Stat 120

Deepak Bastola

2023-02-26

Contents

About

This is a *sample* book written in **Markdown** to guide STAT 120 students interactively explore various class activities and projects in R.

Chapter 1

Class Activity 1

1.1 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),
                          Age = runif(n=10, 20, 60))
```

Click for answer

```
example_data
```

```
   ID Greeting  Male      Age
1   1    Hello  TRUE 39.02883
2   2    Hello FALSE 42.54085
3   3    Hello  TRUE 45.13284
4   4    Hello FALSE 37.52349
5   5    Hello  TRUE 58.24304
6   6  Goodbye FALSE 46.34264
7   7  Goodbye  TRUE 28.33329
8   8  Goodbye FALSE 31.15743
9   9  Goodbye  TRUE 32.76741
10 10  Goodbye FALSE 58.03955
```

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 age 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.

1.2 Your Turn 2

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

Click for answer

```
# read in the data
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)
```

```
              Country EducationExpenditure Literacy
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     99.0
```

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.

1.3 Quiz

1. Cases are a set of individual units where the measurements are taken.

A. TRUE

B. FALSE

Click for answer

TRUE

2. The characteristic that is recorded for each case is called a

A. ledger

B. caseholder

C. placeholder

D. variable

Click for answer

variable

3. Variables can be either categorical or quantitative.

A. TRUE

B. FALSE

Click for answer

TRUE

Chapter 2

Class Activity 2

2.1 Your Turn 1

This exercise is about finding the average word length in Lincoln's Gettysburg's address.

2.2 Your Turn 2

2.2.1 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.

2.3 Your Turn 3

2.3.1 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]  51 204  99 149  92 180 213 241 214 129
```

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
51        51    9  dedicated,
```

```
204      204    4        from
99        99    2          we
149      149    2          we
92        92    2          It
180      180    4        here
213      213    4        that
241      241    6     nation,
214      214    5       cause
129      129   11 consecrated
```

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]  9  4  2  2  2  4  4  6  5 11
```

Then find the mean of these values:

```
mean(mysize)
```

```
[1] 4.9
```

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.

2.3.2 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).

2.4 Quiz

1. A group of researchers investigated the effect of media usage (whether or not subjects watch television or use the Internet) in the bedroom on "Tiredness" during the day (measured on a 50 point scale). The explanatory and response variables are

A. Explanatory is media usage in the bedroom and response is "tiredness"

B. Explanatory is "tiredness" and response is media usage in the bedroom

Click for answer

The correct answer is A.

2. An October 2016 Gallup poll estimates that 60% of US adults support legalizing the use of marijuana. Their results were based on a "random sample of 1,017 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia". The population for this study is

A. all adults (18 and older) living in the U.S. (including D.C)

B. the 1,017 adults (18 and older) living in the U.S. (including D.C) who were sampled

C. the 1,017 adults (18 and older) living in the U.S. (including D.C) who were sampled and support legalizing marijuana

D. all adults (18 and older) living in the U.S. (including D.C) who support legalizing marijuana

Click for answer

The correct answer is A.

3. An October 2016 Gallup poll estimates that 60% of US adults support legalizing the use of marijuana. Their results were based on a "random sample of 1,017 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia". Which statement below regarding bias is true?

A. The results are biased because Gallup only contacted a small fraction of people in the population.

B. The results may be biased because people may not have answered a survey question about marijuana truthfully

Click for answer

The correct answer is B.

Chapter 3

Class Activity 3

3.1 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, Level of Depression (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

3.2 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.

(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

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

3.3 Quiz

1. A third variable that is associated with both the explanatory variable and the response variable is called a confounding variable.

A. TRUE

B. FALSE

Click for answer

TRUE

2. The different levels of an explanatory variable are known as

A. treatments

B. local groups

C. response

D. cases

Click for answer

treatments

3. Causality can always be inferred from observational studies.

A. TRUE

B. FALSE

Click for answer

FALSE

Chapter 4

Class Activity 4

4.1 Your Turn 1

4.1.1 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?

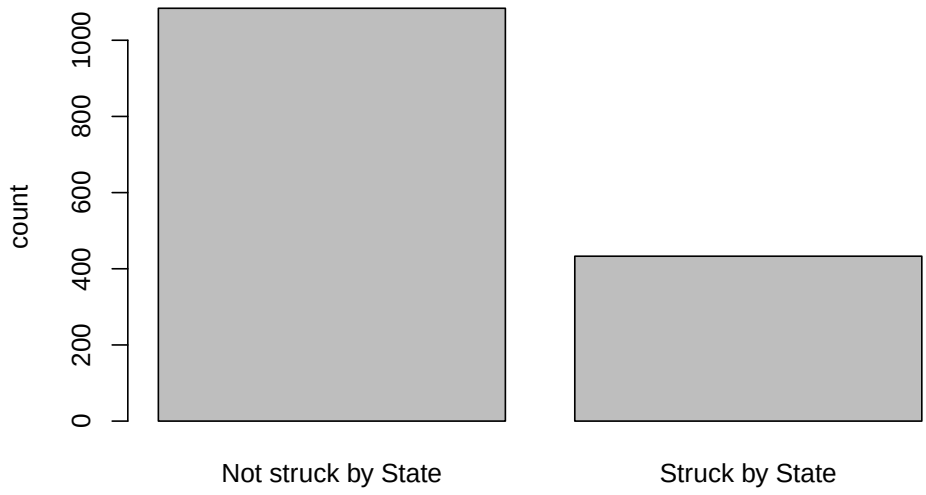
Click for answer

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

(c). Bar graph for one variable

You can create a simple bar graph for one categorical variable with the `barplot` command. Here we visualize the distribution of struck status for all eligible jurors:

```
barplot(counts, ylab = "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

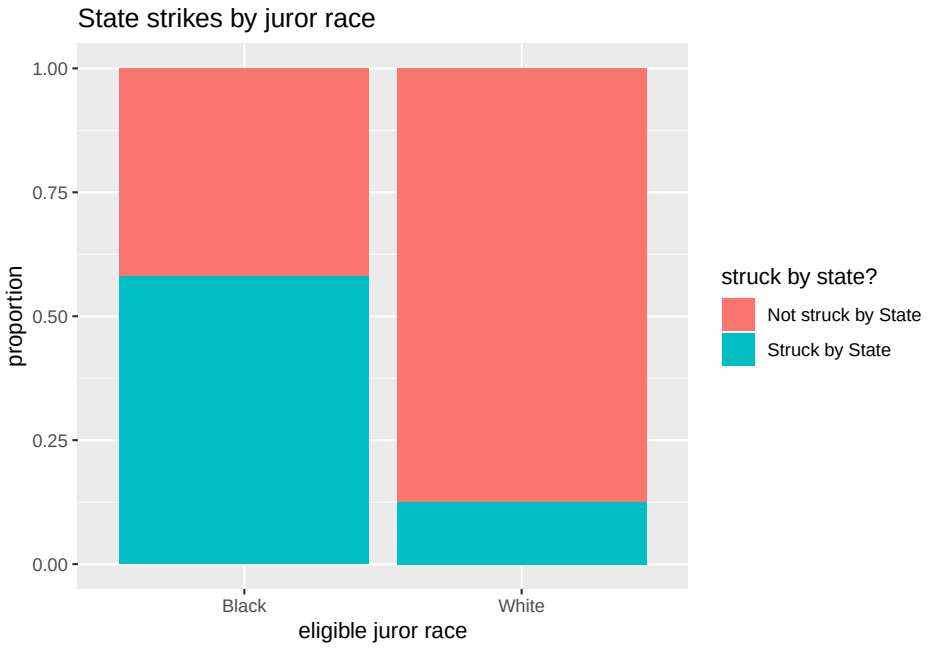
(f). Stacked bar graph for two variables

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:

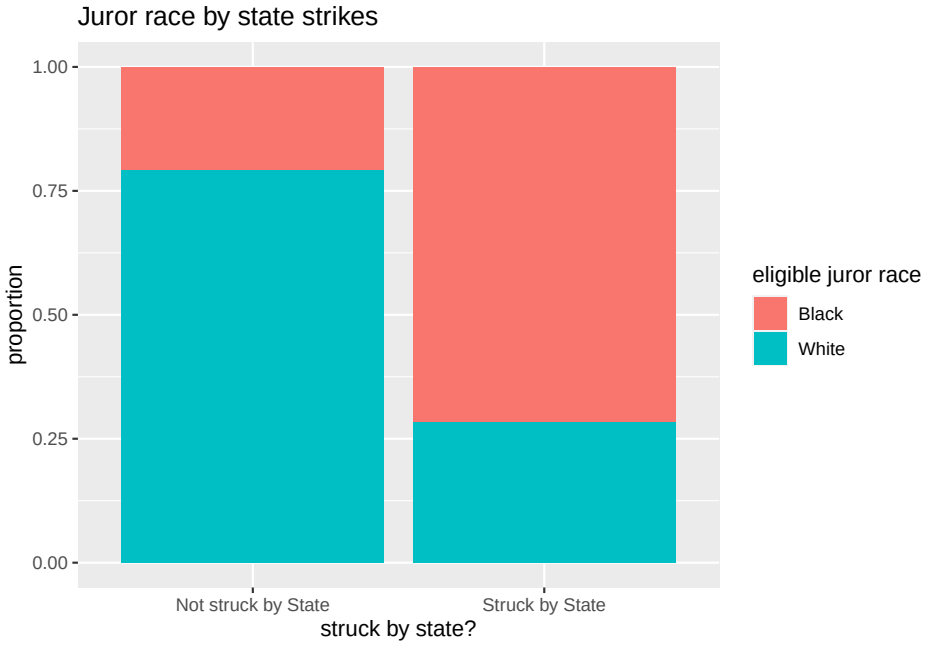
```
prop.table(mytable, margin = 2)
```

```
       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

---

4.2 Your Turn 2

4.2.1 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/BerkeleyGrad.csv")
```

```
# 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?

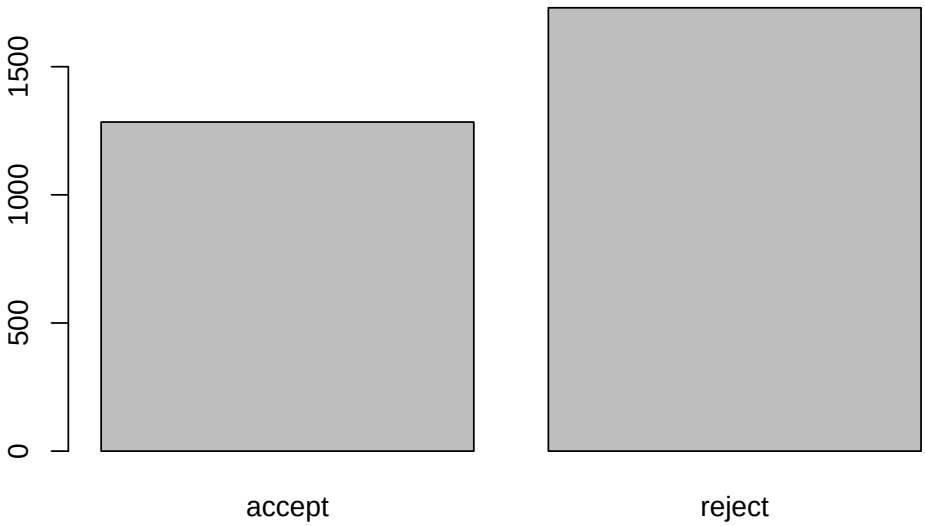
Click for answer

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

(d). Bar graph for one variable

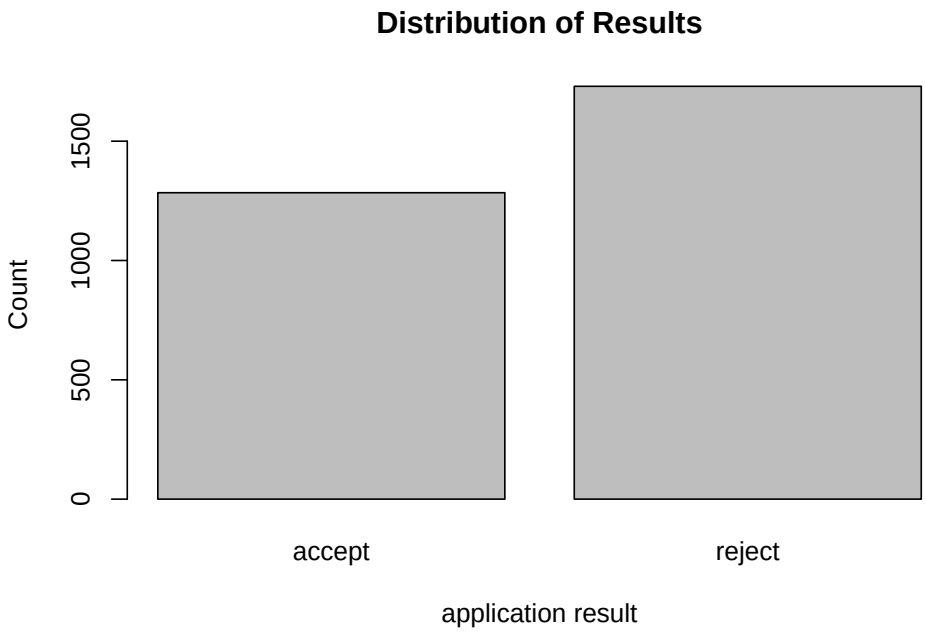
You can create a simple bar graph for one categorical variable with the `barplot` command. Here we visualize the distribution of result:

```
barplot(table(grad$result))
```



