with the case omitted? Explain why it changes.

Click for answer

Answer: The z-score is closer to 0 with the case removed (with: $z = -0.51$, without: $z = -0.061$)

$$z = \frac{-0.01488 - 0}{0.24569} = -0.061$$

The z-score is closer to 0 because the sample mean difference is closer to 0 with the case removed, and this change is greater than the relatively small decrease in SE that we noted with the case removed.

- Does the p-value get smaller or larger (or doesn't change) with the case omitted? Explain why it changes.

Click for answer

Answer: The p-value is larger with the case removed (with: p-value = 0.617, without: p-value = 0.944). This is because the observed difference is closer to 0 (fewer SE away) with the case omitted.

Chapter 22

Class Activity 15

22.1 Example 1: SAT Verbal scores

Suppose that the verbal SAT scores in a population are normally distributed with a mean $\mu = 580$ and standard deviation $\sigma = 70$. If X is shorthand for a verbal SAT score, then we can write this as $X \sim N(580, 70)$.

22.1.0.1 (a) What proportion of scores are above 650?

Click for answer

Answer: About 15.9% of the scores are above 650.

```
pnorm(650,mean=580,sd=70) # proportion below
```

```
[1] 0.8413447
```

```
1-pnorm(650,mean=580,sd=70) # proportion above
```

```
[1] 0.1586553
```

22.1.0.2 (b) What is the 25th percentile (Q1)?

Click for answer

Answer: The score of about 533 is the 25th percentile, meaning 25% of the scores are below this value.

```
qnorm(.25,mean=580,sd=70)
```

```
[1] 532.7857
```

22.1.0.3 (c) What is the IQR for verbal SAT scores in this population? (Hint: find Q1 and Q3)

Click for answer

Answer: The 25th percentile (Q1) is 533 and the 75th percentile (Q3) is 627. The IQR for this normally distributed variable is about 94 points.

```
q1 <- qnorm(.25,mean=580,sd=70);q1
```

```
[1] 532.7857
```

```
q3 <- qnorm(.75,mean=580,sd=70);q3
```

```
[1] 627.2143
```

```
q3-q1
```

```
[1] 94.42857
```

22.1.0.4 (d) What score, high or low, will be deemed an outlier according the boxplot rules for outliers?

Click for answer

Answer: Using the 1.5IQR's boxplot rule gives a lower fence of 392 and an upper fence of 768. So any score below 392 and above 768 will be called an outlier according to this rule.

```
1.5*94
```

```
[1] 141
```

```
q1 - 1.5*94
```

```
[1] 391.7857
```

```
q3 + 1.5*94
```

```
[1] 768.2143
```

22.1.0.5 (e) What percent of the population will be deemed an outlier?

Click for answer

Answer: We need to find the proportion of scores below 392 and above 768. With this symmetric distribution, we find about 0.004 in both tails. About 0.8% of the population will be deemed outliers according to the boxplot rule.

```
pnorm(392,mean=580,sd=70)
```

```
[1] 0.003618747
```

```
1-pnorm(768,mean=580,sd=70)
```

```
[1] 0.003618747
```

22.2 Example 2: Standard Normal

The standard normal distribution has a mean of 0 and standard deviation of 1.

22.2.0.1 (a) What percent of SAT scores are at least 1 standard deviation above average?

Click for answer

```
pnorm(1)  # proportion below
```

```
[1] 0.8413447
```

```
1-pnorm(1) # proportion above
```

```
[1] 0.1586553
```

Answer: About 16% of scores will be at least 1 standard deviation above average. (Note that the score of 580+70 = 650 is 1 standard deviation above average.)

22.2.0.2 (b) How many standard deviations away from average is the 25th percentile of SAT scores?

Click for answer

Answer: The 25th percentile of SAT scores (or any normally distributed values) is 0.67 standard deviations below average. We could also find this value using our answer to (1b):

```
qnorm(.25)
```

```
[1] -0.6744898
```

$$z = \frac{533 - 580}{70} = -0.67$$

```
(533 - 580)/70
```

```
[1] -0.6714286
```

Chapter 23

Class Activity 16

23.1 Example 1: Is Divorce Morally Acceptable?

In a study, we find that 67% of women in a random sample view divorce as morally acceptable. Does this provide evidence that more than 50% of women view divorce as morally acceptable? The standard error for the estimate assuming the null hypothesis is true is 0.021.

23.1.0.1 (a) What are the null and alternative hypotheses for this test?

Click for answer

Answer: If p denotes the proportion of woman who view divorce as morally acceptable in the population, then our hypotheses are

$$H_0 : p = 0.5 \quad H_A : p > 0.5$$

23.1.0.2 (b) What is the standardized test statistic?

Click for answer

Answer: The observed sample proportion is 0.67 with a standard error of 0.021. If the null is true, then we would expect the sampling distribution of the sample mean to be (approximately) normally distributed with a center of 0.50 and SE of 0.021. The standardized score for the sample proportion is then

$$z = \frac{\text{statistic} - \text{null parameter}}{SE} = \frac{0.67 - 0.50}{0.021} = 8.10$$

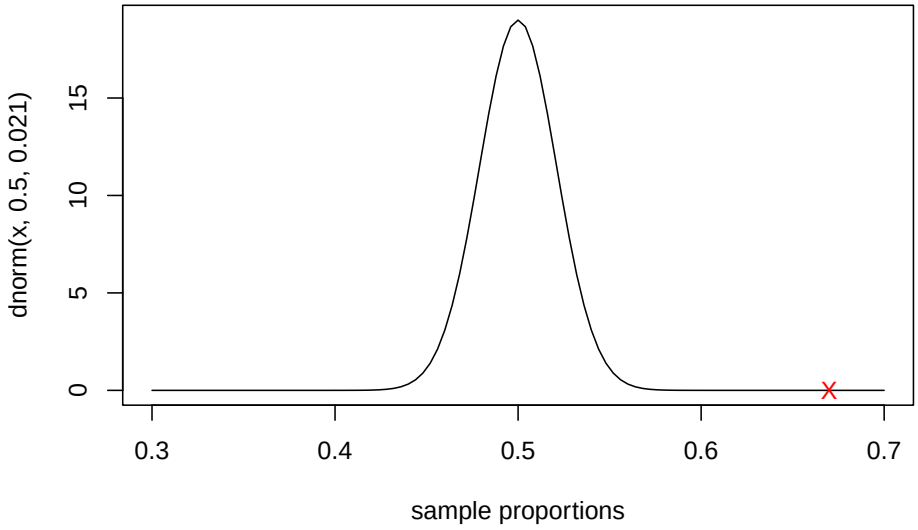
The observed proportion is 8.1 SEs above the hypothesized value of 0.5.

```
(0.67 - 0.5)/0.021
```

```
[1] 8.095238
```

Note that the randomization distribution should look roughly like this (with the observed proportion denoted with a red X):

```
curve(dnorm(x,0.5,.021),from=.3,to=.7,xlab="sample proportions")
points(0.67,0,pch="X",col="red")
```



23.1.0.3 (c) Use the normal distribution to find the p-value.

Click for answer

Answer: As we can see in the normal plot above, the p-value will be very small because the alternative is looking for big sample proportions. The p-value is the proportion of times we get a sample proportion as big, or bigger than, 0.67; or equivantly, the proportion of times we get a sample proportion that is at least 8.1 SEs above the hypothesized proportion. We would report a p-value that is less than 0.0001.

```
1-pnorm(8.10,0,1)
```

```
[1] 2.220446e-16
```

23.1.0.4 (d) What is the conclusion of the test?

Click for answer

Answer: The p-value is very small so we have very strong evidence that more than 50% of all women view divorce as morally acceptable.

23.1.0.5 (e) Use the normal distribution to find a 99% confidence interval for the proportion of all women who view divorce as morally acceptable. Interpret your answer.

Click for answer

Answer: Without knowing the bootstrap SE, our best guess at it would be from the randomization distribution SE which is given as 0.021. Our 99% confidence interval will look like:

$$statistic \pm z^* SE = 0.67 \pm z^*(0.021)$$

The z^* for a 99% CI corresponds to the 99.5th percentile (90% in middle + 0.5% in the left tail). With $z^* = 2.576$, we get a 99% confidence interval of 0.616 to 0.724.

```
qnorm(0.995)
```

```
[1] 2.575829
```

```
0.67 - 2.576*0.021
```

```
[1] 0.615904
```

```
0.67 + 2.576*0.021
```

```
[1] 0.724096
```

23.2 Example 2: Do Men and Women Differ in Opinions about Divorce?

In the same study described above, we find that 71% of men view divorce as morally acceptable. Use this and the information in the previous example to test whether there is a significant difference between men and women in how they view divorce. The standard error for the difference in proportions under the null hypothesis that the proportions are equal is 0.029.

23.2.0.1 (a) What are the null and alternative hypotheses for this test?

Click for answer

Answer: Using the same notation as (3a), except denoting male/female populations, we get

$$H_0 : p_f = p_m \quad H_A : p_f \neq p_m$$

23.2.0.2 (b) What is the standardized test statistic?

Click for answer

Answer: Suppose we look at the difference $p_m - p_f$. The observed difference is then 0.04 (0.71 - 0.67). This value is about 1.4 SEs above the hypothesized difference of 0:

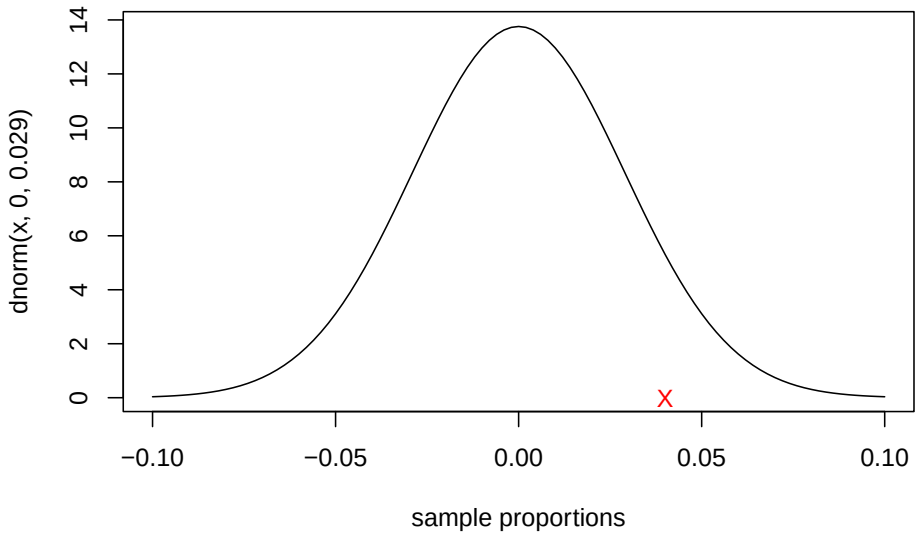
$$z = \frac{\text{statistic} - \text{null parameter}}{SE} = \frac{(0.71 - 0.67) - 0}{0.029} = 1.379$$

```
(0.04 - 0)/0.029
```

```
[1] 1.37931
```

Note that the randomization distribution for the difference in sample proportions should look roughly like this (with the observed proportion difference denoted with a red X):

```
curve(dnorm(x,0,.029),from=-.1,to=.1,xlab="sample proportions")
points(0.04,0,pch="X",col="red")
```



23.2.0.3 (c) Use the normal distribution to find the p-value.

Click for answer

Answer: This is a two-tail test. Since the observed difference is less than 2 SEs away from 0 we know that the (two-tailed) p-value should be bigger than 0.05. We see that the p-value is $2(0.084) = 0.168$.

```
1-pnorm(1.379,0,1) # proportion above z=1.379
```

```
[1] 0.08394738
```

```
2*(1-pnorm(1.379,0,1)) # p-value for two-sided
```

```
[1] 0.1678948
```

23.2.0.4 (d) What is the conclusion of the test?

Click for answer

Answer: The p-value is larger than a 5% significance level, so we do not find evidence of a difference between men and women in the proportion that view divorce as morally acceptable. About 17% of the time we would observe a difference in male/female views of 4 percentage points or greater just by chance.

Chapter 24

Class Activity 17

24.1 Example 1: Movie Goers are More Likely to Watch at Home

In a random sample of 500 movie goers in January 2013, 320 of them said they are more likely to wait and watch a new movie in the comfort of their own home. Compute and interpret a 95% confidence interval for the proportion of movie goers who are more likely to watch a new movie from home.

Click for answer

Answer: We see that $\hat{p} = \frac{320}{500} = 0.640$ (keep at least 3 decimal spots to ensure accuracy in your SE calculation!) The confidence interval is given by:

$$\text{Statistic } \pm z^* SE$$

$$\begin{aligned}
&\hat{p} \pm z^* \cdot \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} \\
&0.64 \pm 1.96 \cdot \sqrt{\tfrac{0.64(1-0.64)}{500}} \\
&0.64 \pm 0.042 \\
&(0.598, 0.682)
\end{aligned}$$

(Make sure to use proportions in your CI, then convert to % at the end if you prefer a percentage interpretation.) We are 95% sure that the proportion of all movie goers who are more likely to wait and watch a new movie at home is between 0.598 and 0.682.

24.2 Example 2: Sample Size and Margin of Error for Movie Goers

(a) What sample size is needed in example 2 if we want a margin of error within ±2%? (Use the sample proportion from the original sample.)

Click for answer

Answer:

$$\begin{aligned} 0.02 &= z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ n &= \left(\frac{z^*}{0.02}\right)^2 \hat{p}(1-\hat{p}) \\ &= \left(\frac{1.96}{0.02}\right)^2 0.64(1-0.64) = 2212.76 \end{aligned}$$

We need a sample size of at least n = 2,213 to have a margin of error this small. This is substantially more than the sample size of 500 used in the actual survey.

(b) What sample size is needed if we want a margin of error within ±2%, and if we use the conservative estimate of p = 0.5?

Click for answer

Answer:

$$n = \left(\frac{1.96}{0.02}\right)^2 0.5(1-0.5) = 2401$$

We need a sample size of at least n = 2,401 to have a margin of error this small. Notice that if we have less knowledge of the actual proportion, we need a larger sample size to arrive at the same margin of error.

24.3 Example 3: Mendel's green peas?

One of Gregor Mendel's famous genetic experiments dealt with raising pea plants. According to Mendel's genetic theory, under a certain set of conditions the proportion of pea plants that produce smooth green peas should be p=3/16 (0.1875). A sample of n=556 plants from the experiment had 108 with smooth green peas. Does this provide evidence of a problem with Mendel's theory and that the proportion is different from 3/16? Show all details of the test.

Click for answer

Answer: We are testing $H_0 : p = 0.1875$ vs $H_a : p \neq 0.1875$ where p represents the proportion of pea plans with smooth green peas. The sample proportion is $\hat{p} = \frac{108}{556} = 0.1942$ and the sample size is $n = 556$. The test statistic is:

$$z = \frac{\text{Statistic } - \text{ Null}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1942 - 0.1875}{\sqrt{\frac{0.1875(1-0.1875)}{556}}} = 0.405$$

This is a two-tail test, and we see that the area to the right of 0.405 in a normal distribution is 0.343 `(1-pnorm(0.405))`, so the p-value is $2(0.343) = 0.686$.The R command is: `2*(1-pnorm(0.405))`

We do not reject H0 and conclude that this sample does not provide evidence that the proportion of smooth green pea plants is different from the 3/16 that Mendel's theory predicts. (It is worth pointing out that this does not "prove" Mendel's theory, since we don't "accept" H_0– we just find a lack of sufficient evidence to refute it.)

Chapter 25

Class Activity 18

25.1 Example 1: Change in gun ownership

A 2016 study described in The Guardian found that a random sample of US adults in 1994 found a female rate of gun ownership of 9%. A similar random sample in 2015 found the rate of female gun ownership rose to 12%. In the section 3.2 handout, we assumed that the SE for the difference in these two sample proportions is 2%. Show how this SE was computed using the appropriate SE formula from chapter 6. Assume that the sample sizes in both 1994 and 2015 were 500.

Click for answer

Answer: We have a 1994 sample proportion of $\hat{p}_{1994} = 0.09$ and a 2015 sample proportion of $\hat{p}_{2015} = 0.12$. The SE of the difference in two sample proportions for a confidence interval is given by:

$$SE = \sqrt{\frac{\hat{p}_{1994}(1-\hat{p}_{1994})}{n_{1994}} + \frac{\hat{p}_{2015}(1-\hat{p}_{2015})}{n_{2015}}} = 0.0194 \approx 0.02$$

25.2 Example 2: Accuracy of Lie Detectors

Participants in a study to evaluate the accuracy of lie detectors were divided into two groups, with one group reading true material and the other group reading false material, while connected to a lie detector. Both groups received electric shocks to add stress. The two way table indicates whether the participants were lying or telling the truth and also whether the lie detector indicated they were lying or not.

	Detector Says Lying	Detector Says Not	Total
Person Lying	31	17	48
Person Not	27	21	48
Total	58	38	96

25.2.1 (a) Are the conditions met for using the normal distribution?

Click for answer

Answer: Yes (all cell counts at least 10)

25.2.2 (b) Find the three sample proportions for the proportion of times the lie detector says the person is lying (the proportion for the lying people, the proportion for the truthful people, and the pooled proportion).

Click for answer

Answer: We see that the proportion for the lying people is $\widehat{p}_L = \frac{31}{48} = 0.6458$, the proportion for the not lying people is $\widehat{p}_N = 0.5625$, and the pooled proportion for all 96 people is $\widehat{p} = \frac{58}{96} = 0.6042$.

25.2.3 (c) Test to see if there is a difference in the proportion of times the lie detector says the person is lying, depending on whether the person is lying or telling the truth. Show all details of the hypothesis test.

Click for answer

Answer:

We are testing $H_0 : p_L = p_N$ vs $H_a : p_L \neq p_N$. The test statistic is

$$z = \frac{statistic - null}{SE} = \frac{(\widehat{p}_L - \widehat{p}_N) - 0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_l} + \frac{\widehat{p}(1-\widehat{p})}{n_N}}} = \frac{0.6458 - 0.5625}{\sqrt{\frac{0.6042(1-0.6042)}{48} + \frac{0.6042*(1-0.6042)}{48}}} = 0.834$$

This is a two-tail test, and the area to the right of 0.834 in a normal distribution is 0.202 (1-pnorm(0.834)), so the p-value is 2(0.202) = 0.404. The R command is: 2*(1-pnorm(0.834))

```
pL_hat = 31/48
pN_hat = 27/48
pooled_p = 58/96
nL = 48
nN = 48
SE = sqrt(pooled_p*(1-pooled_p)*(1/nL + 1/nN))
z = (pL_hat - pN_hat) / SE
p_value = 2*(1-pnorm(z))
p_value
```

```
[1] 0.4038223
```

We fail to reject H0 and conclude that there is not enough evidence that a lie detector can tell whether a person is lying or telling the truth.

25.2.4 (d) Calculate a 95% confidence interval for the difference in proportions of people correctly identified by the lie detector.

Click for answer

```
conf_level = 0.95
z_star = qnorm(1-(1-conf_level)/2)
margin_of_error = z_star * SE
CI_lower = (pL_hat - pN_hat) - margin_of_error
CI_upper = (pL_hat - pN_hat) + margin_of_error
CI = c(CI_lower, CI_upper)
CI
```

```
[1] -0.1123154  0.2789821
```

The 95% confidence interval for the difference in proportions is (-0.299, 0.207). Since, the confidence interval includes the null hypothesized value of 0, we do not reject the null hypothesis, and conclude that there is not enough evidence that a lie detector can tell whether a person is lying or telling the truth.

25.3 Example 3: Smoking and Pregnancy Rate?

Does smoking negatively affect a person's ability to become pregnant? A study collected data on 678 women who were trying to get pregnant. The two-way table shows the proportion who successfully became pregnant during the first cycle trying and smoking status.

25.3.1 (a). Find a 90% confidence interval for the difference in proportion of women who get pregnant, between smokers and non-smokers. Interpret the interval in context.

	Smoker	Non-smoker	Total
Pregnant	38	206	244
Not Pregnant	97	337	434
Total	135	543	678

Click for answer

The conditions are met for using the normal distribution (at least 10 values in each cell of the table). We see that the proportion of smokers who got pregnant is $38/135 = 0.281$ while the proportion of non-smokers who got pregnant is $206/543 = 0.379$. The confidence interval is given by:

$$statistic \pm z^* \cdot SE$$

$$(\hat{p}_S - \hat{p}_N) \pm z^* \cdot \sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{n_S} + \frac{\hat{p}_N(1-\hat{p}_N)}{n_N}}$$

$$(0.281 - 0.379) \pm 1.645 \cdot \sqrt{\frac{0.281(1-0.281)}{135} + \frac{0.379(1-0.379)}{543}}$$

$$-0.098 \pm 0.072 = (-0.170, -0.026)$$

We are 90% sure that the proportion of smokers who get pregnant in the first cycle is between 0.170 and 0.026 less than the proportion of non-smokers who get pregnant on the first cycle. Note that if we had subtracted the other way, the interval would have only positive values, but the interpretation would be the same.

25.3.2 (b). Now, repeat the above analysis using the hypothesis test approach.

Click for answer

We are testing $H_0 : p_S = p_{NS}$ vs $H_a : p_S \neq p_{NS}$. The test statistic is:

```
pS_hat = 38/135
pNS_hat = 206/543
pooled_p2 = (38+206)/(135+543)
nS = 135
```

```
nNS = 543
SE2 = sqrt(pooled_p2*(1-pooled_p2)*(1/nS + 1/nNS))
z2 = (pS_hat - pNS_hat) / SE2
p_value2 = 2*(pnorm(z2))
p_value2
```

```
[1] 0.03394234
```

Based on the p-value, we reject H_0 and conclude that there is a difference in the proportion of women who get pregnant between smokers and non-smokers.

Chapter 26

Class Activity 19

26.1 Example 1: Florida Lakes pH

The textbook dataset `FloridaLakes` contains data on 53 lakes in Florida. We want to know if the average pH of lakes in Florida is different from a neutral value of 7.

```
lakes <- read.csv("http://www.lock5stat.com/datasets1e/FloridaLakes.csv")
head(lakes)
```

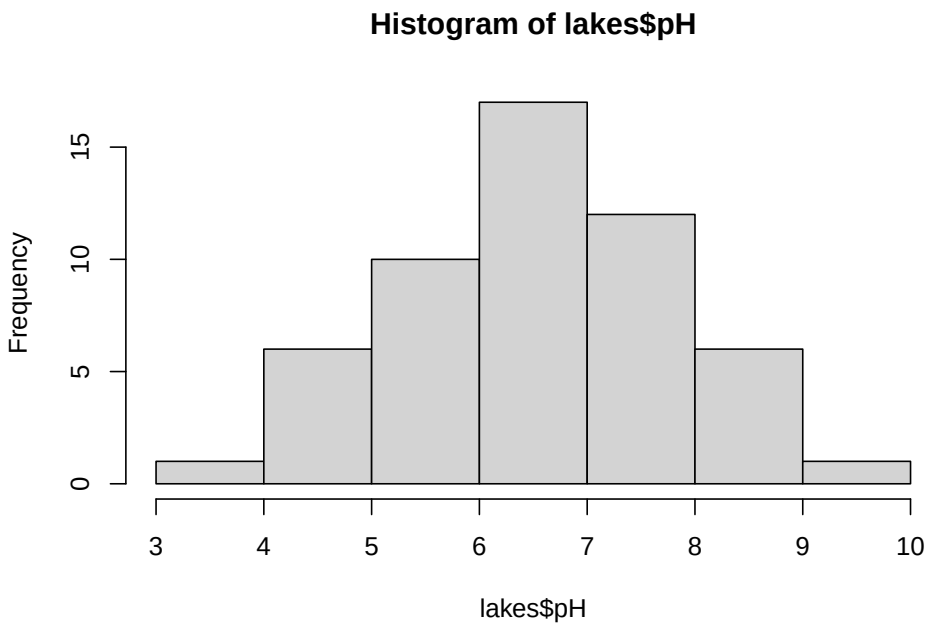
```
  ID         Lake Alkalinity  pH Calcium Chlorophyll
1  1    Alligator        5.9 6.1     3.0         0.7
2  2        Annie        3.5 5.1     1.9         3.2
3  3       Apopka      116.0 9.1    44.1       128.3
4  4 Blue Cypress       39.4 6.9    16.4         3.5
5  5        Brick        2.5 4.6     2.9         1.8
6  6       Bryant       19.6 7.3     4.5        44.1
  AvgMercury NumSamples MinMercury MaxMercury
1       1.23          5       0.85       1.43
2       1.33          7       0.92       1.90
3       0.04          6       0.04       0.06
4       0.44         12       0.13       0.84
5       1.20         12       0.69       1.50
6       0.27         14       0.04       0.48
  ThreeYrStdMercury AgeData
1              1.53       1
2              1.33       0
3              0.04       0
4              0.44       0
```

```
5              1.33       1
6              0.25       1
```

26.1.0.1 (a) EDA

Always plot your data and get summary stats:

```
hist(lakes$pH)
```




```
mean(lakes$pH)
```

```
[1] 6.590566
```

```
sd(lakes$pH)
```

```
[1] 1.288449
```

- What are the sample mean and standard deviation? Use appropriate notation.
- Can we use t-inference methods with the pH variable?

Click for answer

Answer: The average pH was $\bar{x} = 6.591$ with a standard deviation of $s = 1.288$. The distribution of pH is symmetric with no outliers, so we can use t-inference methods.

26.1.0.2 (b) SE for the sample mean

What is the estimated SE for the sample mean?

Click for answer

Answer: The estimated SE for the sample mean is $SE_{\bar{x}} = 0.1770$.

```
sd(lakes$pH)/sqrt(53)
```

```
[1] 0.1769821
```

26.1.0.3 (c) t-test statistic

Using your SE from (b) to compute the t-test statistic for testing if the population mean pH is equal, or not, to 7. Write down your hypotheses then show how the t test statistic is calculated. Interpret this value in context.

Click for answer

Answer: The hypotheses are $H_0 : \mu = 7$ vs $H_A : \mu \neq 7$. The test stat is

$$t = \frac{6.591 - 7}{1.288/\sqrt{53}} = -2.3134$$

The observed mean of 6.591 is 2.3 SEs below the hypothesized mean of 7.

```
(mean(lakes$pH) - 7)/(sd(lakes$pH)/sqrt(53))
```

```
[1] -2.31342
```

26.1.1 (d) One-sample t-test

The function `t.test(x, mu=)` can be used for a one sample test comparing the sample mean of `x` to the hypothesized value given to `mu=`. Here we are testing whether the population mean is equal to 7 or not:

```
t.test(lakes$pH, mu = 7)
```

```

    One Sample t-test

data:  lakes$pH
t = -2.3134, df = 52, p-value = 0.02469
```

```
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.235425 6.945707
sample estimates:
mean of x
 6.590566
```

- What is the t test stat given in the output? Verify that it matches your answer to (c), within reasonable rounding error.

Click for answer

Answer: The test stat is -2.31.

- What is the p-value for the test? Interpret this value.

Click for answer

Answer: The p-value is 0.025. If the mean pH of all lakes is 7, then we would see a sample mean that is at least 2.31 SEs away from 7 about 2.5% of the time in samples of 53 lakes.

- What is your test conclusion?

Click for answer

Answer: There is a statistically significant difference between the observed mean pH of 6.591 and the hypothesized mean of 7 (t=-2.31, df=52, p=0.025).

26.1.1.1 (e) One-sample t confidence interval

What is the 95% confidence interval for the population mean pH? Interpret this CI.

Click for answer

Answer: We are 95% confident that the mean pH of all lakes in Florida is between 6.24 and 6.95.

26.1.1.2 (f) `qt` and `pt`

Show how to compute the p-value for your test in (d) using the `pt` command. Then show how the confidence interval in (e) is computed with a `qt` value.