We can add in a title and x and y axis labels too:

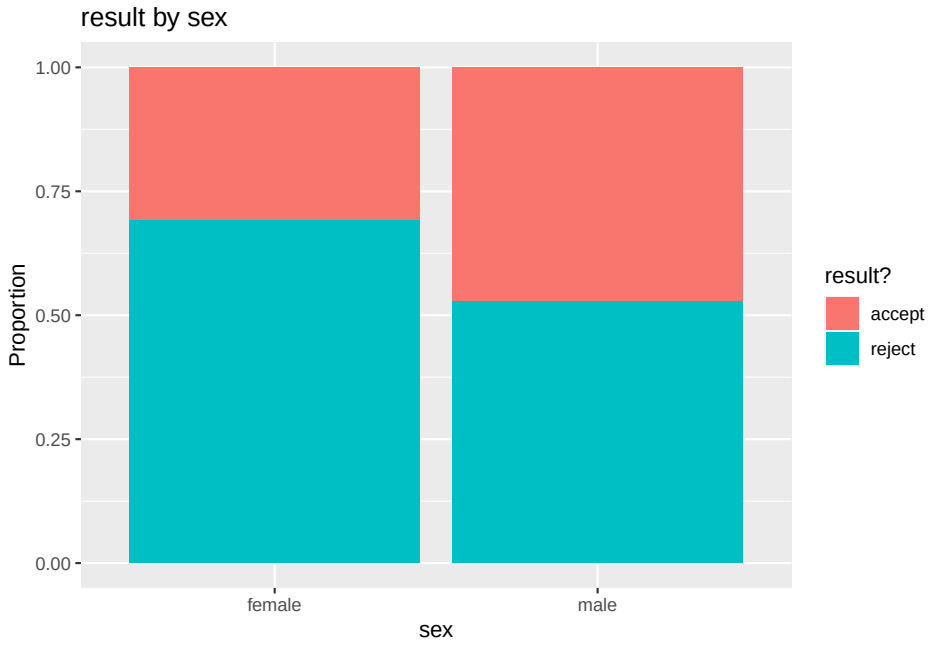
```
barplot(table(grad$result), xlab="application result",
        ylab="Count", main = "Distribution of Results")
```



(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 progra
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 program
head(grad.p2)
```

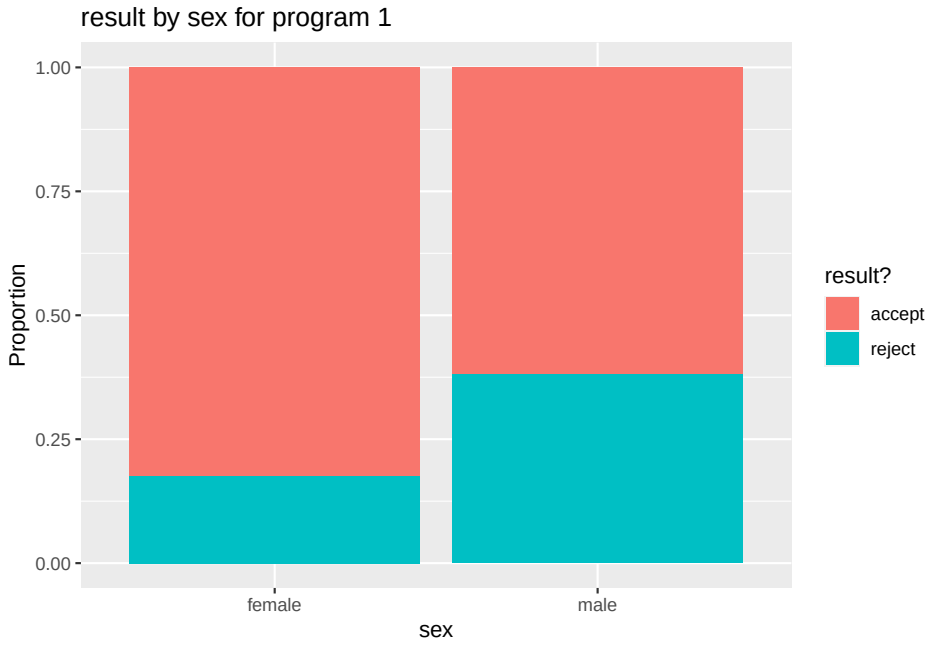
```
   program  sex result
1 program2 male accept
2 program2 male accept
3 program2 male accept
4 program2 male accept
5 program2 male accept
6 program2 male accept
```

(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)
```

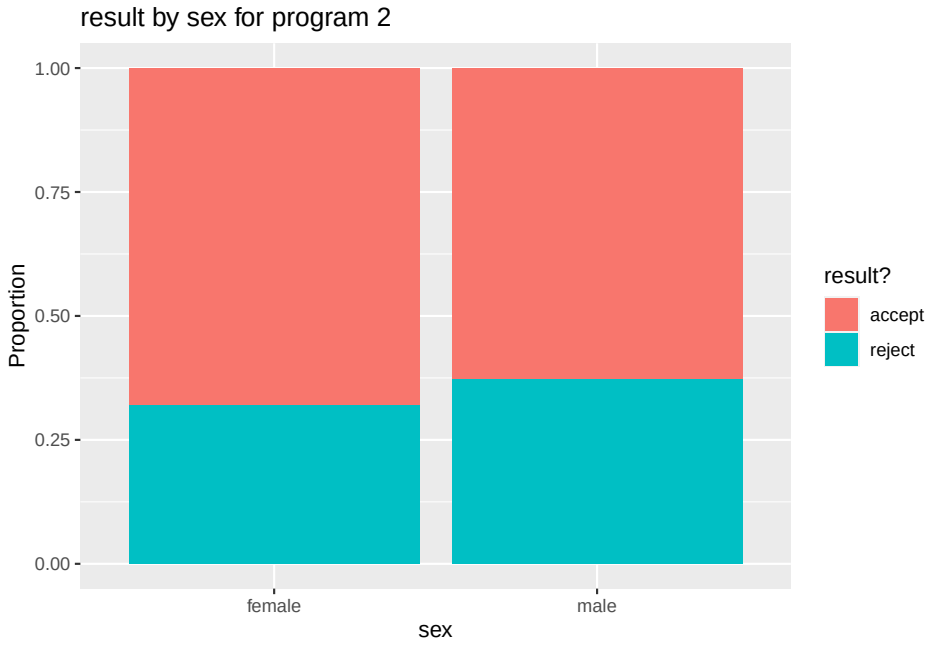
```
          accept    reject
female 0.8240741 0.1759259
male   0.6193939 0.3806061
```

(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)
```

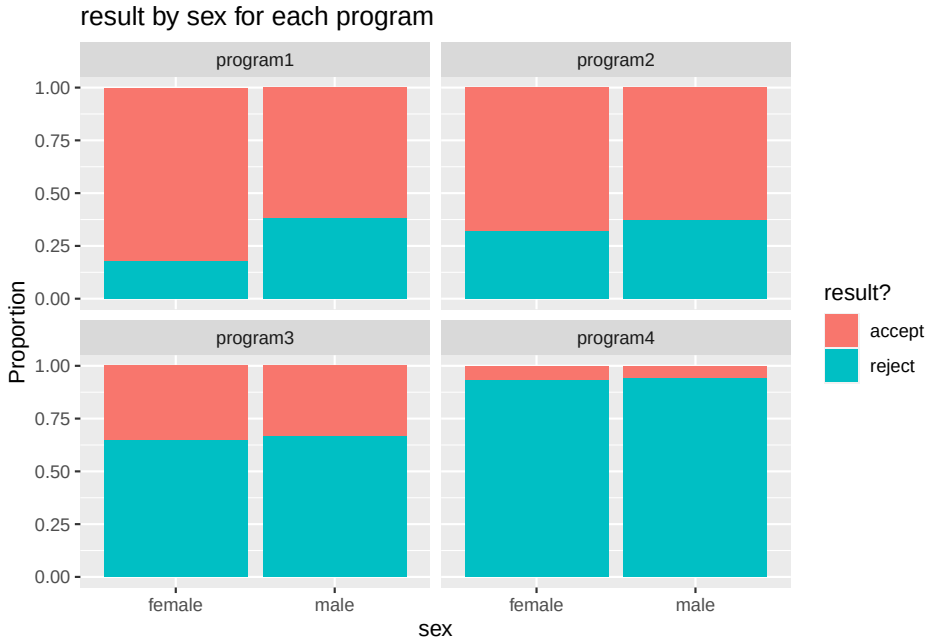
```
              accept     reject
  program1 0.64308682 0.35691318
  program2 0.63076923 0.36923077
  program3 0.34398977 0.65601023
  program4 0.06442577 0.93557423
```

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.

4.3 Quiz

1. A two-way table is shown for two groups, 1 and 2, and two possible outcomes, A and B.

	Outcome A	Outcome B	Total
Group 1	40	10	50
Group 2	30	120	150
Total	70	130	200

What proportion of all cases are in Group 1?

A. 0.33

B. 0.20

C. 0.25

D. 0.75

Click for answer

C. 0.25

2. A disruption of a gene called DYXC1 on chromosome 15 for humans may be related to an increased risk of developing dyslexia. Researchers studied the gene in 109 people diagnosed with dyslexia and in a control group of 195 others who had no learning disorder. The DYXC1 break occurred in 10 of those with dyslexia and in 5 of those in the control group. Is this an experiment or an observational study?

A. Experiment

B. Observational Study

Click for answer

Observational Study

3. The data from question 2 can be summarized in a two way table as:

	Gene Break	No Break	Total
Dyslexia Group	10	99	109
Control Group	5	190	195
Total	15	289	304

What is the proportion of Dyslexia group who have the break on the DYXC1 gene? Round your answer to 3 significant digits after the decimal.

A. 0.026

B. 0.667

C. 0.127

D. 0.092

Click for answer

D. 0.092

Chapter 5

Class Activity 5

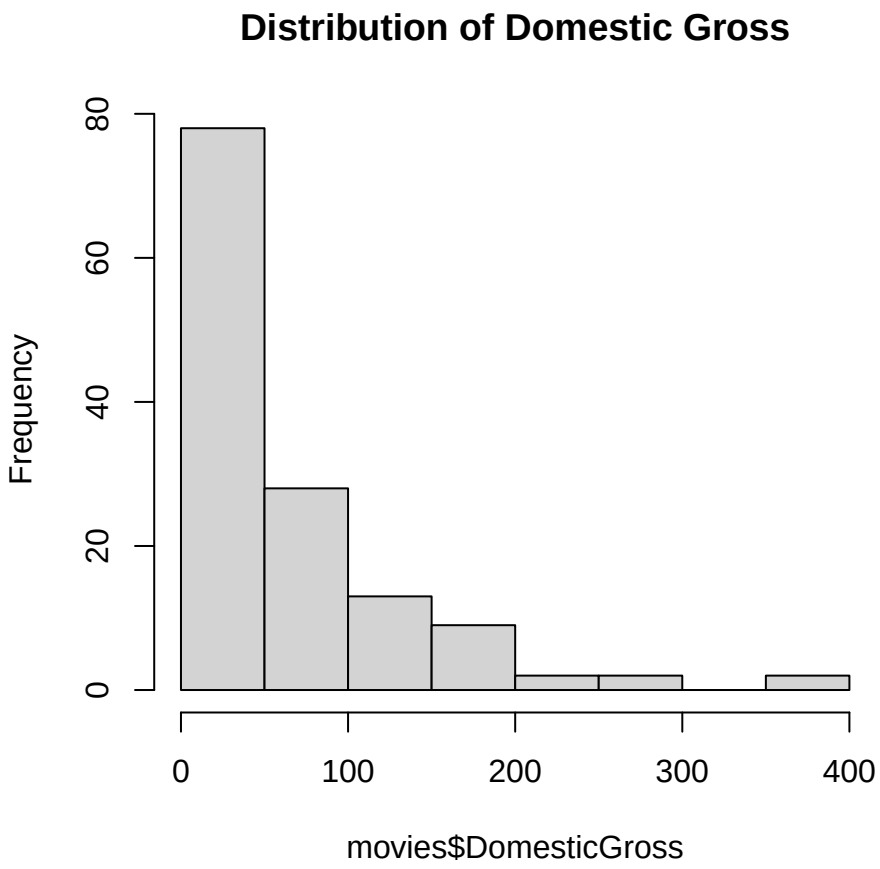
5.1 Your Turn 1

5.1.1 Hollywood Movies Domestic Gross

The dataset `HollywoodMovies2011` provides information on 136 movies that came out of Hollywood in 2011. We will look at the variable `DomesticGross`, which gives US domestic gross income for a movie from all viewers (in millions of dollars).

```
movies <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/HollywoodMovies20
```

```
hist(movies$DomesticGross, main="Distribution of Domestic Gross")
```



(a). Describe the shape of the distribution.

Click for answer

Answer: Skewed to the right

(b). Do there appear to be any outliers? If so, which values?

Click for answer

Answer: Yes, it looks like there are a few high outliers above 300 million.

(c). Finding outliers

We can find the row numbers of cases (movies) that have `DomesticGross` greater than 300 (300 million dollars):

```
which(movies$DomesticGross > 300)
```

```
[1]  4 14
```

Run the `which` command to verify that rows 4 and 14. Then find out which movies these are by subsetting the data frame:

```
movies[c(4,14), ]
```

```
                                        Movie
4  Harry Potter and the Deathly Hallows Part 2
14              Transformers: Dark of the Moon
            LeadStudio RottenTomatoes AudienceScore   Story
4          Warner Bros             96            92 Rivalry
14 DreamWorks Pictures             35            67   Quest
     Genre TheatersOpenWeek BOAverageOpenWeek DomesticGross
4  Fantasy             4375              38672        381.01
14  Action             4088              23937        352.39
   ForeignGross WorldGross Budget Profitability
4        947.10   1328.111    125     10.624888
14       770.81   1123.195    195      5.759974
   OpeningWeekend
4          169.19
14          97.85
```

Note that the `c(4,14)` part of this command creates a **vector** of the numbers 4 and 14 (the `c` stands for combine). Which movies are the outliers?

Click for answer

Answer: Harry Potter and the Deathly Hallows Part 2 and Transformers: Dark of the Moon.

(d). Use the histogram to answer: Is the median less than 100 million, about 100 million, above 100 million?

Click for answer

Answer: It is the point with half the data to the left and half to the right. The median is less than 100 since 100 roughly 110 (80 + 30) cases below it which is well over half the movies in the data set.

(e). Do you expect the mean to be greater than or less than the median. Explain.

Click for answer

Answer: Because the distribution is skewed to the right, we expect the mean to be larger than the median. The large outliers will pull the mean up and won't have much effect on the median.

(f). Computing the mean and median

You can get the mean and median a number of ways. Run these three commands:

```
mean(movies$DomesticGross)
```

```
[1] NA
```

```
median(movies$DomesticGross)
```

```
[1] NA
```

```
summary(movies$DomesticGross)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.02   19.03   37.35   63.22   80.46  381.01       2
```

What does `NA` stand for? How many movies have missing `DomesticGross`? You can subset the data to show you which cases have `NA` values for `DomesticGross`:

```
movies[is.na(movies$DomesticGross), ]
```

```
                            Movie LeadStudio
134                          Hugo  Paramount
136 Never Back Down 2: The Beatdown       Sony
    RottenTomatoes AudienceScore   Story     Genre
134             93            84         Adventure
136             NA            44 Rivalry    Action
    TheatersOpenWeek BOAverageOpenWeek DomesticGross
134             1277              8899            NA
136               NA                NA            NA
    ForeignGross WorldGross Budget Profitability
134           NA         NA     NA            NA
136           NA         NA      3             0
    OpeningWeekend
134          11.36
136           8.60
```

Click for answer

Answer: The `NA` value stands for "Not Available" which is used to code missing values. We can inspect the data frame and see that Hugo and Never Back Down 2 are the two movies that do not have domestic gross values.

(g). Missing data

There are some commands in R that "fail" as a default when missing data (`NA`) are present (`mean`, `median` and `sd` are examples). We can easily turn off this failure feature with the argument `na.rm=TRUE`

```
mean(movies$DomesticGross, na.rm=TRUE)
```

```
[1] 63.22276
```

```
median(movies$DomesticGross, na.rm=TRUE)
```

```
[1] 37.355
```

(h). Stats without outliers

There are a number of ways to "remove" outliers from an analysis. Here we use the square bracket `[]` notation along with a minus `-` to remove row 4 (Harry Potter) from the variable `DomesticGross` before our summary stat calculations:

```
summary(movies$DomesticGross[-4])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.02   18.88   37.30   60.83   80.36  352.39       2
```

Why does the mean change more than the median when this case is removed? (compare (g) and (h) mean and median values)

Click for answer

Answer: Both values go down after removing the highest grossing movie of the year, but the drop in the mean is more substantial. The mean drops by almost 4% when Harry Potter is removed while the median only drops by about 0.1%.

```
100*(60.83 - 63.22276)/63.22276 # percent change in the mean
```

```
[1] -3.78465
```

```
100*(37.30 - 37.355)/37.355 # percent change in the median
```

```
[1] -0.147236
```

(i). Computing standard deviation

The standard deviation command is `sd`. We need to add the `na.rm` argument to obtain the SD for `DomesticGross`:

```
sd(movies$DomesticGross, na.rm=TRUE)
```

```
[1] 69.41799
```

Look again at the distribution of `DomesticGross` shown in the histogram. Why is SD (variation around the mean) an inadequate measure of variation for this type of distribution?

Click for answer

Answer: The is much more variation (spread) to the data above the mean than below it. Because the distribution is strongly skewed right, we can't use one measure of variation when describing how `DomesticGross` values vary around some central value (like a mean).

(j). Stats by `Genre`

The `tapply(y, x, stat)` command gives the `stat` value of `y` for each level of `x`. Here we get the `summary` of `DomesticGross` for each type of `Genre`:

```
tapply(movies$DomesticGross, movies$Genre, summary)
```

```
$Action
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.54   24.96   40.26   91.02  161.53  352.39       1

$Adventure
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
     NA      NA      NA     NaN      NA      NA       1

$Animation
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.39   51.41  115.67  104.62  142.86  191.45

$Comedy
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.79   23.21   37.41   56.51   69.75  254.46

$Drama
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.38    4.40   13.30   32.37   51.16  169.22

$Fantasy
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.32   96.24  191.16  191.16  286.09  381.01
```

```
$Horror
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.02   17.69   24.05   34.87   38.18  127.00

$Romance
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.03   18.51   39.05   61.40   70.26  260.80

$Thriller
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.02   31.18   40.49   41.44   62.50   79.25
```

- Which movies genre has the highest median domestic gross?
- Why are there no summary stats for the adventure genre?

Click for answer

Answer: To help answer these questions you really should explore the number of movies in each genrewith the table command.

- The fantasy genre has the highest median domestic gross ($381 million). But note that only two movies have this classification in 2011. The action genre was second highest at $352 million and there were 12 movies in this category.
- The adventure genre only has one movie (Hugo) and this movie is also missing a value for `DomesticGross`!

```
table(movies$Genre)
```

```
   Action Adventure Animation    Comedy     Drama   Fantasy
       32         1        12        27        21         2
   Horror   Romance  Thriller
       17        11        13
```

```
which(movies$Genre == "Adventure")
```

```
[1] 134
```

```
movies[134, ]
```

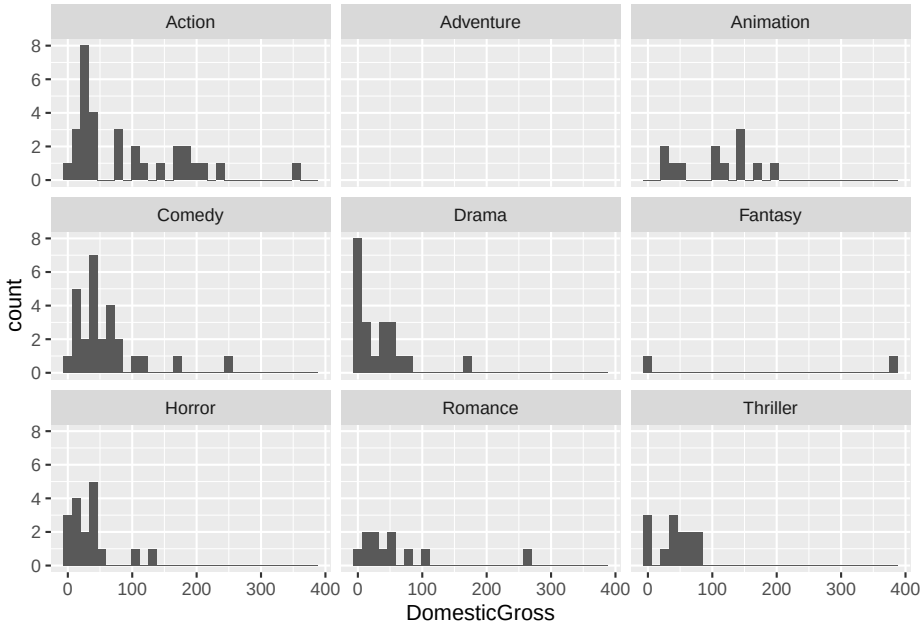
```
    Movie LeadStudio RottenTomatoes AudienceScore Story
134  Hugo  Paramount            93            84
```

```
        Genre TheatersOpenWeek BOAverageOpenWeek
134 Adventure             1277              8899
    DomesticGross ForeignGross WorldGross Budget
134            NA           NA         NA     NA
    Profitability OpeningWeekend
134            NA          11.36
```

(k). Extra: Histogram of `DomesticGross` by `Genre`

(Not in Lab Manual) The `ggplot2` package allows you to create histograms separated by a categorical variable using the `facet_wrap` command. Assuming that `ggplot2` is already installed, all you need to do is load it with `library` then create your graph:

```
library(ggplot2)
ggplot(movies, aes(x=DomesticGross)) +
  geom_histogram() +
  facet_wrap(~Genre)
```



Which genre has the most variability in domestic gross?

Click for answer

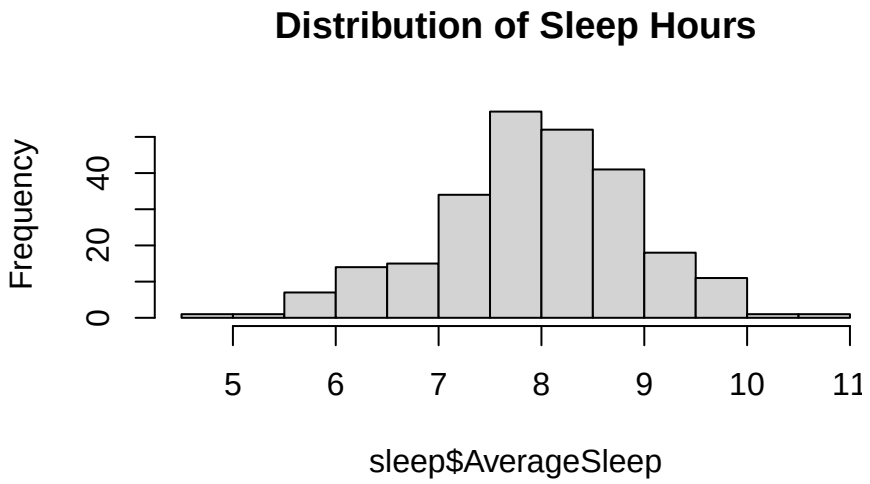
Answer: The action genre has the largest range of values.

5.2 Your turn 2

5.2.1 Example 2: Sleep

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

```
sleep <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/SleepStudy.csv")
hist(sleep$AverageSleep, main="Distribution of Sleep Hours")
```



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

Click for answer

Answer: The mean is around 8 hours

(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
```

5.2.2 Example 3: 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.

(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.

(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.

5.3 Example 4: 5 number summaries

For each five number summary below, indicate whether the data appear to be symmetric, skewed to the right, or skewed to the left.

(a). (2, 10, 15, 20, 69)

```
my_vector1 <- c(1, 10, 15, 20, 69)
summary(my_vector1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1      10      15      23      20      69
```

Click for answer

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

(b). (10, 57, 85, 88, 93)

```
my_vector2 <- c(10, 57, 85, 88, 93)
summary(my_vector2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.0    57.0    85.0    66.6    88.0    93.0
```

Click for answer

Answer: Skewed left since mean is less than median.