Click for answer

Answer: For the two-sided test, the p-value is twice the proportion below the test stat $t = -2.313$ under a t-distribution with $df = 53 - 1 = 52$

```
2*pt(-2.313,df=52)
```

```
[1] 0.02471195
```

For a 95% CI, we get the 97.5th percentile from the same t-distribution

```
qt(.975,52)
```

```
[1] 2.006647
```

26.2 Example: Nutrition Study

The dataset `NutritionStudy` contains data on daily calorie intake and other variables for 315 individuals. We want to know if the average daily calorie intake is different from the recommended 2000 calories.

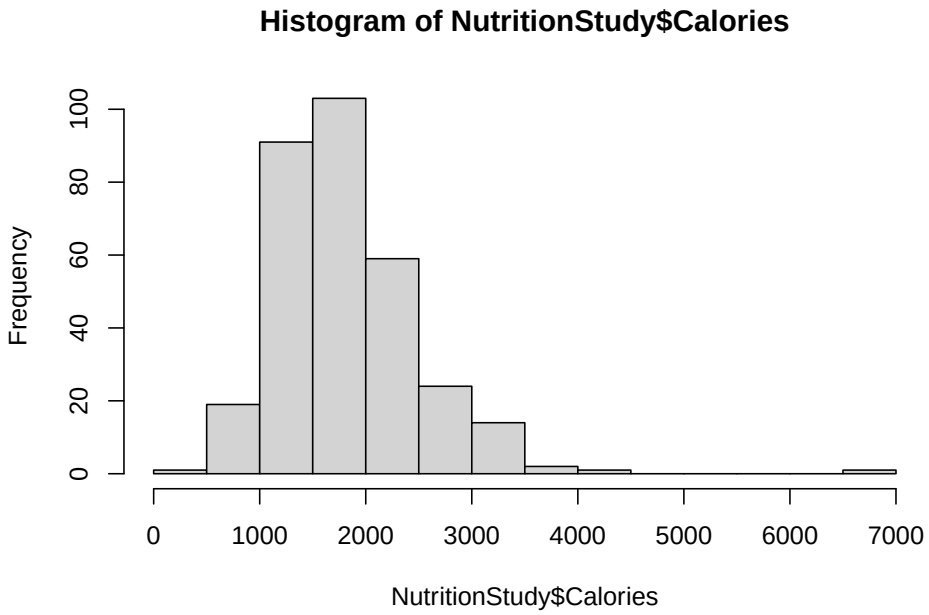
```
#install.packages("Lock5Data")
library(Lock5Data)
library(dplyr)
data(NutritionStudy)
glimpse(NutritionStudy)
```

```
Rows: 315
Columns: 17
$ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1~
$ Age           <int> 64, 76, 38, 40, 72, 40, 65, 58, 35, ~
$ Smoke         <chr> "No", "No", "No", "No", "No", "No", ~
$ Quetelet      <dbl> 21.4838, 23.8763, 20.0108, 25.1406, ~
$ Vitamin       <int> 1, 1, 2, 3, 1, 3, 2, 1, 3, 3, 1, 2, ~
$ Calories      <dbl> 1298.8, 1032.5, 2372.3, 2449.5, 1952~
$ Fat           <dbl> 57.0, 50.1, 83.6, 97.5, 82.6, 56.0, ~
$ Fiber         <dbl> 6.3, 15.8, 19.1, 26.5, 16.2, 9.6, 28~
$ Alcohol       <dbl> 0.0, 0.0, 14.1, 0.5, 0.0, 1.3, 0.0, ~
$ Cholesterol   <dbl> 170.3, 75.8, 257.9, 332.6, 170.8, 15~
$ BetaDiet      <int> 1945, 2653, 6321, 1061, 2863, 1729, ~
$ RetinolDiet   <int> 890, 451, 660, 864, 1209, 1439, 802,~
$ BetaPlasma    <int> 200, 124, 328, 153, 92, 148, 258, 64~
$ RetinolPlasma <int> 915, 727, 721, 615, 799, 654, 834, 8~
$ Sex           <chr> "Female", "Female", "Female", "Femal~
$ VitaminUse    <chr> "Regular", "Regular", "Occasional", ~
$ PriorSmoke    <int> 2, 1, 2, 2, 1, 2, 1, 1, 1, 2, 2, 1, ~
```

26.2.1 (a) EDA

Always plot your data and get summary stats:

```
hist(NutritionStudy$Calories)
```




```
mean(NutritionStudy$Calories)
```

```
[1] 1796.655
```

```
sd(NutritionStudy$Calories)
```

```
[1] 680.3474
```

26.2.2 (b) SE for the sample mean

What is the estimated SE for the sample mean?

Click for answer

Answer:

```
n <- length(NutritionStudy$Calories)
SE <- sd(NutritionStudy$Calories) / sqrt(n)
SE
```

```
[1] 38.33324
```

26.2.3 (c) t-test statistic

Compute the t-test statistic for testing if the population mean calorie intake is equal, or not, to 2000.

Click for answer

Answer:

```
t_stat <- (mean(NutritionStudy$Calories) - 2000) / SE
t_stat
```

```
[1] -5.304676
```

26.2.4 (d) One-sample t-test

Perform a one-sample t-test to test whether the population mean calorie intake is equal to 2000 or not.

Click for answer

Answer:

```
t.test(NutritionStudy$Calories, mu = 2000)
```

```
    One Sample t-test

data:  NutritionStudy$Calories
t = -5.3047, df = 314, p-value = 2.135e-07
alternative hypothesis: true mean is not equal to 2000
95 percent confidence interval:
 1721.232 1872.077
sample estimates:
mean of x
 1796.655
```

26.2.5 (e) One-sample t confidence interval

What is the 95% confidence interval for the population mean calorie intake?

Click for answer

Answer:

```
ci <- t.test(NutritionStudy$Calories, mu = 2000)$conf.int
ci
```

```
[1] 1721.232 1872.077
attr(,"conf.level")
[1] 0.95
```

26.2.6 (f) qt and pt

Show how to compute the p-value for the test in (d) using the pt command. Then show how the confidence interval in (e) is computed with a qt value.

Click for answer

Answer:

```
p_value <- 2 * pt(-abs(t_stat), df = n - 1)
p_value
```

```
[1] 2.135134e-07
```

```
t_star <- qt(0.975, df = n - 1)
t_star
```

```
[1] 1.967548
```

Chapter 27

Class Activity 20

27.1 Example 1: API

The Academic Performance Index (API) is computed for all California schools. It is a number, ranging from a low of 200 to a high of 1000, that reflects a school's performance on a statewide standardized test (http://api.cde.ca.gov). We have a SRS of 200 schools and are interested in how a school's performance is related to the wealth of its students. The variable `growth` measures the growth in API from 1999 to 2000 (API 2000 - API 1999).

```
api <- read.csv("http://people.carleton.edu/~kstclair/data/api.csv")
```

27.1.0.1 (a) Categorizing wealth

Let's define a school as "low wealth" if over 50% of its students are eligible for subsidized `meals` and "high wealth" otherwise. We can use an `ifelse` command to create a variable `wealth` that measures this:

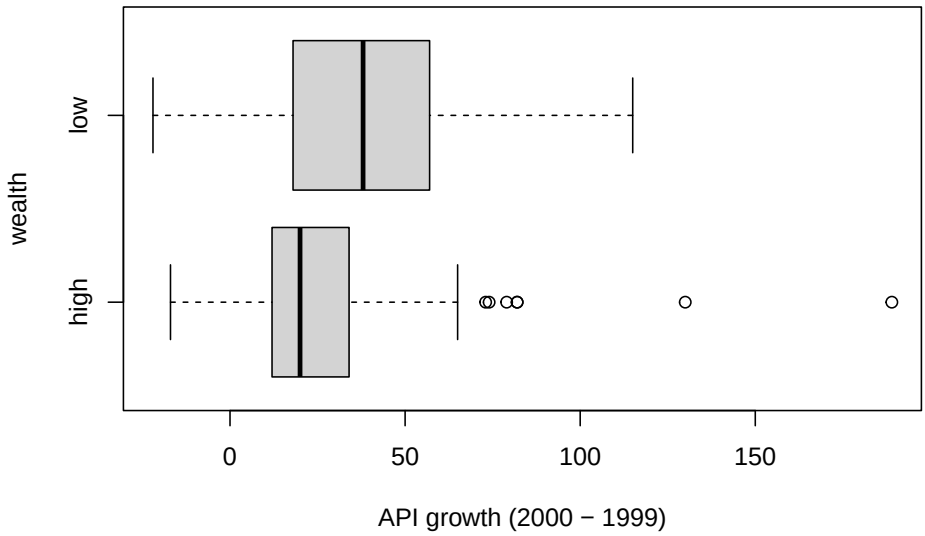
```
api$wealth <- ifelse(api$meals > 50, "low","high")
table(api$wealth)
```

```
high  low
 102   98
```

```
library(dplyr)
api %>% group_by(wealth) %>% summarize(mean(growth), sd(growth))
```

```
# A tibble: 2 x 3
  wealth `mean(growth)` `sd(growth)`
  <chr>           <dbl>        <dbl>
1 high             25.2         28.8
2 low              38.8         30.0
```

```
boxplot(growth ~ wealth, data=api, xlab="API growth (2000 - 1999)" , horizontal=T)
```



- How many schools are "low" and "high" wealth.
- Are wealth and API growth related?
- What is the observed difference in mean API growth between high and low wealth schools. Use correct notation.
- Can we use t-inference methods to compare mean growths?
  Click for answer

Answer: There are $n_h = 102$ "high" wealth and $n_l = 98$ "low" wealth schools. The low wealth schools tend to have higher (and more variable) growth than high wealth schools. The difference in observed mean API growth between high and low growth schools is $\bar{x}_h - \bar{x}_l = 25.24510 - 38.82653 = -13.58$. We can use t-methods since both samples sizes (98 and 102) can be deemed large and there isn't severe skewness, but there are two extreme outliers that will be addressed below.

27.1.0.2 (b) SE for the sample mean difference

What is the estimated SE for the sample mean difference?

Click for answer

Answer: The SE for the mean difference is 4.1544:

$$SD_{\bar{x}_h - \bar{x}_l} = \sqrt{\frac{28.75380^2}{102} + \frac{29.95048^2}{98}} = 4.1544$$

```
sqrt(28.75380^2/102 +  29.95048^2/98)
```

```
[1] 4.154404
```

27.1.0.3 (c) t-test statistic

Using your SE from (b) to compute the t-test statistic that can be used to determine if mean API growth differs for low and high wealth schools. Write down your hypotheses then show how the t test statistic is calculated. Interpret this value in context.

Click for answer

Answer: The hypotheses are $H_0 : \mu_h - \mu_l = 0$ vs $H_A : \mu_h - \mu_l \neq 0$. The test stat is

$$t = \frac{(25.24510 - 38.82653) - 0}{4.154404} = -3.2692$$

The observed mean difference is 3.3 SEs below the hypothesized mean difference of 0.

```
((25.24510 - 38.82653) - 0)/4.154404
```

```
[1] -3.269164
```

27.1.0.4 (d) Two-sample t-test

Is there evidence that mean API growth differs for low and high wealth schools? Give the hypotheses for this test, then run the `t.test(y ~ x, data=)` command below to conduct a t-test to give a p-value and conclusion.

```
t.test(growth ~ wealth, data=api)
```

```
    Welch Two Sample t-test

data:  growth by wealth
t = -3.2692, df = 196.71, p-value = 0.001273
alternative hypothesis: true difference in means between group high and group low is no
95 percent confidence interval:
 -21.774321  -5.388544
sample estimates:
mean in group high  mean in group low
          25.24510           38.82653
```

- What is the t test stat given in the output? Verify that it matches your answer to (c), within reasonable rounding error.

Click for answer

Answer: The test stat matches, $t = -3.2692$.

- What is the p-value for the test? Interpret this value.
  Click for answer

Answer: The p-value is 0.001273. If there is no difference between mean growth in the two populations, then there is just a 0.13% chance of seeing a sample mean difference that is 3.27 standard errors or more away from 0.

- What is your test conclusion?
  Click for answer

Answer: We have strong evidence to suggest that the average API growth in low and high wealth schools are not the same.

27.1.0.5 (e) Consider outliers

The boxplot in (a) shows a number of outliers for the `high` wealth group, but two cases in particular were very high. Suppose we omitted these two (most) extreme cases when running the test in (d). Will the p-value for this test be smaller or larger than the p-value computed in part (d)? Explain.

Click for answer

Answer: Removing the two large outliers which will both reduce the mean in the high group and reduce the SD in the high group. Both actions will magnify the difference in mean growth between the high and low groups (increasing the difference and decreasing the SE), so the test stat will increase in magnitude and the p-value will decrease.

27.1.0.6 (f) Check outlier influence

To omit these cases we have to find their row numbers, then `subset` them out of the data:

```
which(api$growth > 120 )
```

```
[1]  74 119
```

```
api %>% slice(74,119)  # another dplyr package command
```

```
           cds stype            name                 sname
1 5.471911e+13     E Lincoln Element    Lincoln Elementary
2 1.975342e+13     E Washington Elem Washington Elementary
  snum                   dname dnum       cname cnum flag
1 5873 Exeter Union Elementary  226      Tulare   53   NA
2 2543   Redondo Beach Unified  585 Los Angeles   18   NA
  pcttest api00 api99 target growth sch.wide comp.imp both
1      98   693   504     15    189      Yes      Yes  Yes
2     100   745   615      9    130      Yes      Yes  Yes
  awards meals ell yr.rnd mobility acs.k3 acs.46 acs.core
1    Yes    50  18   <NA>        9     18     NA       NA
2    Yes    41  20   <NA>       16     19     30       NA
  pct.resp not.hsg hsg some.col col.grad grad.sch avg.ed
1       93      28  23       27       14        8   2.51
2       81      11  26       32       16       16   2.99
  full emer enroll api.stu    pw  fpc wealth
1   91    9    196     177 30.97 6194   high
2  100    3    391     313 30.97 6194   high
```

```
t.test(growth ~ wealth, data = api, subset = -c(74,119))
```

```
    Welch Two Sample t-test

data:  growth by wealth
t = -4.395, df = 174.97, p-value = 1.916e-05
alternative hypothesis: true difference in means between group high and group low is not equal to
95 percent confidence interval:
 -23.571116  -8.961945
sample estimates:
mean in group high  mean in group low
          22.56000           38.82653
```

- How does the t-test stat change when omitting these two changes? Why does it change in this direction?
- Check your answer here with your anwer in part (e)!
  Click for answer

Answer: Without these outliers, the p-value decreases to 0.00001916 and we have even stronger evidence for a difference in mean API growth. Why does the p-value decrease? Omitting the two outliers will decrease the sample SD for the high group, which in turn will (slightly) decrease the SE for the difference in means. Omitting the two outliers will also decrease the sample mean for the high group (from 25.24510 to 22.56000), which will make the observed difference in means larger in magnitude (from -13.58 to -16.27). The test stat gets even further from 0 (drops from -3.2692 to -4.395), meaning the observed difference with outliers omitted is further away from 0 (in terms of SEs) than it was when all data points were included. This means that the p-value will decrease (from 0.0013 to 0.00002) since the data is deemed more "extreme" under the null hypothesis.