(c). (200, 300, 400, 500, 600)

```
my_vector3 <- c(200, 300, 400, 500, 600)
summary(my_vector3)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    200     300     400     400     500     600
```

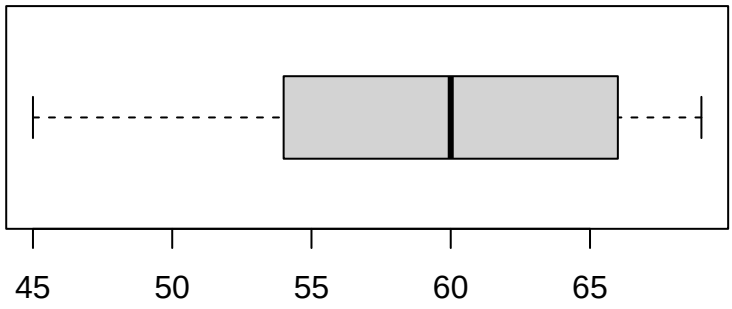
Click for answer

Answer: Symmetric since mean is same as median.

5.4 Example 5: 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")
boxplot(hotdogs$HotDogs, xlab="Number of Hot Dogs Consumed", horizontal=T)
```



(a). Use the boxplot to estimate the 5 number summary and IQR for this data.

Click for answer

Answer: min = 45, Q1 = 54, m = 60, Q3 = 66, max = 69. IQR is about 66-54 or 12 hotdogs

(b). Computing 5 number summaries

R doesn't have '5 number summary' command, but `summary` gives you a "6" number summary by adding the mean to the 5 number summary. You can also use `IQR` to get the IQR:

```
summary(hotdogs$HotDogs)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  45.00   54.00   60.00   58.64   65.00   69.00
```

```
IQR(hotdogs$HotDogs)
```

```
[1] 11
```

How close were your guesses from the boxplot to the values given by this command?

Click for answer

(Answers will vary) Within one hotdog of the R values.

(c). Use the boxplot outlier rule to verify that there are no outliers in this data.

Click for answer

Answer:

- $1.5IQR = 18$ hotdogs.

- Lower fence: $Q1 - 1.5IQR = 54 - 18 = 32 < min$ so there are no low outliers.
- Upper fence: $Q3 + 1.5IQR = 65 + 18 = 83 > max$ so there are no high outliers.

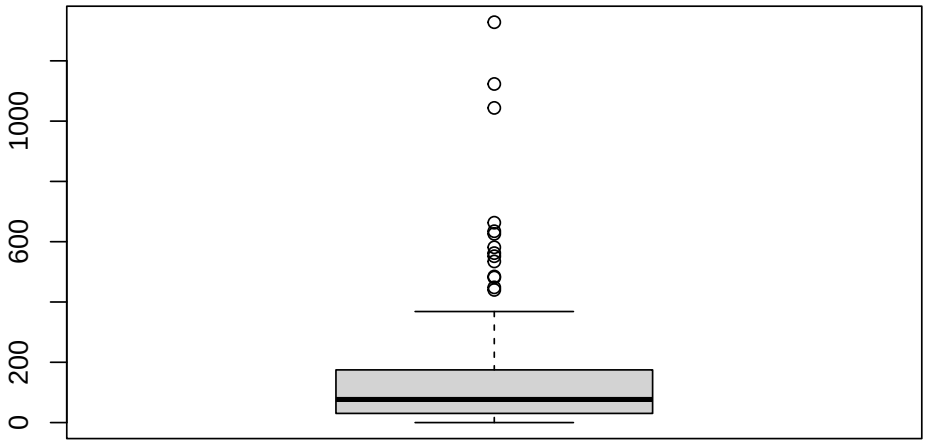
5.5 Examples 6: Hollywood Movies World Gross revisited

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

```
movies <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/HollywoodMovies20
```

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

```
boxplot(movies$WorldGross)
```



How many movies are identified as outliers for world gross?

Click for answer

Answer: Just using the boxplot, there looks to be about 10 movies that are high outliers

(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.

```
summary(movies$WorldGross)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.025   30.706   76.659  150.742  173.691 1328.111
    NA's
       2
```

```
IQR(movies$WorldGross, na.rm = TRUE)
```

```
[1] 142.985
```

Click for answer

- $1.5IQR = 1.5(142.985) = 214.48$ hundred million dollars
- Lower fence: $Q1 - 1.5IQR = 30.710 - 214.48 = -183.8 < min$ so there are no low outliers.
- Upper fence: $Q3 + 1.5IQR = 173.7 + 214.48 = 388.18 < max$ so there are high outliers.
- The upper whisker extends to the largest movie value that is below the fence of 388.18. You could look at the data spreadsheet and find which movie comes closest to this fence, but a quicker way is to use R. First we can use which to find out the row numbers of the movies with less than 388.18 in WorldGross. Then use this set to find out the max of the WorldGross within this group of movies, which turns out to be 368.404 hundred million dollars.

```
1.5*IQR(movies$WorldGross, na.rm = TRUE)
[1] 214.4775
30.710 - 214.48
[1] -183.77
173.7 + 214.48
[1] 388.18
```

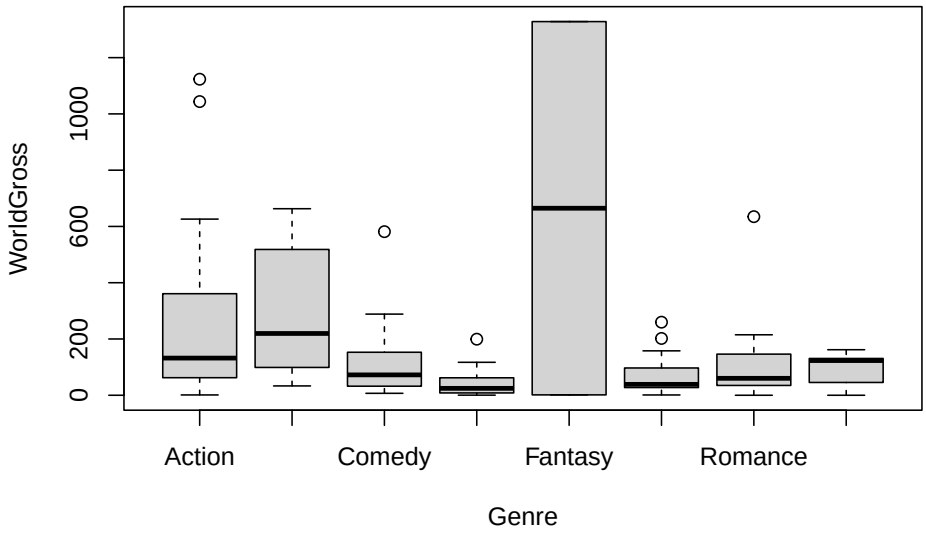
```
notoutliers <- which(movies$WorldGross < 388.18)
max(movies$WorldGross[notoutliers])
[1] 368.404
which(movies$WorldGross == 368.404)
[1] 49
movies[49,]
                                Movie LeadStudio
49 Captain America: The First Avenger     Disney
```

```
   RottenTomatoes AudienceScore         Story  Genre
49             78            75 Metamorphosis Action
   TheatersOpenWeek BOAverageOpenWeek DomesticGross
49             3715             17512        176.65
   ForeignGross WorldGross Budget Profitability
49       191.75    368.404    140      2.631457
   OpeningWeekend
49          65.06
```

(c). Side-by-side boxplot

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

```
boxplot(WorldGross ~ Genre, data=movies)
```



- 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 fantacy genre only has 2 movies in it

- What is one issue with the default version of this graph?

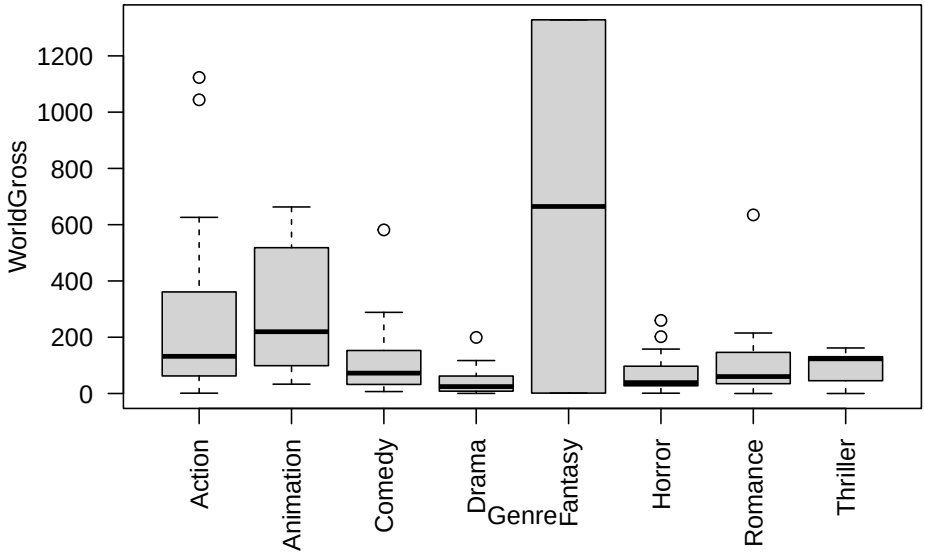
Click for answer

Answer: The genre labels are not all present.

(d). Improving the default boxplot

There are many values in `Genre` for this data and their values (levels) have longer names. This can cause issues when using these names to label graphs, like the x-axis in your boxplot. The are many (many, many) ways to modify graphs in R. Here is one way to change the label orientation on your x-axis.

```
boxplot(WorldGross ~ Genre, data=movies, las=2)
```



The `las` arguments let's you change the orientation of the axis labels relative to the axis. The value of 2 makes the labels perpendicular to the axis.

5.6 Example 8: Ants on a Sandwich

The number of ants climbing on a piece of a peanut butter sandwich left on the ground near an anthill for a few minutes was measured 7 different times and the results are: $43, 59, 22, 25, 36, 47, 19$

(a). Calculate the mean number of ants.

Click for answer

Answer: $\bar{x} = 35.857$

(b). Calculate the median number of ants.

Click for answer

Answer: Order data then find middle value: $19, 22, 25, 36, 43, 47, 59$. Then $m = 36$

(c). Calculate the quartiles for the number of ants.

Click for answer

Answer: Since $m = 36$, the first quartile will be the median of $19, 22, 25 : Q1 = 22$. The third quartile will be the median of $43, 47, 59 : Q3 = 47$.

Chapter 6

Class Activity 6

6.1 Your Turn 1

6.1.1 Beer Example

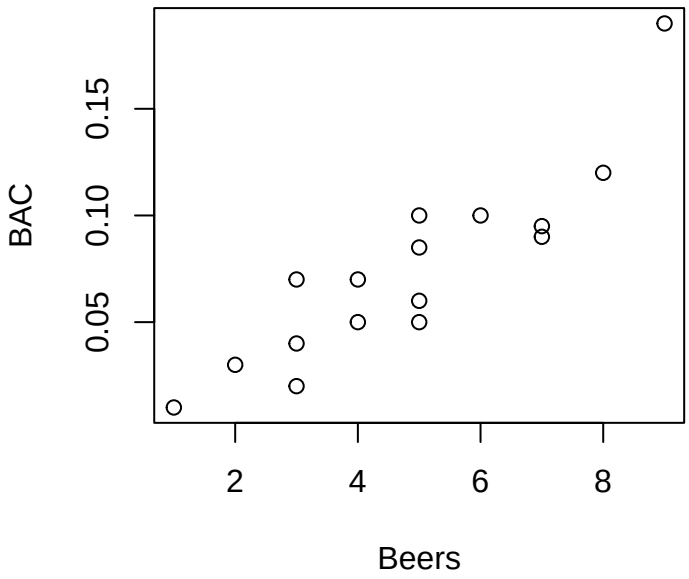
A study of 16 Ohio State University students looked at the relationship between the number of beers a student consumes and their blood alcohol content (BAC) 30 minutes after their last beer. The regression information from R to predict BAC from number of beers consumed is given below.

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

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

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

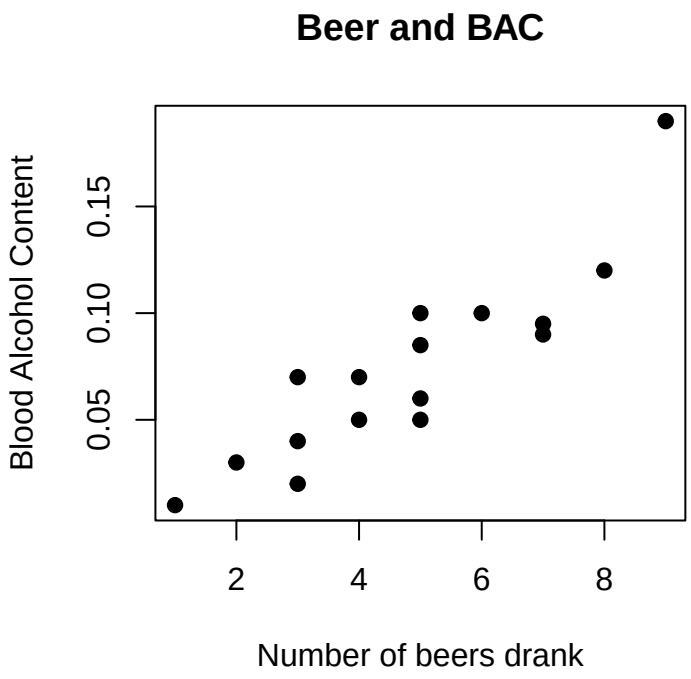
```
plot(BAC ~ Beers, data=bac)
```



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

You can modify this basic graph by adding a title and changing the plotting symbol. The `pch=19` argument changes the symbols to filled circles.