27.1.0.7 (g) 95% confidence interval

Compare the two 95% CI given in the output (with and without outliers). Explain how and why the CIs change after omitting these two outliers.

Click for answer

Answer: Without outliers: -23.57 to -8.96 and with outliers: -21.77 to -5.39. An mentioned above, omitting the two points makes the difference in means further away from 0. This shifts the CI further from a difference of 0. Removing the outliers also decrease the SE of our sample difference, so the margin of error for the interval without outliers is, roughly, 7 while the margin of error with outliers is, roughly, 8.

27.1.0.8 (h) Interpret two-sample CI

Using the results without the two outliers, interpret the 95% CI given in this output. Do not use the word "difference'' in your answer.

Click for answer

Answer: We are 95% confident that the mean API growth between 1999 and 2000 for all low wealth schools is anywhere from 8.96 points to 23.57 points higher than the mean API growth for all high wealth schools in California.

27.2 Example 2: Matched Pairs

A study is conducted to determine the effect of a home meter for helping diabetics control their blood glucose levels. Researchers would like to determine if the home meter is effective in helping patients reduce their blood glucose levels. A random sample of 36 diabetics had their blood glucose levels measured before they were taught to use the meter and again after they had utilized the meter for 2 weeks. Researchers observed an average decrease (before - after) of blood glucose level of 2.78 mmol/liter with a standard deviation of 6.05 mmol/liter. Analysis results are shown below:

```
Sample mean:  2.78 ; sample standard deviation:  6.05 ; sample size: 36
Standard error:  1.0083
95 percent confidence interval for true mean:  1.0763  , Infinity
Hypothesis test H0: mu =  0  Alternative is  greater
t statistic =  2.757 ; degrees of freedom =  35 ; p-value= 0.0046
```

27.2.0.1 (a) What conditions need to be met by this data to use t inference procedures?

Click for answer

Answer: There is a moderate sample size of $n = 36$ so we need to assume that the observed differences (before-after) are not strongly skewed and that there are no outliers. If these assumptions are not met, then the t-inference procedures above may not be appropriate.

27.2.0.2 (b) Define the unknown parameter of interest (be very specific), then state the null and alternative hypotheses for this test. Make sure your hypotheses agree with the output!

Click for answer

Answer: Let the μ represent the population mean decrease in glucose levels measured before and after the treatment (before - after). A positive value of μ implies that the home meter is effective in reducing blood glucose levels. The alternative hypothesis (the research statement) will be that μ is greater than 0 and the null statement will be that μ is equal to 0, meaning there is no benefit to using the treatment.

$$H_0 : \mu = 0 \text{ vs. } H_A : \mu > 0$$

27.2.0.3 (c) What is the test statistic value for this test? What does this value indicate?

Click for answer

Answer: The test stat value is 2.757. The mean glucose level decrease in the sample was 2.757 SE's above the hypothesized mean decrease of 0.

27.2.0.4 (d) Is there sufficient evidence to claim that the monitor is effective in helping patients reduce their blood glucose levels?

Click for answer

Answer: You reject H_0 when the p-value is small. Since the P-value of 0.3% is quite small, we can conclude that there is strong evidence that the use of home meters lowers blood glucose levels, on average (H_A).

27.2.0.5 (e) What type of error (1 or 2) could you have made in part (d)? If you did make this error, what are its implications for people with diabetes?

Click for answer

Answer: Since we rejected, we may have made a type 1 error of rejecting the null when it is actually true. This means we would have claimed that the home meter was useful in reducing blood glucose levels, on average, when in fact it doesn't reduce levels. People with diabetes would be encouraged to use these meters (at a cost to themselves or their insurance company) to help control their glucose levels and not see any real benefit.

27.2.0.6 (f) Compute and interpret a 95% confidence interval for the true average decrease in blood glucose levels. (Note that this CI is not given above, the CI given in the output is a "one-sided" CI.)

Click for answer

Answer: The 95% CI for the population mean decrease in glucose level is

$$\bar{x} \pm t^*_{n-1}\frac{s}{\sqrt{n}} = 2.78 \pm 2.042\frac{6.05}{\sqrt{36}} = 2.78 \pm 2.017 = (0.72, 4.84)$$

where t^* is based on 36-1=35 degrees of freedom. Using the green table, we round df down to 30 so we get $t^*_{30} = 2.042$. Or using R command `qt(.975,df=35)` we get the exact value $t^*_{35} = 2.0301$. We are 95% confident that, after learning to use a home meter, the average decrease in blood glucose in this population is between 0.72 and 4.84 mmol/liter.

Chapter 28

Class Activity 21

28.1 Example 1: Food poisoning

Suppose in an outbreak, 447 of the 998 individuals who ate beef curry were observed to have food poisoning symptoms. As researchers, suppose we want to test the hypothesis that the probability (long run proportion) of a "random individual who ate beef curry" having food poisoning is 0.1. Conduct an appropriate hypothesis test.

Click for answer

*Answer:*The set of hypotheses are:

$\mathsf{H}_0 : p_{FP} = 0.1, \qquad p_{NFP} = 0.9$ H_a : One proportion is different

where p_{FP} is the proportion of people who had food poisoning and p_{NFP} is the proportion who did not have food poisoning. The expected count assuming the null hypothesis is true is $n * p_{FP} = 998 * 0.1 = 99.8$ and $n * p_{NFP} = 998 * 0.9 = 898.2$, respectively. The expected count is larger than 5, so we can proceed with the chi-square test. The observed count is 447 and 551 respectively. So, the test statistics can be constructed as

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(447-99.8)^2}{99.8} + \frac{(551-898.2)^2}{898.2} = 1342.105$$

```
(447 - 99.8)^2/99.8 + (551 - 898.2)^2/898.2
```

```
[1] 1342.105
```

The degrees of freedom corresponding to this test is 1. So, the p-value can be calculated to be 0 as:

```
1 - pchisq(1342.105, df = 1)
```

```
[1] 0
```

We can also do the test in R using the `chisq.test` function.

```
chisq.test(x = c(447, 551), p = c(0.1, 0.9))
```

```
    Chi-squared test for given probabilities

data:  c(447, 551)
X-squared = 1342.1, df = 1, p-value < 2.2e-16
```

We reject the null hypothesis ($\chi^2 = 1342.105, df = 1, p - value \approx 0$). There is a significant evidence that the proportion of individuals who eat beef curry and get sick is not 0.1

28.2 Example 2: Candy flavors

We have bags of candy with five flavors in each bag. We collect a random sample of ten bags. Each bag has 100 pieces of candy and five flavors. Use Chi-square goodness of fit test to test if the proportions of the five flavors in each bag are the same. The data table below shows the combined flavor counts from all 10 bags of candy. Fill in the details below:

Click for answer

Flavor	Observed Count (O)	Expected Count (E)	$O - E$	$(O - E)^2$	$(O - E)^2/E$
Apple	180	200	-20	400	2
Lime	250	200	50	2500	12.5
Cherry	120	200	-80	2500	32
Orange	225	200	25	625	3.125
Grape	225	200	25	625	3.125

Answer:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$$
$$H_a : \text{at least one } p_i \text{ not equal to } 1/5$$

```
1 - pchisq(52.75, df = 5-1)
```

```
[1] 9.612522e-11
```

The observed test statistics is:

$$\begin{aligned}\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} &= \frac{(180-200)^2}{200} + \frac{(250-200)^2}{200} \\ &+ \frac{(120-200)^2}{200} + \frac{(225-200)^2}{200} + \frac{(225-200)^2}{200} \\ &= 52.75\end{aligned}$$

```
chisq.test(x = c(180, 250, 120, 225, 225), p = rep(1/5,5))
```

```

    Chi-squared test for given probabilities

data:  c(180, 250, 120, 225, 225)
X-squared = 52.75, df = 4, p-value = 9.613e-11
```

We reject the null hypothesis ($\chi^2 = 52.75, df = 4, p - value \approx 0$). We have significant evidence to claim that at least one proportion of flavors is not the same as others.

Chapter 29

Class Activity 22

29.1 Example 1: Does political comfort level depend on religion?

Consider survey questions about political comfort level and religion. We want know if the response to the comfort level question is associated with their religious practice. To test this question about **two categorical** variables with one variable containing at least **3 levels**, we must conduct a chi-square test for association.

29.1.0.1 (a) Hypotheses

State the hypotheses for this test.

Click for answer

Answer: The null can be stated a couple of equivalent ways: There is no association between religion and comfort level; the variables comfort level and religion are independent of one another; the distribution of comfort level is the same for all three religion types.

The alternatives are just "not the null" statements: There is an association between religion and comfort level; the variables comfort level and religion are dependent; the distribution of comfort level is the different for at least one religion type.

29.1.0.2 (b) Data

Does the data suggest that there is an association between comfort level and religion?

```
library(dplyr)
library(ggplot2)

# read the data
survey <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Survey

# and drop the rows containing missing values using the tidyr package
survey <- survey %>% tidyr::drop_na()

# rename comfort level using fct_recode() from the forcats package
survey <- survey %>% mutate(comfortness = forcats::fct_recode(Question.9,
                          `rarely` = "rarely, if ever, comfortable",
                          `sometimes` = "sometimes comfortable",
                          `almost always` = "almost always comfortable"),
                          comfortness = forcats::fct_relevel(comfortness,
                                                             "almost always",
                                                             "sometimes",
                                                             "rarely"))

# rename comfort level using fct_recode() from the forcats package
survey <- survey %>%mutate(religiousness = forcats::fct_recode(Question.8,
                          `not religious` = "not religious",
                          `religious not active` = "religious but not actively practici
                          `religious active` = "religious and actively practicing my re
                          religiousness = forcats::fct_relevel(religiousness,
                                                               "not religious",
                                                               "religious not active",
                                                               "religious active"))
```

```
# Make a two way table
library(kableExtra)
counts <- table(survey$religiousness, survey$comfortness)
prop.table(counts,1)  # dist of comfort level given religious level
```

```
                     almost always  sometimes     rarely
  not religious          0.53092784 0.39175258 0.07731959
  religious not active   0.39393939 0.41414141 0.19191919
  religious active       0.31578947 0.42105263 0.26315789
```

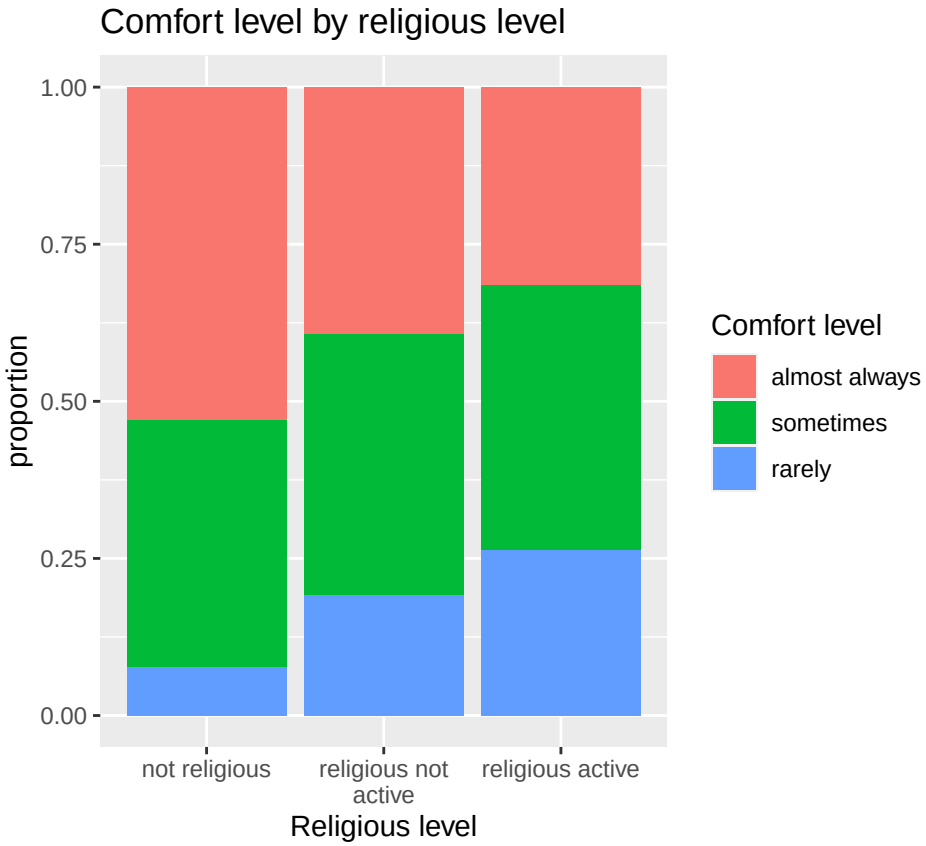
29.1.0.3 Formatted tables in R

Table 29.1: A two way table of religious preference and political comfortness

	almost always	sometimes	rarely
not religious	103	76	15
religious not active	39	41	19
religious active	18	24	15

```
kableExtra::kable(table(survey$religiousness, survey$comfortness),
                  caption = "A two way table of religious preference and political comfortness")
  kable_styling(position = "center")
```

```
ggplot(survey, aes(x=religiousness, fill=comfortness)) +
  geom_bar(position="fill") +
  labs(fill = "Comfort level", x = "Religious level", y = "proportion",
       title="Comfort level by religious level") +
    scale_x_discrete(labels = function(x) stringr::str_wrap(x, width = 16))
```



Click for answer

Answer: Yes, there is a much higher rate of "almost always" comfortable for the not religious respondents (53.1%) than those that are religious (not active: 39.4%; active: 31.6%).

29.1.0.4 (c) Expected counts

Compute the expected number of "not religious" people who are "almost always comfortable".

Click for answer

Answer: There are 194 "not religious" respondents and the overall rate (ignoring religion) of "almost always" comfortable is about 45.7%. If the null is true (and religion doesn't relate to comfort level), the expected number is about

$$194 \times \frac{160}{350} = 88.686$$

```
table(survey$religiousness)
```

```
       not religious religious not active
                 194                   99
    religious active
                  57
```

```
table(survey$comfortness)
```

```
almost always     sometimes        rarely
          160           141            49
```

29.1.0.5 (d) Chi-square contribution

What is the contribution to the chi-square test statistic from the "not religious"/"almost always comfortable" cell?

Click for answer

Answer: The contribution to the chi-square test stat from this category is 2.31.

$$\frac{(103 - 88.6857143)^2}{88.6857143} = 2.31$$

29.1.0.6 (e) Chi-square test

The `chisq.test(x,y)` can be used to give chi-square test results. For this version, `x` and `y` are categorical variables from a data set.

```
ComfortReligion <- chisq.test(survey$religiousness, survey$comfortness)
ComfortReligion
```

```
    Pearson's Chi-squared test

data:  survey$religiousness and survey$comfortness
X-squared = 19.33, df = 4, p-value = 0.0006768
```

- What is the chi-square test stat value?

Click for answer

Answer: The test stat value is 19.33

- How is the degrees of freedom of 4 calculated?

Click for answer

Answer: There are 3 categories for each variable, so the degrees of freedom will be $df = (3-1)(3-1) = 4$.

- Interpret the p-value for this test.

Click for answer

Answer: If there is no association between comfort level and religiousness, then we would see a chi-square test stat of 19.33, or one even larger, only about 0.07% of the time.

29.1.0.7 (f) Conclusion

What is your conclusion for this test?

Click for answer

Answer: We have strong evidence that there is an association between political comfort level and religiousness ($\chi^2 = 19.33$, df $= 4$, p-value $= 7 \times 10^{-4}$).

29.1.0.8 (g) Expected counts

Are the expected counts large enough to use the chi-square distribution to compute the p-value?

```
ComfortReligion$expected
```

```
                      survey$comfortness
survey$religiousness   almost always sometimes rarely
  not religious             88.68571  78.15429  27.16
  religious not active      45.25714  39.88286  13.86
  religious active          26.05714  22.96286   7.98
```

Click for answer

Answer: Yes, all expected counts are 5 or greater.

29.1.0.9 (h) Simulated p-value

If you were concerned that the expected counts weren't large enough to trust using a chi-square distribution to compute a p-value, you can add a `simulate.p.value = TRUE` argument to use a randomization distribution to compute the p-value:

```
chisq.test(survey$religiousness, survey$comfortness, simulate.p.value = TRUE)
```

```
    Pearson's Chi-squared test with simulated p-value
    (based on 2000 replicates)

data:  survey$religiousness and survey$comfortness
X-squared = 19.33, df = NA, p-value = 0.0004998
```

The p-value is slightly different, but your conclusion should be the same.

29.1.0.10 (i) Where is the difference?

Use the grouped bar graph and conditional percents from part (b) to describe the association you (should have) found in part (f). To help quantify differences, compute a 95% confidence interval for the difference in the true proportions of "rarely comfortable" people in the not religious and actively religious groups.

Click for answer

Answer: The largest test stat contributions comes from the not religious/rarely comfortable group and the active religious/rarely comfortable group. We can see that the not religious respondents have a low "rarely" comfortable level compared to religious groups (7.7% vs. 26.3% for active and 19.2% for not active) and they have a very high almost always comfortable level compared to religious groups (53.1% vs. 31.6% for active and 39.4% for not active).

If $p_{not.rel}$ and $p_{active.rel}$ denote the true proportions of "rarely comfortable" for the not religious and active religious groups. We want a 95% CI for $p_{not.rel} - p_{active}$. The sample proportions are computed from the `counts` table (or the `prop.table` output). Of the 194 "not religious" respondents, 15 are rarely comfortable so

$$\hat{p}_{not.rel} = \frac{15}{194} = 0.077$$

$$\hat{p}_{active.rel} = \frac{15}{57} = 0.263$$

So a 95% CI for the difference in the true rates of rarely comfortable is

$$\begin{aligned} CI &= (0.077 - 0.263) \pm 1.96 \cdot \sqrt{\frac{0.077(1-0.077)}{194} + \frac{0.263(1-0.263)}{57}} \\ &= (-0.306, -0.066) \end{aligned}$$

```
round((0.077 - 0.263) + c(-1,1)* 1.96* sqrt(0.077*(1-0.077)/194  + 0.263*(1-0.263)/57)
```

```
[1] -0.306 -0.066
```

- I am 95% confident that the percentage of all non-religious students who are rarely comfortable is between 7 and 31 percentage points lower than the actively religious students.

29.2 Example 2: Perry Preschool Project

In a 1962 social experiment, 123 3- and 4-year-old children from poverty-level families in Ypsilanti, Michigan, were randomly assigned either to a treatment group receiving 2 years of preschool instruction or to a control group receiving no preschool. The participants were followed into their adult years. The following table shows how many in each group were arrested for some crime by the time they were 19 years old. (*Time*, July 29, 1991).

	Arrested	Not Arrested	Total
Preschool	19	42	61
Control	32	30	62
Total	51	72	123

Is a statistically significant difference between the rate of arrest (or no arrest) in the two treatment groups.

29.2.0.1 (a) Test choice

There are two categorical variables, each with two levels. We could either use a two sample test to compare proportions (groups: treatment, response: arrest outcome) OR we could use a chi-square test of independence. These tests will give identical results. For this example, we will use the chi-square test. State your hypotheses needed to test the question above.

Click for answer

Answer: The null hypothesis is that the treatment (preschool/control) is not related to the arrest outcome.

29.2.0.2 (b) Chi-square test with summarized data

This example differs from example 2 because we have data in a summarized two-way table. (Example 2 had the raw categorical variables available.) To run the chi-square test, we first must create a matrix of counts using the `cbind` command that **binds** together **c**olumns of counts:

```
counts <- cbind(c(19,32), c(42,30))
colnames(counts) <- c("arrested", "not arrested") # adds column names
rownames(counts) <- c("preschool", "control")  # adds row names
counts
```

```
          arrested not arrested
preschool       19           42
control         32           30
```

We then use this in the `chisq.test` command:

```
preschool.test <- chisq.test(counts)
preschool.test
```

```
    Pearson's Chi-squared test with Yates' continuity
    correction

data:  counts
X-squared = 4.4963, df = 1, p-value = 0.03397
```

- Are the expected counts large enough to trust these results?

Click for answer

Answer: Yes, they are all above 25.

```
51/123  # overall arrest rate
```

```
[1] 0.4146341
```

```
72/123  # overall non arrest rate
```

```
[1] 0.5853659
```

```
preschool.test$expected
```

```
          arrested not arrested
preschool 25.29268     35.70732
control   25.70732     36.29268
```

- What is your conclusion?

Click for answer

Answer: There is some evidence of an association between the treatment (preschool/control) and the arrest outcome ($\chi^2 = 4.50$, df=1, p-value=0.034).

29.2.0.3 (c) How different?

How do the arrest rates differ for each treatment group? Compute a 95% confidence interval for the difference in arrest rates between those who had the preschool and control treatments.

```
prop.table(counts,1)
```

```
           arrested not arrested
preschool 0.3114754    0.6885246
control   0.5161290    0.4838710
```

Click for answer

Answer: About 52% of the control group were arrested while only about 31% of the preschool group were arrested.

$$\hat{p}_{control} = \frac{32}{62} = 0.516129, \quad \hat{p}_{preschool} = \frac{19}{61} = 0.3114754$$

The 95% for the difference in true arrest rates $p_{control} - p_{preschool}$ is

$$0.516129 - 0.3114754 \pm 1.96\sqrt{\frac{0.516129(1-0.516129)}{62} + \frac{0.3114754(1-0.3114754)}{61}}$$

$$0.2046536 \pm 1.96(0.0868549)$$

$$(0.034418, 0.3748892)$$

We are 95% confident that the true rate of arrest for the preschool treatment is 3.4 to 37.5 percentage points lower than the arrest rate for the control group. This is evidence that the preschool treatment lowered the risk of arrest.

29.2.0.4 Comment

The `chisq.test` command uses a test stat "correction" when both of your categorical variables have only 2 levels. With this correction, your chi-square test results won't exactly match a two-sample test for the difference of two proportions. If you turn off the correct with `correct=FALSE` you will obtain identical results.

```
chisq.test(counts, correct=FALSE) # exact same as two-sample proportion test
```

```
    Pearson's Chi-squared test

data:  counts
X-squared = 5.3059, df = 1, p-value = 0.02125
```

29.3 (Optional) Example 3: College graduates and exercise

A survey of college graduates was done to study how frequently they exercised. The survey was completed by 470 graduates. They were asked where they lived their senior year. Use the following data to determine whether there is an association between exercise on campus and students' living arrangements.

	No regular exercise	Sporadic exercise	Regular exercise	Total
Dormitory	32	30	28	90
On-Campus Apartment	74	64	42	180
Off-campus Apartment	110	25	15	150
At Home	39	6	5	50
Total	255	125	90	470

```
counts3 <- cbind(c(32, 74, 110, 39), c(30,64,25,6), c(28,42,15,5))
colnames(counts3) <- c("No regular exercise", "Sporadic exercise", "Regular exercise")
rownames(counts3) <- c("Dormitory", "On-Campus Apartment", "Off-campus Apartment", "At Home")
```

```
knitr::kable(counts3)
```

	No regular exercise	Sporadic exercise	Regular exercise
Dormitory	32	30	28
On-Campus Apartment	74	64	42
Off-campus Apartment	110	25	15
At Home	39	6	5

```
test3 <- chisq.test(counts3)
```

Click for answer

Answer:

29.3.1 Step 1:

H_0 : exercise and living arrangements independent of each other
H_A : exercise and living arrangements dependent of each other

29.3.2 Step 2:

The observed and expected values from the chi square test are:

```
test3$observed
```

```
                    No regular exercise Sporadic exercise
Dormitory                           32                30
On-Campus Apartment                 74                64
Off-campus Apartment               110                25
At Home                             39                 6
                    Regular exercise
Dormitory                         28
On-Campus Apartment               42
Off-campus Apartment              15
At Home                            5
```

```
round(test3$expected,2)
```

```
                    No regular exercise Sporadic exercise
Dormitory                         48.83             23.94
On-Campus Apartment               97.66             47.87
Off-campus Apartment              81.38             39.89
At Home                           27.13             13.30
                    Regular exercise
Dormitory                      17.23
On-Campus Apartment            34.47
Off-campus Apartment           28.72
At Home                         9.57
```

All of the expected counts are greater than 5.

29.3.3 Step 3:

The test statistics is calculated as:

$$
\begin{aligned}
\chi^2 &= \sum \frac{(O-E)^2}{E} \\
&= \frac{(32-48.83)^2}{48.83} + \frac{(30-23.94)^2}{23.94} + \frac{(28-17.23)^2}{17.23} + \\
&\quad \frac{(74-97.66)^2}{97.66} + \frac{(64-47.87)^2}{47.87} + \frac{(42-34.47)^2}{34.47} + \\
&\quad \frac{(110-81.38)^2}{81.38} + \frac{(25-39.89)^2}{39.89} + \frac{(15-28.72)^2}{28.72} + \\
&\quad \frac{(39-27.13)^2}{27.13} + \frac{(6-13.30)^2}{13.30} + \frac{(5-9.57)^2}{9.57} \\
&= 5.80 + 1.53 + 6.73 + 5.73 + 5.44 + 1.64 + 10.06 + 5.56 + 6.55 + 5.19 + 4.01 + 2.18 \\
&= 60.42
\end{aligned}
$$

```
(32 - 48.83)^2/48.83 + (30 - 23.94)^2/23.94 + (28 - 17.23)^2/17.23 + (74 - 97.66)^2/97.66 + (64 -
```

```
[1] 60.43885
```

```
5.80 + 1.53 + 6.73 + 5.73 + 5.44 + 1.64 + 10.06 + 5.56 + 6.55 + 5.19 + 4.01 + 2.18
```

```
[1] 60.42
```

The degree of freedom of χ^2 is $df = (4-1) * (3-1) = 6$.

```
test3
```

```
    Pearson's Chi-squared test

data:  counts3
X-squared = 60.439, df = 6, p-value = 3.664e-11
```

29.3.4 Step 4:

The p-value can also be calculated as

```
1 - pchisq(60.43, df = 6)
```

```
[1] 3.680733e-11
```

29.3.5 Step 5:

There is significant evidence of an association between the exercise and living arrangements ($\chi^2 = 60.43$, df=6, p-value ≈ 0).

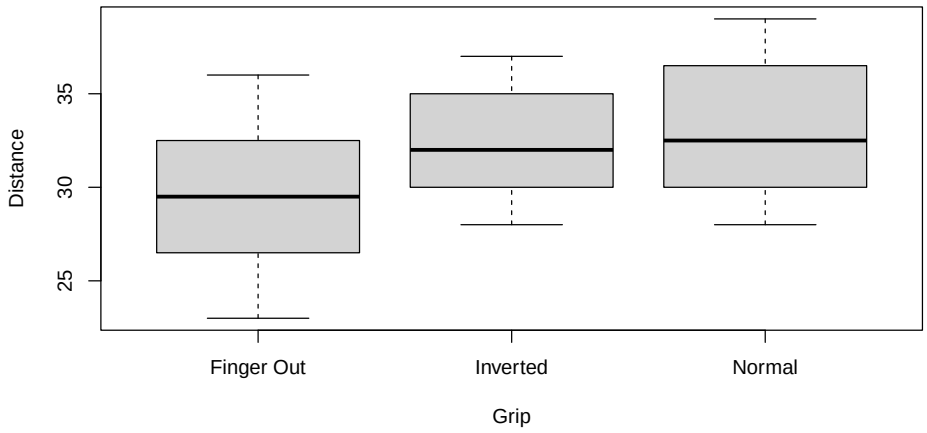
Chapter 30

Class Activity 23

30.1 Example 1: Frisbee grip

The data set `Frisbee.csv` contains data on `Distance` thrown (in paces) for three different frisbee `Grip` types. There are 24 difference cases (throws) Here we can compare responses to this question by the religiousness of the respondent:

```
frisbee <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Frisbee.csv")
boxplot(Distance ~ Grip, data = frisbee)
```




```
tapply(frisbee$Distance, frisbee$Grip, summary)
```

```
$`Finger Out`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  23.00   26.75   29.50   29.50   32.25   36.00

$Inverted
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.00   30.00   32.00   32.38   34.50   37.00

$Normal
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.00   30.00   32.50   33.12   36.25   39.00
```

The question we want to answer is whether or not the differences in observed mean distance thrown are statistically significant. To test this question comparing **means** for a **quantitative** response broken up into **at least 2 groups**, we can conduct a **one-way ANOVA test**.