```
plot(BAC ~ Beers, data=bac, pch=19,
     main="Beer and BAC", xlab="Number of beers drank", ylab = "Blood Alcohol Content")
```



(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
```

(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
```

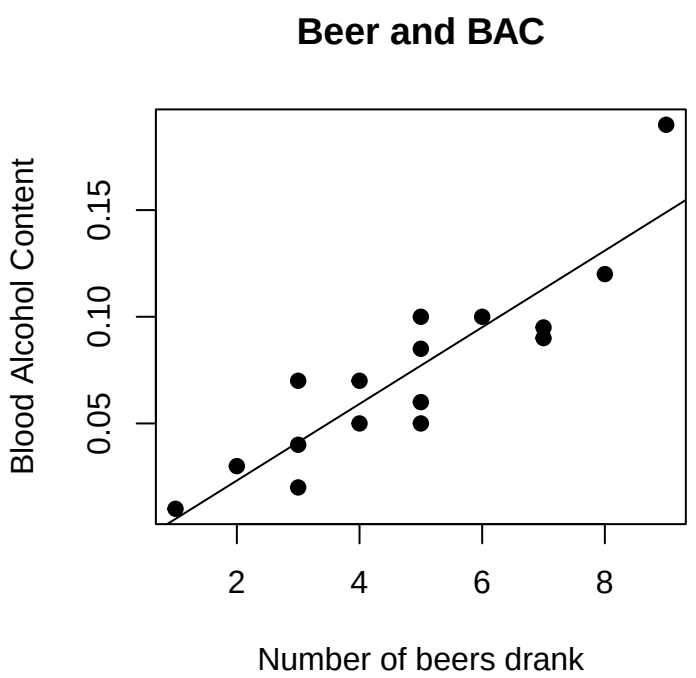
- After running the `lm` command above in your R console, check the **Environment** tab to see that the object `bac.lm` is now one of the objects stored in R's memory (for this session of Rstudio).
- Write down the fitted regression equation to predict BAC from number of beers.

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:

```
# Need to call the plot function again!!

plot(BAC ~ Beers, data=bac, pch=19,
     main="Beer and BAC", xlab="Number of beers drank", ylab = "Blood Alcohol Content")
abline(bac.lm) # adds regression line to the plot above
```



(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.

(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.

(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
```

(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
```

(h). Getting residuals in R

Click for answer

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

```
# part h
resid(bac.lm)
```

```
           1            2            3            4
 0.022881795  0.006773080  0.041026747 -0.011009491
           5            6            7            8
-0.001190682 -0.018045729  0.028809318 -0.017118205
           9           10           11           12
-0.021190682 -0.027118205  0.010845557  0.004918033
          13           14           15           16
 0.007881795 -0.023045729  0.004736842 -0.009154443
```

Notice that case 2 in the data drank 2 beers and had a BAC recorded as 0.03. We can see that their residual value matches our answer to (g) up to some rounding error.

```
# part h
bac$BAC[2]
```

```
[1] 0.03
```

```
bac$Beers[2]
```

```
[1] 2
```

```
resid(bac.lm)[2]
```

```
         2
0.00677308
```

(i). Getting R^2 value

Click for answer

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)
```

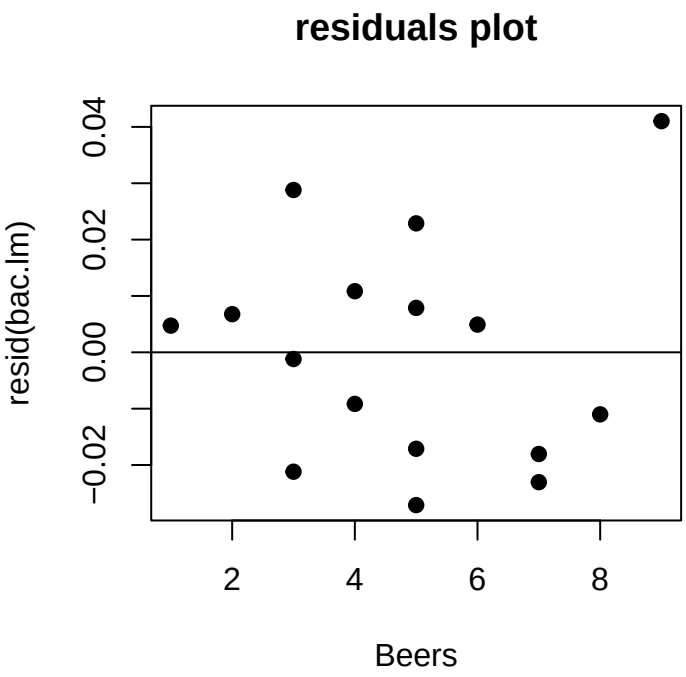
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

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
```

(j). Making a residuals plot

Click for answer

The regression of `BAC` on `Beers` has a residuals plot that plots the model's residuals on the y-axis and the explantory ("predictor") on the x-axis. We add a horizontal reference line (the detrended regression line) with the `abline(h=0)` command:

```
# code for residual plot
plot(resid(bac.lm) ~ Beers, data=bac, pch=19, main = "residuals plot")
abline(h=0)
```



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?

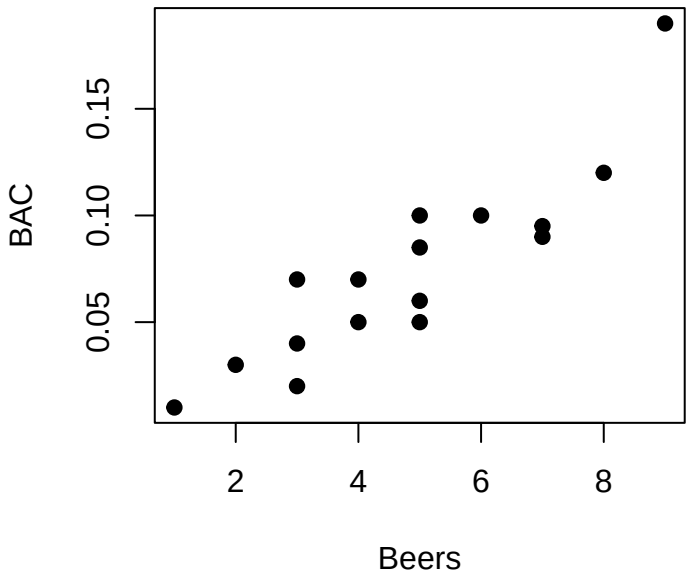
Click for answer

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.

(k). Identifying points The `which` command can be used to identify points by their row number in a scatterplot.

We can use `==` to see which case drank exactly 9 beers. Which is the row number of the case that drank 9 beers?

```
plot(BAC ~ Beers, data=bac, pch=19)
```




```
which(bac$Beers == 9)
```

```
[1] 3
```

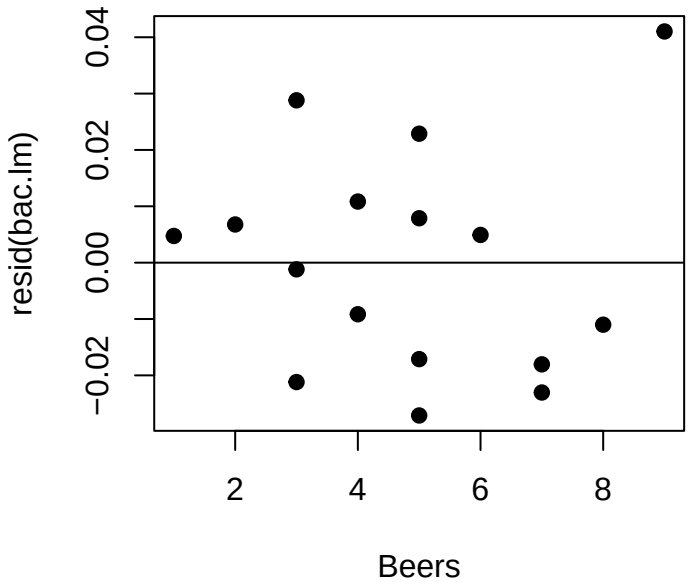
Click for answer

Answer: Row 3.

What is the row number of the case with the most negative residual?

Click for answer

```
plot(resid(bac.lm) ~ Beers, data=bac, pch=19)
abline(h=0)
```



We could eyeball the graph to see that the most negative residual is less than -0.02:

```
# which case has resid less than -0.02?

resid(bac.lm)[which(resid(bac.lm) < -0.02)]
```

```
          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
```

(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):

Click for answer

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

```r
# 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
```

```r
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.

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.

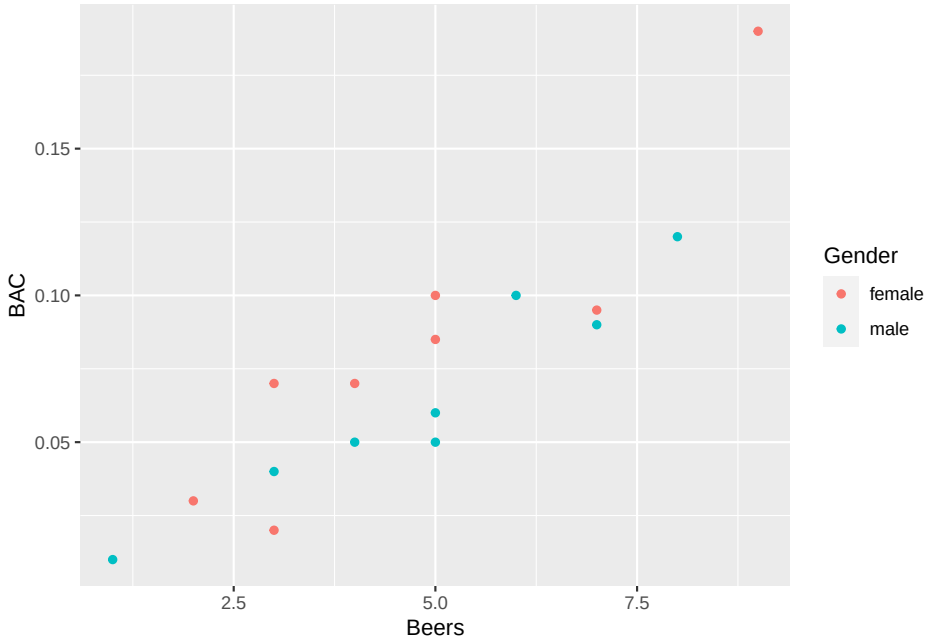
Click for answer

Answer: The R^2 decreases 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.