30.1.0.1 (a) One-way ANOVA hypotheses

State the hypotheses for this test.

Click for answer

Answer: Let μ be the true mean distance thrown using a certain grip. Then $H_0 : \mu_{fout} = \mu_{invert} = \mu_{normal}$ vs. H_A : at least one mean is different.

30.1.0.2 (b) One-way ANOVA test

You can obtain the one-way ANOVA table and test results with the `aov(y ~ x, data=)` command. Running the `summary` function on this anova result gives you the ANOVA table:

```
frisbee.anova <- aov(Distance ~ Grip, data = frisbee)
summary(frisbee.anova)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Grip         2  58.58   29.29   2.045  0.154
Residuals   21 300.75   14.32
```

- What is the F test stat value?

Click for answer

Answer: $F = 2.045$

- Interpret the p-value.

Click for answer

Answer: If grip does not affect distance thrown, then we would see mean differences as larger, or larger, than those observed about 15.4% of the time.

- What is your conclusion?

Click for answer

Answer: This study does not provide evidence that these three grips affect the mean distance thrown.

30.1.0.3 (c) Checking assumptions

Can we trust the p-value obtained above using the F distribution?

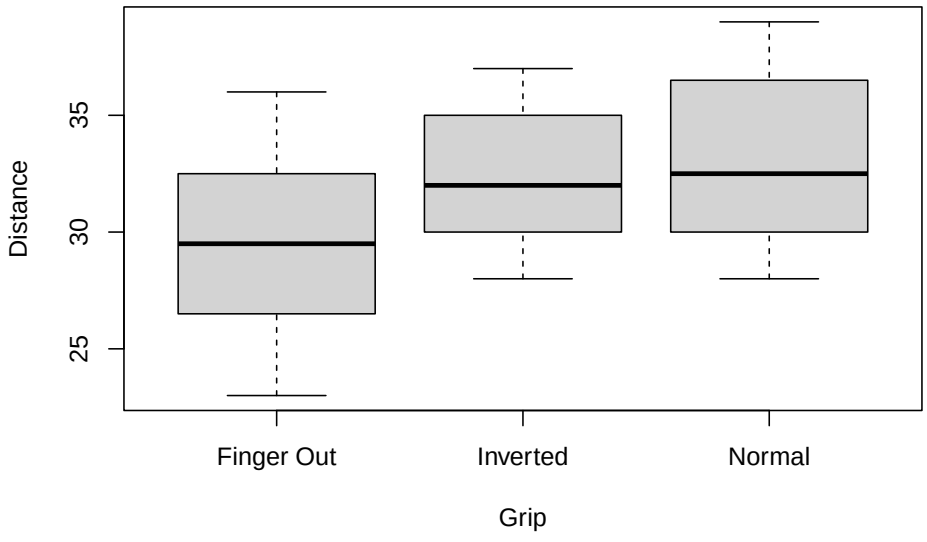
```
table(frisbee$Grip)  # check n's
```

```
Finger Out   Inverted     Normal
         8          8          8
```

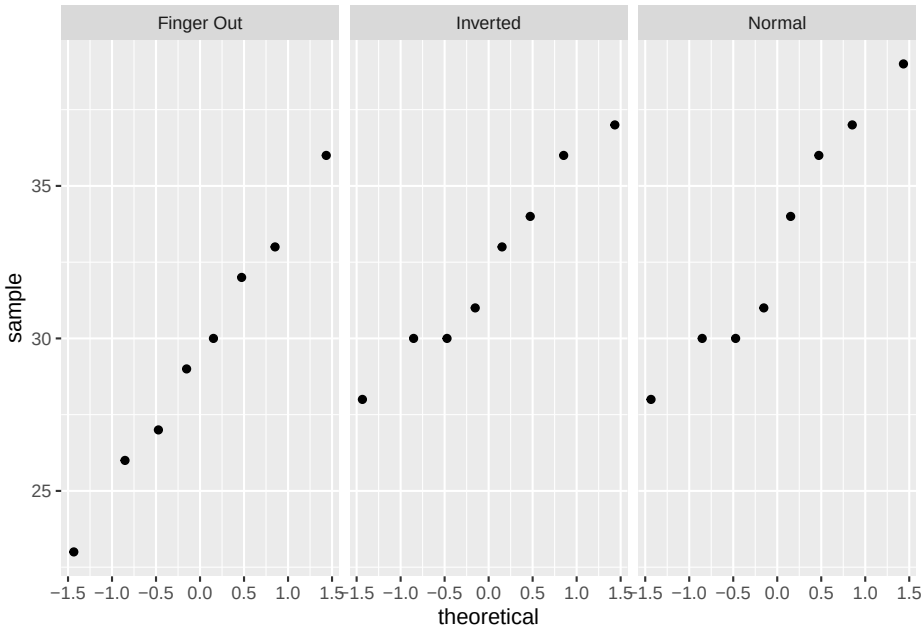
```
tapply(frisbee$Distance, frisbee$Grip, sd)  # similar SD's?
```

```
Finger Out   Inverted     Normal
  4.174754   3.159453   3.943802
```

```
library(ggplot2)  # shape?
boxplot(Distance ~ Grip, data = frisbee)
```



```
ggplot(frisbee, aes(sample = Distance)) + geom_qq() + facet_wrap(~Grip)
```



Click for answer

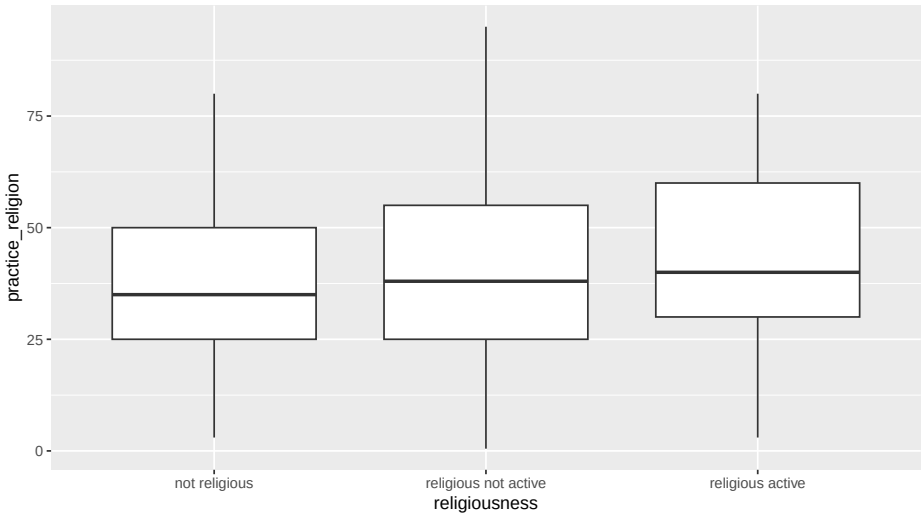
Answer: Sample sizes in all three groups are small (8) but the observed distances thrown within each group are roughly normally distributed. There are small differences in variation of the three groups, but the SD rule is met since largest SD (4.17) is less than twice the smallest SD (3.16). The assumptions are met.

30.2 Example 2: Comparing % religious guess by religion

One of the class survey questions asked respondents to give their best guess at the percentage of students at Carleton who practice a religion. Here we can compare responses to this question by the religiousness of the respondent:

```
library(dplyr)
# read the data
survey <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Survey

# and drop the rows containing missing values using the tidyr package
survey <- survey %>% tidyr::drop_na()
```

```
# make a new variable called `practice_religion_percentage` (more informative variable name)
survey <- survey %>%  mutate(practice_religion = Question.7)

# rename comfort level using fct_recode() from the forcats package
survey <- survey %>%mutate(religiousness = forcats::fct_recode(Question.8,
                           `not religious` = "not religious",
                           `religious not active` = "religious but not actively practicing",
                           `religious active` = "religious and actively practicing my religion"),
                           religiousness = forcats::fct_relevel(religiousness,
                                                                "not religious",
                                                                "religious not active",
                                                                "religious active"))
ggplot(data = survey) +
    geom_boxplot(aes(x = religiousness, y = practice_religion))
```




```
tapply(survey$practice_religion, survey$religiousness, summary)
```

```
$`not religious`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   25.00   35.00   38.05   50.00   80.00

$`religious not active`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.50   25.00   38.00   40.19   55.00   95.00

$`religious active`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   30.00   40.00   41.32   60.00   80.00
```

30.2.0.1 (a) One-way ANOVA hypotheses

We want to determine if the differences in observed mean guesses are statistically significant. State the hypotheses for this test.

Click for answer

Answer: Let μ be the true mean religous % guess in a given religiousness group. Then $H_0 : \mu_{notRelig} = \mu_{Relig,Act} = \mu_{Relig,NotAct}$ vs. H_A : at least one mean is different.

30.2.0.2 (b) Check assumptions

Can use trust the results from a one-way ANOVA test?

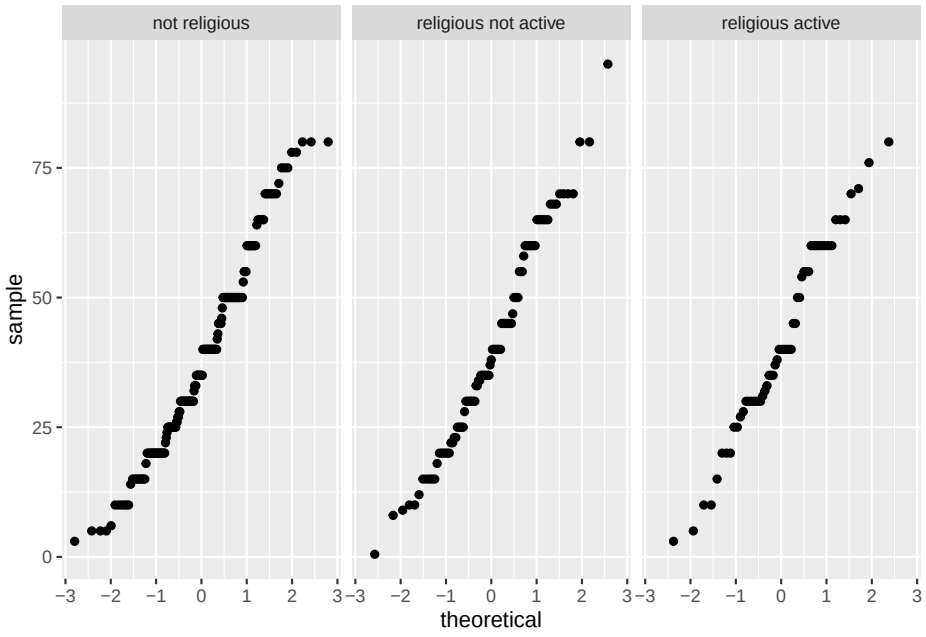
```
table(survey$religiousness)
```

```
       not religious religious not active
                 194                   99
    religious active
                  57
```

```
tapply(survey$practice_religion, survey$religiousness, sd, na.rm=TRUE)  #need na.rm wi
```

```
       not religious religious not active
            17.96535             19.22239
    religious active
            18.33143
```

```
ggplot(survey, aes(sample = practice_religion)) + geom_qq() + facet_wrap(~religiousness)
```



Click for answer

Answer: Yes, the assumptions are met. The distributions within each group are slightly skewed or roughly symmetric, and the sample sizes within each group are all at least 30. In addition, the SD in each group are close to each other (18% to 19.2%).

30.2.0.3 (c) One-way ANOVA test

Assuming part (b) checks out, run the one-way ANOVA test to compare means:

```
guess.aov <- aov(practice_religion ~ religiousness, data = survey)
summary(guess.aov)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
religiousness   2    607   303.6   0.898  0.408
Residuals     347 117321   338.1
```

- What is the F test stat value?

Click for answer

Answer: $F = 0.898$

- Interpret the p-value.

Click for answer

Answer: If there is no difference in true mean guess in all three groups, we would see an F test stat of at least 0.898 about 40.8% of the time.

- What is your conclusion?

Click for answer

Answer: The differences in mean guesses that we've observed in our sample are not statistically significant. We don't have evidence that the true mean guesses for the three religiousness groups are different.

30.2.0.4 (d) Describe the association?

If you found a statistically significant difference in means in part (c), describe how the groups differ. If you did not find a statistically significant difference in part (c), estimate the average guess for all students in the (hypothetical) population of 215 students.

Click for answer

Answer: We didn't find a statistically significant difference in part (c). So what is our best estimate of the average guess for all students, since responses don't seem to differ by religiousness?

```
t.test(survey$practice_religion)
```

```
    One Sample t-test

data:  survey$practice_religion
t = 39.882, df = 349, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 37.25428 41.11927
sample estimates:
mean of x
 39.18677
```

We are 95% confident that the mean guess at the percentage of religious students at Carleton is between 37.3% to 41.1% for all math 215 students.

What if there was a difference?!

Use EDA to describe how the sample means differ. Does it look like all three means are different, or does one mean look different from the rest? The sample mean responses from the two religious groups look similar (active: 41.3%; not active: 40.2%) but the mean response of the not religious group is lower (38.1%).

Chapter 31

Class Activity 24

31.1 Example 1: Cuckoo Eggs

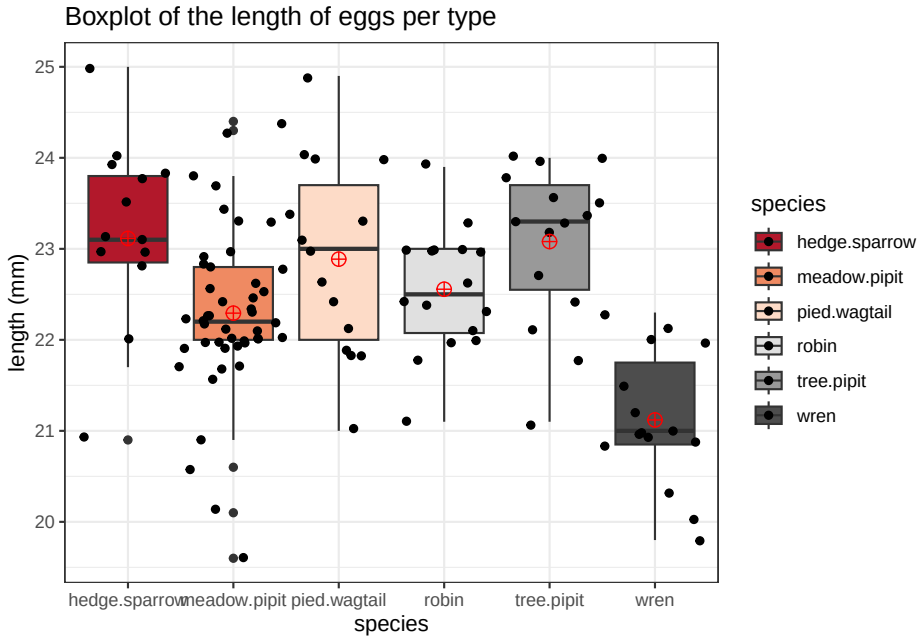
The common cuckoo does not build its own nest: it prefers to lay its eggs in another birds' nest. It is known, since 1892, that the type of cuckoo bird eggs are different between different locations. In a study from 1940, it was shown that cuckoos return to the same nesting area each year, and that they always pick the same bird species to be a "foster parent" for their eggs. Over the years, this has lead to the development of geographically determined subspecies of cuckoos. These subspecies have evolved in such a way that their eggs look as similar as possible as those of their foster parents.

The cuckoo dataset contains information on 120 Cuckoo eggs, obtained from randomly selected "foster" nests. For these eggs, researchers have measured the `length` (in mm) and established the `type` (species) of foster parent. The type column is coded as follows:

- `type=1`: Hedge Sparrow
- `type=2`: Meadow Pit
- `type=3`: Pied Wagtail
- `type=4`: European robin
- `type=5`: Tree Pipit
- `type=6`: Eurasian wren

The researchers want to test if the type of foster parent has an effect on the average length of the cuckoo eggs.

31.1.1 1(a) The boxplot of the length of the eggs across all the species is shown below. Based on these boxplots, do the assumptions of normality and similar variability appear to be met?

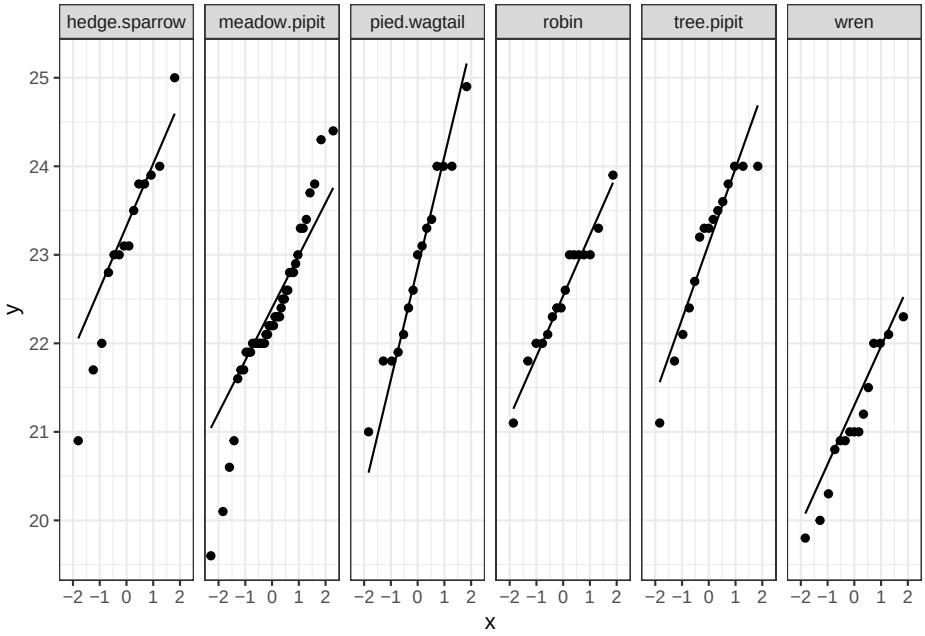


31.1.2 (1b) Formally verify that the assumptions are valid by using the outputs given.

Click for answer

Answer: Based on the qqplot, the data points in each group are close to the line and there are no major deviations towards the center. So, the normality assumption seems to be satisfied.

```
Cuckoo %>%
  ggplot(aes(sample=length)) + geom_qq() + geom_qq_line() + facet_grid(~species) +  the
```



Similarly, based on the statistics below, the ratio of the largest s to the smallest s is 1.57. So, the equal variance assumption is satisfied.

Caution: If the equal variance assumption or the normality assumption is not met in ANOVA, then the results of the one-way ANOVA may not be reliable. This is especially true if the sample sizes between the groups are unequal and the variances between the groups are also unequal.

```
1.0722917/0.6821229
```

```
[1] 1.571992
```

```
library(dplyr)
stat <- Cuckoo %>% group_by(species) %>% summarize(mean(length), sd(length), length(length))
stat <- as.data.frame(stat)
stat
```

```
        species mean(length) sd(length) length(length)
1 hedge.sparrow     23.11429  1.0494373             14
2  meadow.pipit     22.29333  0.9195849             45
3  pied.wagtail     22.88667  1.0722917             15
4         robin     22.55625  0.6821229             16
5    tree.pipit     23.08000  0.8800974             15
6          wren     21.12000  0.7542262             15
```

31.1.3 (1c) Fit an ANOVA model to do a formal hypothesis test. Report the test statistics and conclude your hypothesis test.

```
fit_anova <- aov(length~species, Cuckoo)
summary(fit_anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
species       5  42.81   8.562   10.45 2.85e-08 ***
Residuals   114  93.41   0.819
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Click for answer

Answer: The hypotheses can be stated as:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a : \text{at least one } \mu_i \text{ is different}$$

Let's assume the conditions for the test are approximately met. To find which of the species differ from the rest, we need to construct confidence intervals for the mean length differences between each pair of species.

31.1.4 (1d) First, find a 95% confidence interval for the mean cuckoo egg length in `European robin` nests (Type = 4).

Click for answer

Answer:

95 % confidence interval is:

```
MSE <- 0.8193847
stat[4,2] + c(-1,1)*(qt(1-0.05/2, df=113))*sqrt(MSE)/sqrt(stat[4,4])
```

```
[1] 22.10791 23.00459
```

$$22.556 \pm 1.981 * \frac{\sqrt{0.8194}}{\sqrt{16}}$$
$$= (22.108, 23.005)$$

31.1.5 (1e) Find a 95% CI for the difference in mean egg length between European robin(type = 4) and Eurasian wren(type = 6) nests.

Click for answer

Answer:

```
(stat[4,2] - stat[6,2]) + c(-1,1)* (qt(1-0.05/2, df=113))* sqrt(MSE*(1/stat[4,4] + 1/stat[6,4]))
```

```
[1] 0.79172 2.08078
```

$$(22.556 - 21.120) \pm 1.981 \cdot \sqrt{0.8194\left(\frac{1}{16} + \frac{1}{15}\right)}$$

$= (0.792, 2.081)$

31.1.6 (1f) Find a 95% CI for the difference in mean egg length between Pied Wagtail (type = 3) and European robin(type = 4) nests.

Click for answer

Answer:

```
(stat[3,2] - stat[4,2]) + c(-1,1)* (qt(1-0.05/2, df=113))*sqrt(MSE*(1/stat[3,4] + 1/stat[4,4]))
```

```
[1] -0.3141134  0.9749467
```

$$(22.887 - 22.556) \pm 1.981 \cdot \sqrt{0.8194\left(\frac{1}{15} + \frac{1}{16}\right)}$$
$$= (-0.314, 0.975)$$

31.1.7 (1g) We can use the R function pairwise.t.test to analyze which pair of means are significantly different from one another. Using p.adjust.method = "bonferroni", we will see the p-values adjusted for multiple comparison. These adjusted p-values should still be compared with $\alpha = 0.05$ to find any significant differences.

Based on the R output, which of the pairs are different?

```
pairwise.t.test(Cuckoo$length, Cuckoo$species, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  Cuckoo$length and Cuckoo$species

             hedge.sparrow meadow.pipit pied.wagtail
meadow.pipit 0.05554       -            -
pied.wagtail 1.00000       0.44898      -
robin        1.00000       1.00000      1.00000
tree.pipit   1.00000       0.06426      1.00000
wren         5e-07         0.00045      7e-06
             robin   tree.pipit
meadow.pipit -       -
pied.wagtail -       -
robin        -       -
tree.pipit   1.00000 -
wren         0.00035 5e-07

P value adjustment method: bonferroni
```

Click for answer

Answer:

Based on the adjusted p-values we can say the five pairs of species `6-1`, `6-2`, `6-3`, `6-4`, and `6-5` are different at the significance level of 5%. Here, each pairwise test is testing:

$$H_0 : \mu_i = \mu_j \text{ vs. } H_a : \mu_i \neq \mu_j$$

31.2 (Optional) Example 2: Metal Contamination

An environmental studies student working on an independent research project was investigating metal contamination in a local river. The metals can accumulate in organisms that live in the river (known as bioaccumulation). He collected samples of Quagga mussels at three sites in the river and measured the concentration of copper (in micrograms per gram, or mcg/g) in the mussels. His data are summarized in the provided table and plot. He wants to know if there are any significant differences in mean copper concentration among the three sites.

Site	Mean ($\bar{x}$)	SD (s)	n
1	21.34	3.092	5
2	16.60	2.687	4
3	13.16	4.274	5

31.2.0.1 (a) Assumptions

What do we need to assumption about copper concentrations to use one-way ANOVA to compare means at the three sites?

Click for answer

Answer: With such small sample sizes in each group it would be hard to get a good sense of how they are distributed. We will just need to assume that these measurements are approximately normally distributed.

31.2.0.2 (b) One-way ANOVA hypotheses

State the hypotheses for this test.

Click for answer

Answer: Let μ_i be the true mean copper concentration at location i. Then

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

vs. H_A : at least one mean is different.

31.2.0.3 (c) ANOVA table

Fill in the missing values `A` - `E` from the ANOVA table:

Source	df	SS	MS	F
Groups	`A = 2`	169.05	`C = 84.525`	`E = 6.99`
Error	11	`B = 132.97`	`D = 12.088`	
Total	13	302.02		

Click for answer

Answer:

- A: The group degrees of freedom is always the number of groups minus 1. Here we have 3 groups so $A = 3 - 1 = 2$.

- B: The group and error sum of squares adds up to the total sum of squares. So we have $B = 302.02 - 169.05 = 132.97$.
- C: Mean square values are always sum of squares divided by degrees of freedom. For groups MS: $C = 169.05/2 = 84.525$
- D: Mean square values are always sum of squares divided by degrees of freedom. For error MS: $D = 132.97/11 = 12.088$
- The F test stat is the ratio of the group MS and error MS: $F = 84.525/12.088 = 6.992$.

```
302.02 - 169.05
```

```
[1] 132.97
```

```
169.05/2
```

```
[1] 84.525
```

```
132.97/11
```

```
[1] 12.08818
```

```
84.525/12.088
```

```
[1] 6.992472
```

31.2.0.4 (d) p-value

The command `pf(x, df1=, df2=)` gives the area under the F-distribution below the value `x`. Use this command to get the p-value from this one-way ANOVA test. Interpret this value.

Click for answer

Answer: The p-value is about 1.1%. If the means are the same at the three sites, we would see sample means this different, or even more different, about 1.1% of the time.