(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:

```
library(ggplot2)
ggplot(bac, aes(x=Beers, y=BAC, color=Gender)) + geom_point()
```



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

Click for answer

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

(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:

```
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
```

```
# enter code for the male model

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:

```
bac.female <- subset(bac, sub = Gender == "female")
lm(BAC ~ Beers, data=bac.female)
```

```

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

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

```
bac.male <- subset(bac, sub = Gender == "male")
lm(BAC ~ Beers, data=bac.male)
```

```

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

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

6.2 Your Turn 2

6.2.1 Mice Mass Example

The time of day in which calories are consumed can affect weight gain. At least, that appears to be true in mice. Mice normally eat all their calories at night, but when mice ate some of their calories during the day (when mice are supposed to be sleeping), they gained more weight even though all the mice ate the same

total amount of calories. Here we look at the regression of body mass gain in grams, `BMGain`, against the percent of calories eaten during the day, `DayPct` for a study involving 27 mice. The R commands needed to answer the questions below are:

```
mice <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/MICE.csv")

plot(BMGain ~ DayPct, data=mice, pch=19)
mice.lm <- lm(BMGain ~ DayPct, data=mice)
mice.lm
```

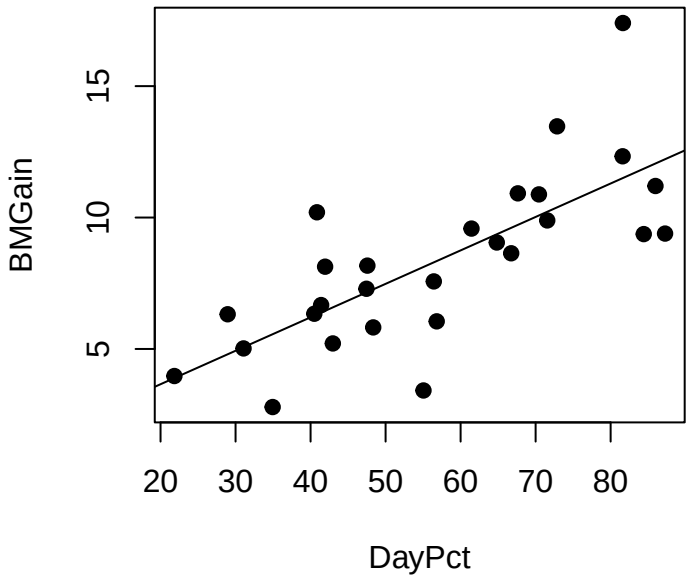
```
Call:
lm(formula = BMGain ~ DayPct, data = mice)

Coefficients:
(Intercept)       DayPct
     1.1128       0.1273
```

```
cor(mice$BMGain, mice$DayPct)
```

```
[1] 0.7398623
```

```
abline(mice.lm) # adds regression line to previously created scatterplot
```



(a). What are the coordinates (roughly) of the case with the largest positive residual?

```
mice[which(resid(mice.lm) == max(resid(mice.lm))),]
```

```
    X Light BMGain Corticosterone DayPct Consumption
25 25    LL   17.4         66.679 81.636       7.177
   GlucoseInt   GTT15  GTT120 Activity
25        Yes 435.644 405.941     6702
```

Click for answer

Answer: The case with the largest residual is located at about 80% calories and 17g body mass gain. We can find which row this corresponds to using the `which` command shown below.

(b). What are the coordinates (roughly) of the case with the most negative residual?

```
mice[which(resid(mice.lm) == min(resid(mice.lm))),]
```

```
    X Light BMGain Corticosterone DayPct Consumption
10 10    DM   3.42         208.26 55.051       3.857
   GlucoseInt   GTT15  GTT120 Activity
10         No 271.717 148.485     1084
```

Click for answer

Answer: The case with the most negative residual is located at about 55% calories and 3g body mass gain. We can find which row this corresponds to using the `which` command shown below. The code below also highlights the cases in (a) with a circle and (b) with a square.

(c). What is the predicted body mass gain for a mouse that eats 50% of its calories during the day?

$$\widehat{BMGain} = 1.1128 + 0.1273(50) = 7.48$$

```
1.1128 + .1273*50
```

```
[1] 7.4778
```

Click for answer

Answer: A mouse that eats 50% of its calories during the day is predicted to gain 7.48 grams.

(d). Find the residual for the mouse who ate 48.3% of its calories during the day and gained 5.82 grams.

Click for answer

Answer: We first find the predicted body mass gain:

$$\widehat{BMGain} = 1.1128 + 0.1273(48.3) = 7.26$$

The residual is then:

$$Residual = BMGain - \widehat{BMGain} = 5.82–7.26 = –1.44.$$

```
1.1128 + .1273*48.3
```

```
[1] 7.26139
```

```
5.82 - (1.1128 + .1273*48.3)
```

```
[1] -1.44139
```

(e). Interpret the slope of the regression line in context.

Click for answer

Answer: The slope is 0.1273. When a mouse eats one more percent of its calories during the day, its predicted body mass gain goes up by 0.1273 grams.

(f). Interpret the intercept of the line in context, if it makes sense to do so.

Click for answer

Answer: The intercept is 1.1128. A mouse who eats 0% of its calories during the day (and all of them at night when a mouse normally eats all its food) is predicted to gain 1.11 grams. But this would be **extrapolation** because the range of observed percents is, roughly, 20-90. It does not make sense to interpret the intercept in this context.

(g). Use the correlation value to compute R^2, then interpret (in context) the R^2 value for this model.

```
r <- 0.7398623
r^2
```

```
[1] 0.5473962
```

(h). Get the value of R^2 from the regression output, then interpret (in context) the R^2 value for this model.

```
summary(mice.lm)
```

```
Call:
lm(formula = BMGain ~ DayPct, data = mice)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6990 -1.1694  0.0728  0.9174  5.8975

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.11280    1.38211   0.805    0.428
DayPct       0.12727    0.02315   5.499 1.03e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.231 on 25 degrees of freedom
Multiple R-squared:  0.5474,    Adjusted R-squared:  0.5293
F-statistic: 30.24 on 1 and 25 DF,  p-value: 1.032e-05
```

Click for answer

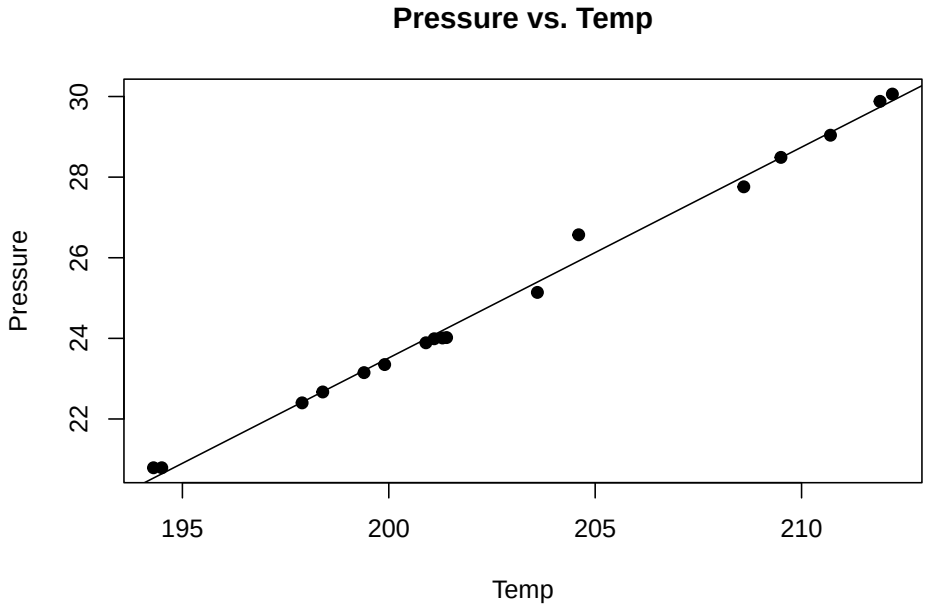
Answer: From `Multiple R-squared`, we get $R^2 = 0.547$. The percent of calories that a mouse eats during the day explains about 55% of the variability in weight gain for this study.

6.2.2 Forbes Example

In the mid 1800s, James D. Forbes conducted a experiments designed to determine if the atmospheric pressure at a given location can just be determined by the boiling temp of water at that location.

(a). Fit the linear regression of `Pressure` on `Temp`:

```
forbes <-  read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/forbes
plot(Pressure ~ Temp, data=forbes, pch=19, main = "Pressure vs. Temp")
forbes.lm <- lm(Pressure ~ Temp, data=forbes)
abline(forbes.lm)
```




```
summary(forbes.lm)
```

```
Call:
lm(formula = Pressure ~ Temp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
Temp          0.52289    0.01011   51.74   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2328 on 15 degrees of freedom
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.9941
F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

- Describe the relationship between pressure and temp (strength, form, direction).

Click for answer

Answer: This is a strong, positive relationship that looks linear.

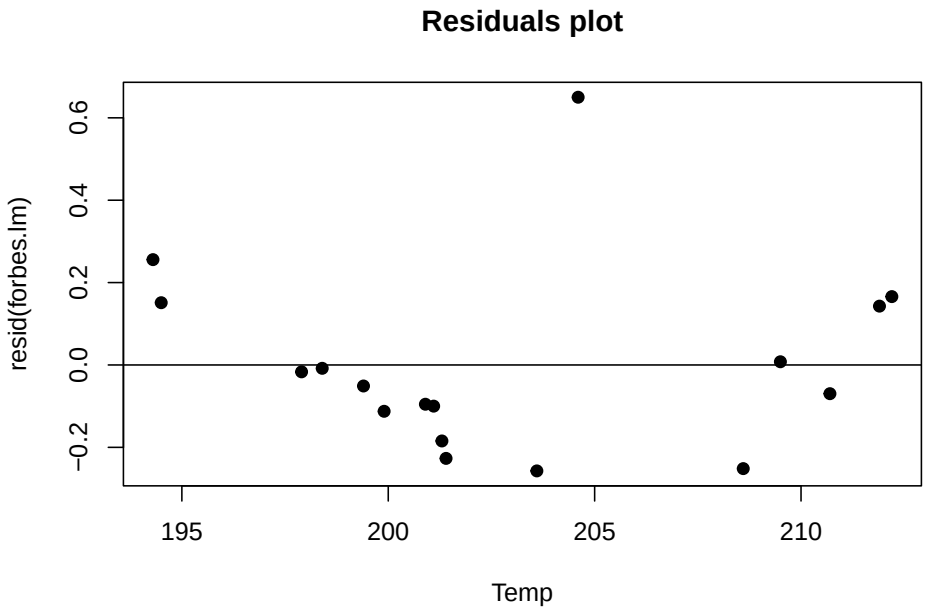
- Interpret the value of R^2

Click for answer

Answer: About 99.4% of the variation observed in pressure can be explained by the boiling point temps.

(b). Check the residuals plot

```
plot(resid(forbes.lm) ~ Temp, data=forbes, pch=19, main = "Residuals plot")
abline(h=0)
```



- Is the relationship between pressure and temp linear?

Click for answer

Answer: No! There is curvature, which means the linear model is systematically underestimating pressure at low and high temps and overestimating pressure at mid-range temps.

- Does the residual plot highlight an unusual case? Explain.

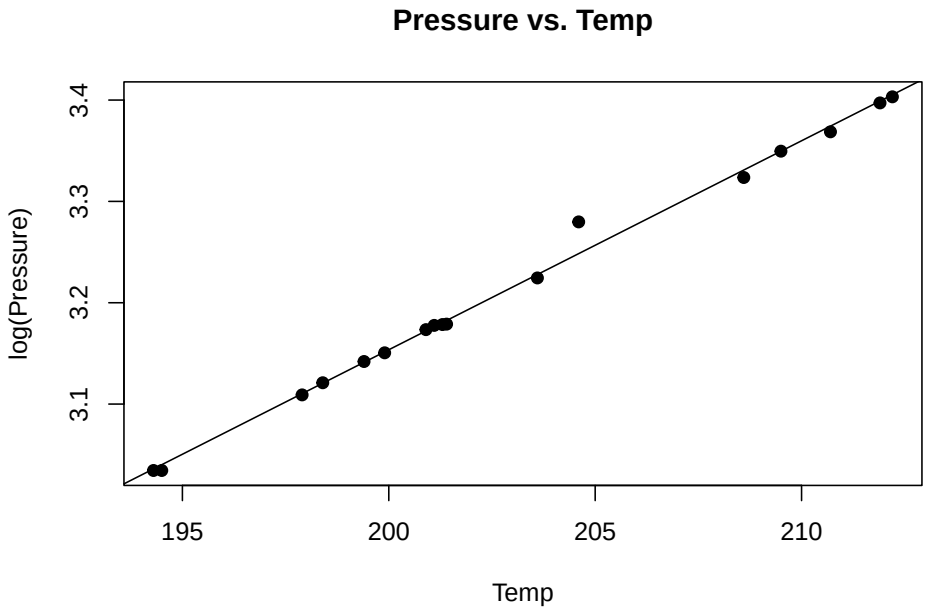
Click for answer

Answer: Yes, there is one case that has an unusually high pressure value given its temp.