```
1-pf(6.992, df1=2, df2=11)
```

```
[1] 0.01097789
```

31.2.0.5 (e) Conclusion

What is your conclusion for this test?

Click for answer

Answer: We have some evidence that at least one of the true mean copper concentration at the three sites is differenct from the others.

31.2.0.6 (f) Confidence interval

Compute a 95% confidence interval for the difference in means between site 1 and 3. Interpret this interval.

Click for answer

Answer: Since we don't have the data, we will have to compute the CI by hand. The degrees of freedom "best guess" (since we aren't letting R approximate it), is 11. The 95% CI for the difference in true means in site 1 and 3 is :

$$(21.34 - 13.16) \pm (2.201))\sqrt{12.088\left(\frac{1}{5}+\frac{1}{5}\right)} = 3.34, 13.02$$

```
(21.34 - 13.16) + c(-1,1)* qt(1-0.05/2, df = 11)*sqrt(12.088*(1/5+1/5))
```

```
[1]  3.340234 13.019766
```

We are 95% confident that the true mean copper concentration at site 1 is 3.34 to 13.02 mcg/g higher than the true mean concentration at site 3.

Chapter 32

Class Activity 25

```
# load necessary libraries
library(readr) # read_csv
library(dplyr) # data manipulation
library(forcats) # categorical variables
library(janitor) # clean_names
library(tidyr) # drop_na
library(tidyverse) # functions form tidy ecosystem
```

32.1 Linear Regression Analysis: Exploring the Relationship between Average Mathematics GPA and Length of Study in Mathematics

This class activity aims to guide you through a data analysis project involving a school score dataset. You will learn how to load a dataset, inspect its contents, manipulate and clean the data, perform exploratory analysis, and apply linear regression to identify relationships. Towards the end, you will deal with outliers and visualize data to comprehend the results better.

32.1.1 Step 1: Load and Inspect the Data

```
# Task: Load the school_scores.csv data file from your local directory
school_scores <- read_csv("~/Desktop/Insync/STAT120_Spring23/class_activities/data/school_scores.
```

```
# Task: Use the glimpse() function to view the structure of your data.
# glimpse(school_scores)
```

32.1.2 Step 2: Data Cleaning

```
# Task: Use janitor's clean_names() function to standardize column names,
# and tidyr's drop_na() to remove missing values.
school_scores_clean <- school_scores %>%
  janitor::clean_names() %>%
  tidyr::drop_na()
```

32.1.3 Step 3: Data Manipulation

```
# Task: Select relevant variables from the dataset for further analysis.
school_new <- school_scores_clean %>% select(year,
                                             state_name,
                                             total_math,
                                             total_verbal,
                                             academic_subjects_mathematics_average_gpa
                                             academic_subjects_mathematics_average_yea
```

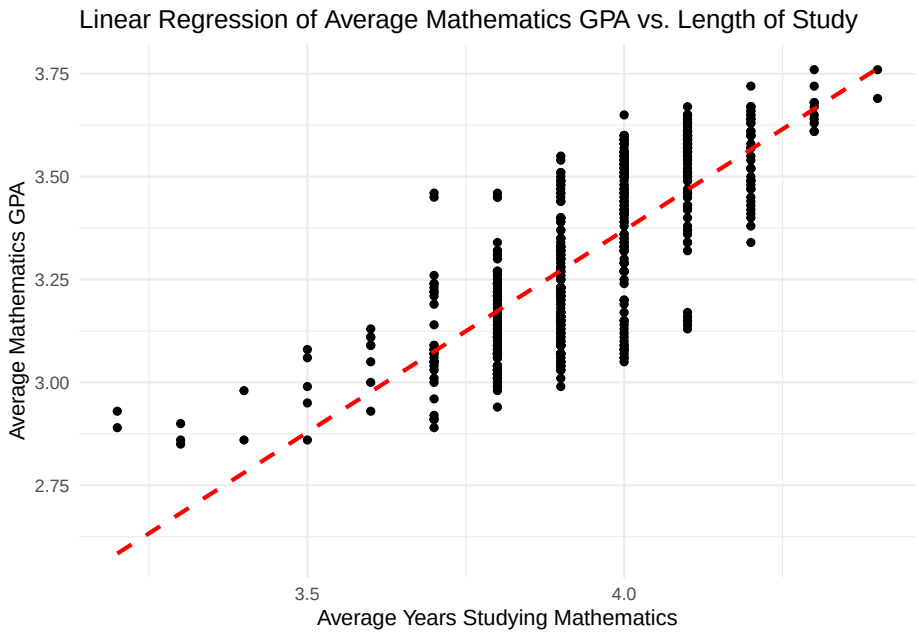
```
# Task: Create a new categorical variable named 'GPA' from
# the 'academic_subjects_mathematics_average_gpa' variable.
school_final <- school_new %>%
  mutate(GPA = case_when(academic_subjects_mathematics_average_gpa < 3 ~ "low",
                         academic_subjects_mathematics_average_gpa >= 3.0 & academic_su
                         academic_subjects_mathematics_average_gpa >= 3.25 & academic_s
                         academic_subjects_mathematics_average_gpa >= 3.5 ~ "excellent
```

```
# Task: Convert the new GPA variable into a factor variable.
school_final <- school_final %>% mutate(GPA = factor(GPA))
```

```
# Task: Collapse the GPA variable levels into two broad categories.
school_final <- school_final %>%
  mutate(collapsed_GPA = forcats::fct_collapse(GPA,
                                               Low = c("low", "satisfactory"),
                                               High = c("good", "excellent")))
```

32.1.4 Step 4: Data Analysis - Linear Regression

```
# Fit a linear model
ggplot(data = school_final, aes(x = academic_subjects_mathematics_average_years,
                                y = academic_subjects_mathematics_average_gpa)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  labs(x = "Average Years Studying Mathematics", y = "Average Mathematics GPA",
       title = "Linear Regression of Average Mathematics GPA vs. Length of Study") +
  theme_minimal()
```

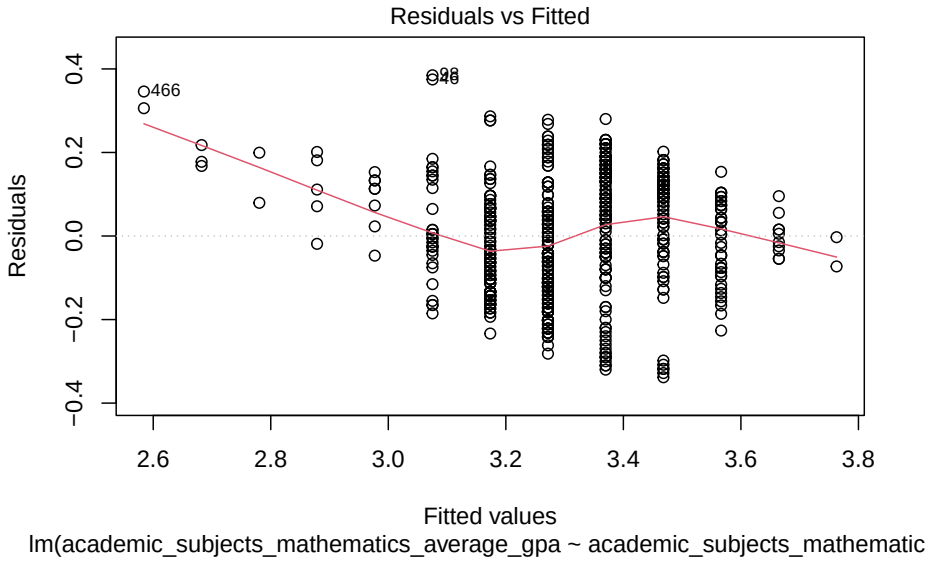



```
# Task: Conduct a linear regression analysis with GPA as the response
# variable and average years studying mathematics as the predictor.
GPA.lm <- lm(academic_subjects_mathematics_average_gpa ~ academic_subjects_mathematics_average_ye
             data = school_final)
summary(GPA.lm)
```

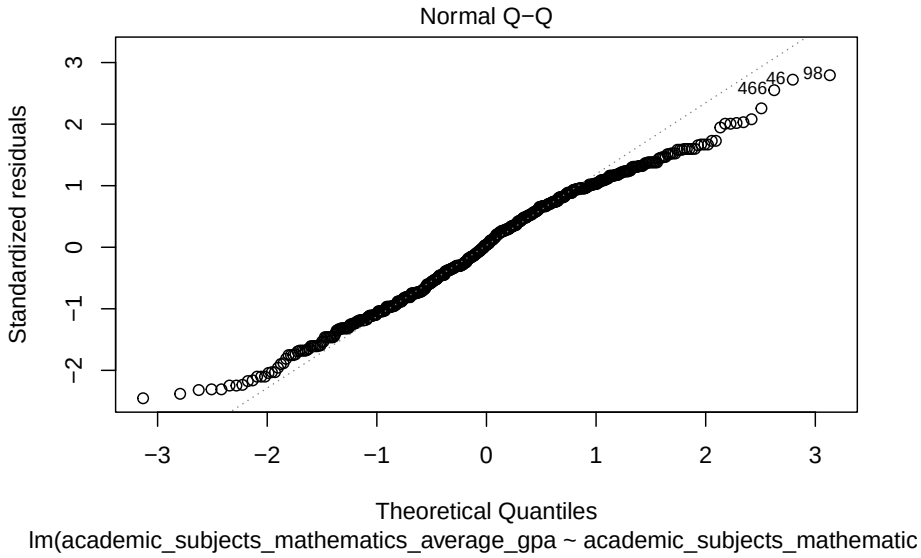
```

Call:
lm(formula = academic_subjects_mathematics_average_gpa ~ academic_subjects_mathematics_average_ye
    data = school_final)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.33812 -0.10344  0.00478  0.11188  0.38478

Coefficients:
                                            Estimate
(Intercept)                                  -0.5592
academic_subjects_mathematics_average_years   0.9823
                                            Std. Error
(Intercept)                                     0.1348
academic_subjects_mathematics_average_years     0.0342
                                            t value
(Intercept)                                  -4.147
academic_subjects_mathematics_average_years  28.725
                                            Pr(>|t|)
(Intercept)                                 3.87e-05 ***
academic_subjects_mathematics_average_years  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.138 on 575 degrees of freedom
Multiple R-squared:  0.5893,    Adjusted R-squared:  0.5886
F-statistic: 825.1 on 1 and 575 DF,  p-value: < 2.2e-16
```

```r
# Task: Generate a Residual plot to assess the regression model's assumptions.
plot(GPA.lm, which = 1)
```



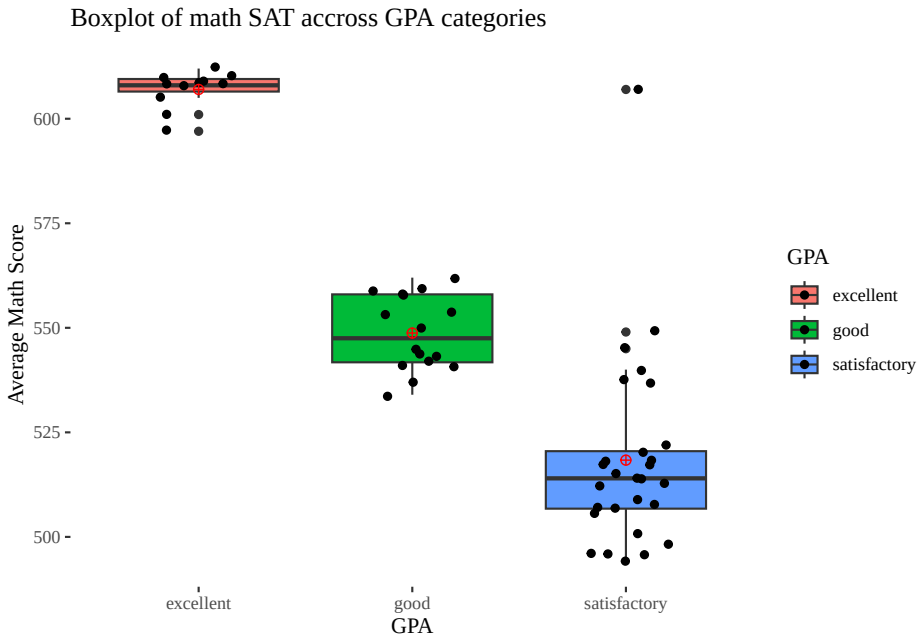
```
# Task: Generate a QQ plot to visualize the normality of residuals.
plot(GPA.lm, which = 2)
```



32.1.5 Step 5: Handling Outliers

```
# Task: Filter the data to only include certain states.
school_selected <- school_final %>%
  filter(state_name %in% c("Alabama", "California", "Montana", "Minnesota", "Nevada"))
```

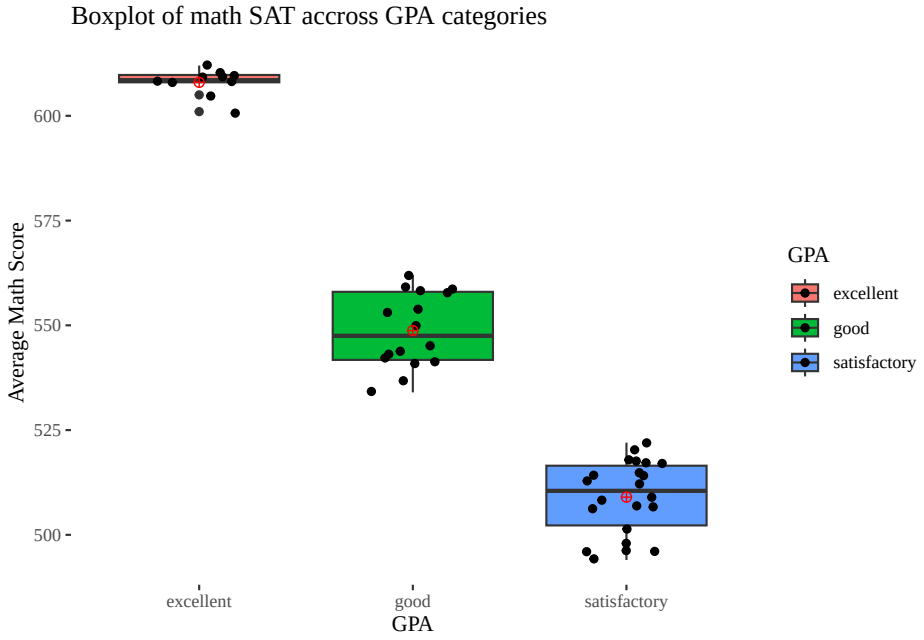
```
# Task: Visualize math SAT scores across GPA categories for the selected
# states using a boxplot, and identify potential outliers.
school_selected %>%
  ggplot(aes(x=GPA,y=total_math,fill=GPA)) +
  theme_bw() +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  labs(title ="Boxplot of math SAT accross GPA categories",
       y = "Average Math Score",
       x = "GPA") +
  stat_summary(fun=mean, geom="point", shape=10,
               size=2, color="red", fill="black") +
  ggthemes::theme_tufte()
```




```
# Task: Remove outliers based on pre-determined conditions.
school_selected_no_outlier <-  school_selected %>%
  filter(GPA == "excellent" & total_math >= 600 &  total_math <= 800 |
           GPA == "good" & total_math >= 525  &  total_math <= 575 |
           GPA == "satisfactory" & total_math >= 425  &  total_math <= 525)
```

32.1.6 Step 6: Visualizing Cleaned Data

```
# Task: Re-visualize the boxplot of math SAT scores across
# GPA categories after removing the outliers.
school_selected_no_outlier %>%
  ggplot(aes(x=GPA,y=total_math,fill=GPA)) +
  theme_bw() +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  labs(title ="Boxplot of math SAT accross GPA categories",
       y = "Average Math Score",
       x = "GPA") +
  stat_summary(fun=mean, geom="point", shape=10,
               size=2, color="red", fill="black") +
  ggthemes::theme_tufte()
```



By the end of this activity, you will have practiced various data analysis techniques, including data manipulation, data cleaning, linear regression, and outlier handling.