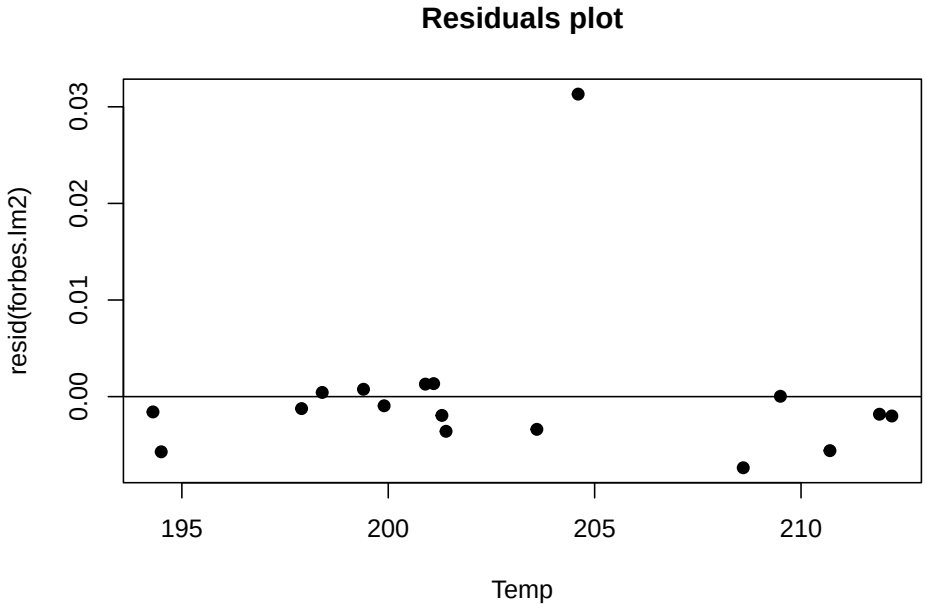
(c). "Fixing" the model

A linear model can be used with this data if we **transform** the response variable to the logarithmic scale. Here `log(y)` gives the natural log of the variable `y`.

```
plot(log(Pressure) ~ Temp, data=forbes, pch=19, main = "Pressure vs. Temp")
forbes.lm2 <- lm(log(Pressure) ~ Temp, data=forbes)
abline(forbes.lm2)
```




```
plot(resid(forbes.lm2) ~ Temp, data=forbes, pch=19, main = "Residuals plot")
abline(h=0)
```



- Has the curvature in the scatterplot and residuals plots been reduced by logging the variables?

Click for answer

Answer: Yes, there is less curvature

- Has the outlier been eliminated by logging the variables?

Click for answer

Answer: No, the outlier is still present.

(d). Removing bad measurement

Identify which case has the large residual value around 0.03.

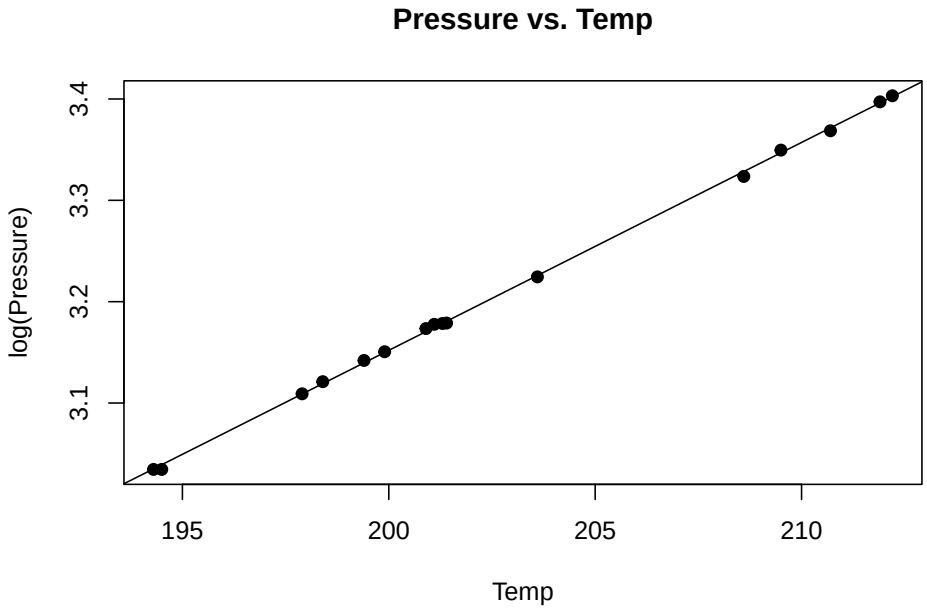
```
resid(forbes.lm2)[which(resid(forbes.lm2) > 0.02)]
```

```
        12
0.03131388
```

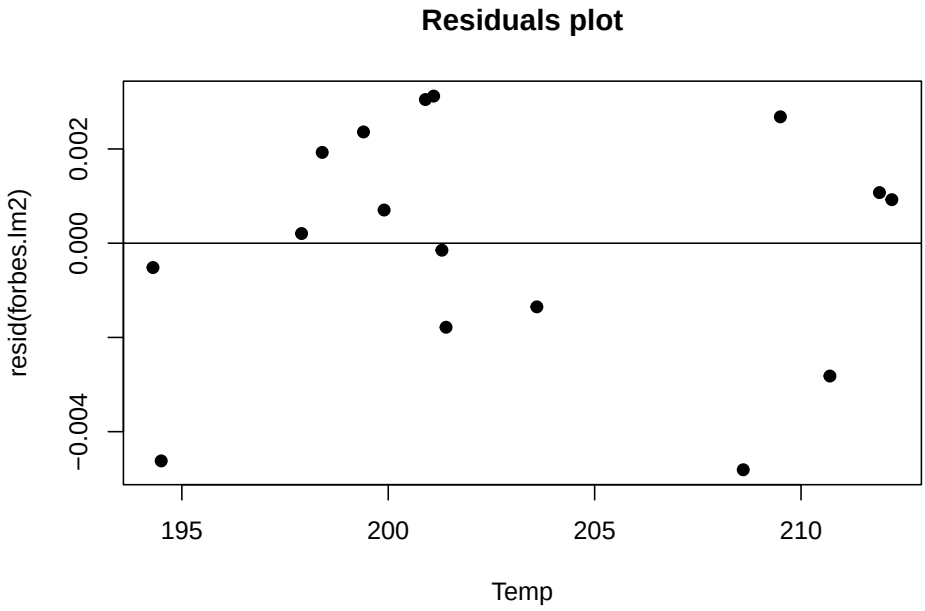
Repeat part (c) but this time remove the case you identified. The easiest way to do this is to create a new version of the data with row 12 removed:

Click for answer

```
forbes2 <- forbes[-12, ]
plot(log(Pressure) ~ Temp, data=forbes2, pch=19, main = "Pressure vs. Temp")
forbes.lm2 <- lm(log(Pressure) ~ Temp, data=forbes2)
abline(forbes.lm2)
```




```
plot(resid(forbes.lm2) ~ Temp, data=forbes2, pch=19, main = "Residuals plot")
abline(h=0)
```


Residuals plot
resid(forbes.lm2)
0.002
0.000
−0.004
195
200
205
210
Temp

Chapter 7

Class Activity 7

7.1 Your Turn 1

7.1.1 Parameters and Statistics

Here are some notations that will be useful for you. Look for the codes to produce this in the associated Rmd file.

	Parameter	Statistic
Mean	μ	$\bar{x}$
Proportion	p	$\hat{p}$
Std. Dev.	σ	s
Correlation	ρ	r
Slope	β	b

7.1.2 Example 1: Parameters and Statistics

For each of the following, state whether the quantity described is a parameter or a statistic, and give the correct notation.

(a). Average household income for all houses in the US, using data from the US census

Click for answer

Answer: This is a parameter since the mean is for all houses in the US, and the notation is μ.

(b). The proportion of all residents in a county who voted in the last presidential election.

Click for answer

Answer: This is a parameter since we have information on all the residents, and the notation is p.

(c). The difference in proportion who have ever smoked cigarettes, between a sample of 500 people who are 60 years old and a sample of 200 people who are 25 years old.

Click for answer

Answer: We use statistics since the proportions are from samples. The notation for the difference in sample proportions is $\hat{p}_1 - \hat{p}_2$

(d). The correlation between weight and height for 5-year old kids.

Click for answer

Answer: If we are looking at all 5-year old kids it is a parameter, and the notation for correlation is ρ.

(e). The mean number of extracurricular activities from a random sample of 50 students at your school.

Click for answer

Answer: This is a statistic since the mean is from a sample, and the notation is μ.

7.1.3 Example 2: 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.

(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
```

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

Click for answer

Answer: Statistic

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

---

7.2 Your Turn 2

7.2.1 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.

```
# Define parameters
set.seed(123) # set seed for reproducibility
pop.prop <- .66 # Population proportion
n.size <- 100  # sample size
```

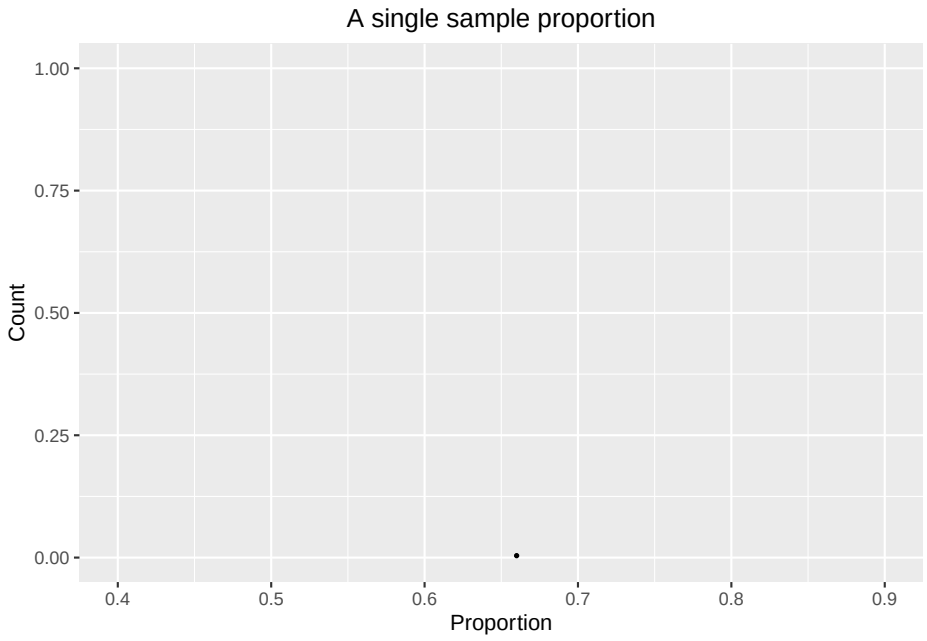
(a). Generate a random sample of size $n = 100$ and plot its sample proportion.

```
# 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
```

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

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

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = sample.prop1)) +
  geom_dotplot(dotsize=0.25, stackratio=0.75, binwidth=0.01) +
  ggtitle("A single sample proportion") +  xlab("Proportion") + ylab("Count")+
  scale_x_continuous(limits = c(0.4, 0.9))+
  theme(plot.title = element_text(hjust = 0.5))
```

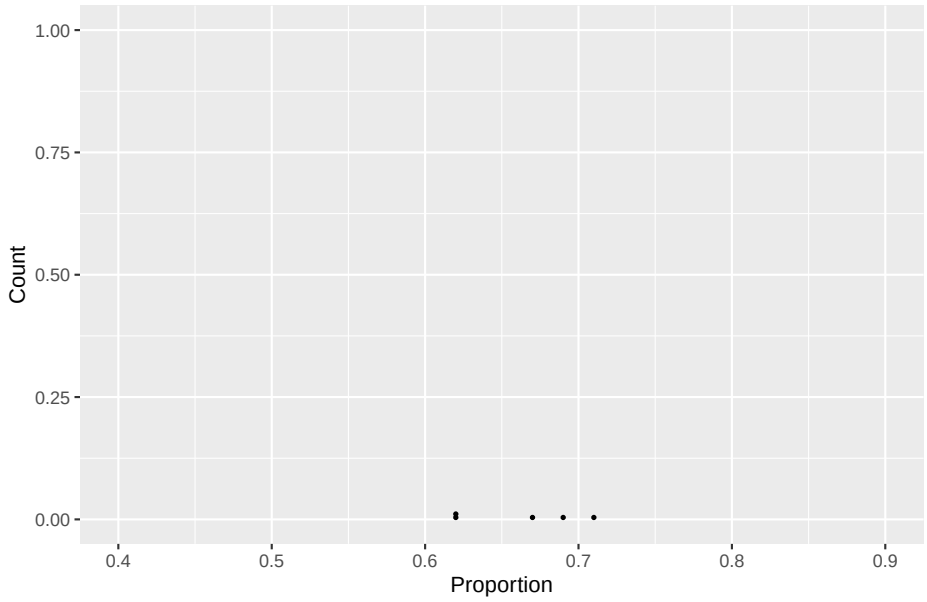


(b). Generate 5 random samples of size $n = 100$ and plot the sample proportions.

```
# 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 = sample.prop5)) +
  geom_dotplot(dotsize=0.25, stackratio=0.9, binwidth=0.01 )+
  ggtitle("") +  xlab("Proportion") + ylab("Count")+
  scale_x_continuous(limits = c(0.4, 0.9))+
  theme(plot.title = element_text(hjust = 0.5))
```

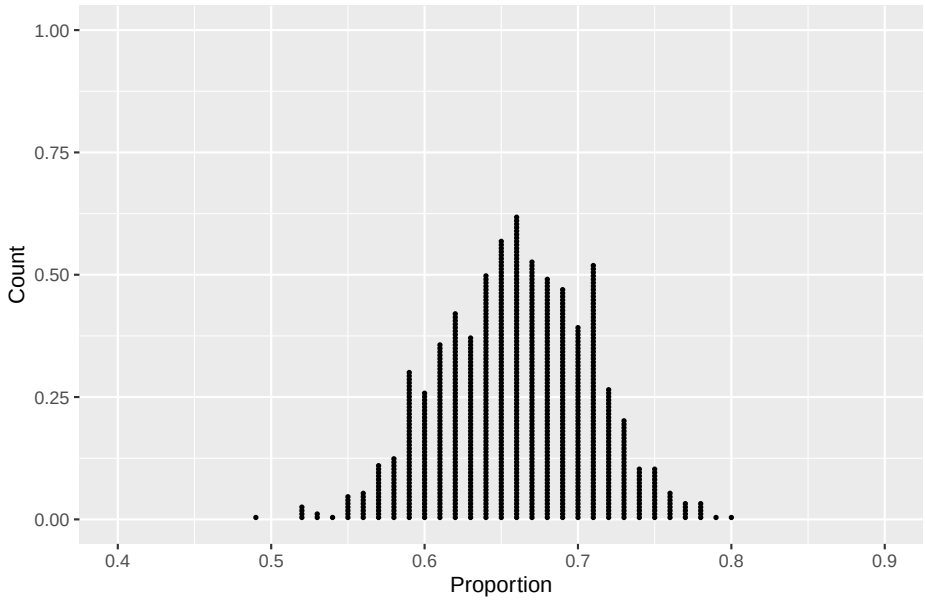


(c). Generate 1000 random samples of size $n = 100$ and plot the sample proportions.

```
# Generate 1000 samples
sample1000 <- rbinom(n = 1000, size = n.size, p = pop.prop)
sample.prop1000 <- sample1000/n.size

data <- data.frame(x = sample.prop1000)

ggplot(data, aes(x = sample.prop1000)) +
  geom_dotplot(dotsize=0.25, method = "histodot", stackratio=0.9, binwidth=0.01) +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0.4, 0.9))+
  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.prop1000)
```

```
[1] 0.65962
```

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

Click for answer

Answer: About 0.03, 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.prop1000)
```

```
[1] 0.0483176
```

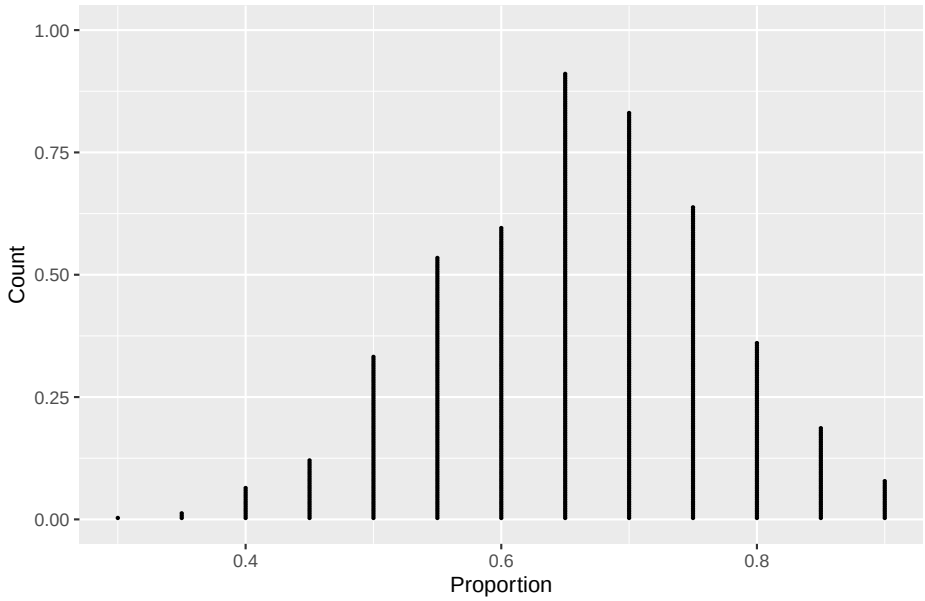
(d). Repeat part(c) with sample size 20 instead of 100. Generate 1000 samples.

```
# Generate 1000 samples
n.size <- 20

sample1000 <- rbinom(n = 1000, size = n.size, p = pop.prop)
sample.prop1000 <- sample1000/n.size

data <- data.frame(x = sample.prop1000)

ggplot(data, aes(x = sample.prop1000)) +
  geom_dotplot(dotsize=0.225, method = "histodot", stackratio=0.8, binwidth=0.01) +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0.3, 0.9))+
  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.prop1000)
```

```
[1] 0.65885
```

```
sd(sample.prop1000)
```

```
[1] 0.1086093
```

(e). Now suppose the population proportion is $p = 0.90$ instead of $p = 0.66$ in part (e). Keep n.size=20.

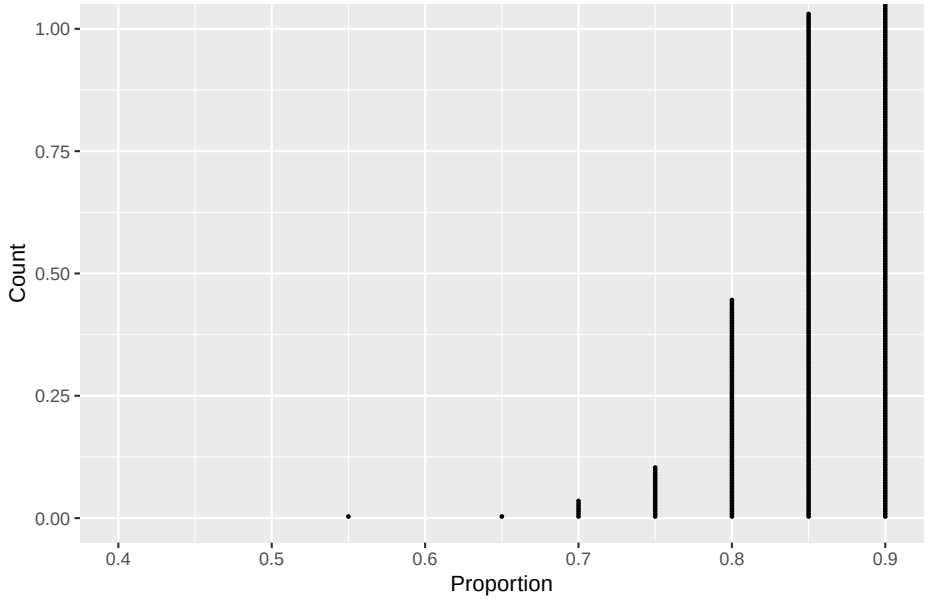
```
# Generate 1000 samples

pop.prop <- 0.90
n.size <- 20
n.size <- 20
```

```
sample1000 <- rbinom(n = 1000, size = n.size, p = pop.prop)
sample.prop1000 <- sample1000/n.size

data <- data.frame(x = sample.prop1000)

ggplot(data, aes(x = sample.prop1000)) +
  geom_dotplot(dotsize=0.21, method = "histodot", stackratio=0.8, binwidth=0.01) +
  ggtitle("") +  xlab("Proportion") + ylab("Count") +
  scale_x_continuous(limits = c(0.4, 0.9))+
  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.prop1000)
```

```
[1] 0.9028
```

```
sd(sample.prop1000)
```

```
[1] 0.06466329
```

7.2.2 Example 4: Simulation for a Sample Mean

We'll look at sampling movies from the population of 134 Hollywood movies made in 2011 and measuring their budget (millions of dollars).

```
# import dataset
library(Lock5Data)
movies <- HollywoodMovies2011
```

(a). What is the population mean of the `Budget`?

```
# r-code
mean(movies$Budget, na.rm = TRUE)
```

```
[1] 53.48134
```

(b). Generate a random sample of size $n = 10$ and plot the sample proportion.

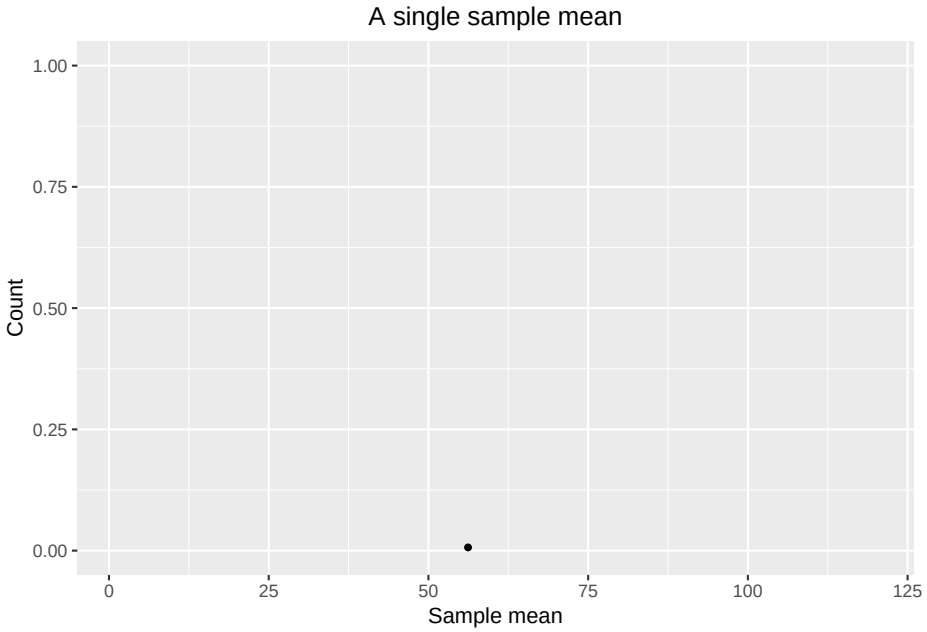
```
# define a data frame
n.size <- 10

Budget <- movies$Budget[!is.na(movies$Budget)] # remove NAs

sample1 <- sample(Budget, size = n.size)
sample.mean1 <- mean(sample1)

mydata <- data.frame(x = sample.mean1)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = sample.mean1)) +
  geom_dotplot(dotsize=1, stackratio=0.9, binwidth=1) +
  ggtitle("A single sample mean") +  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  theme(plot.title = element_text(hjust = 0.5))
```



(c). Generate 5 random samples of size $n = 10$ and plot the sample means.

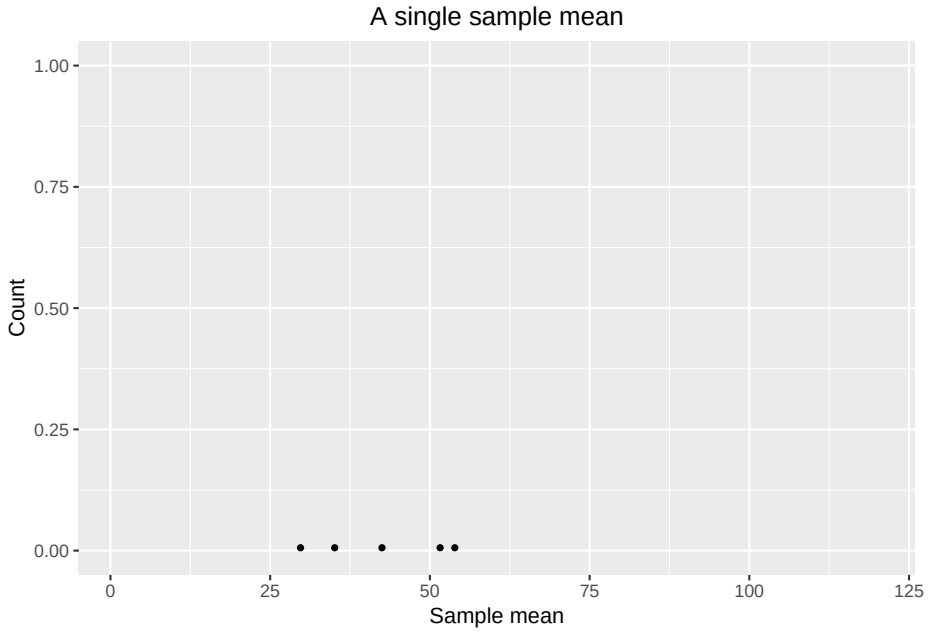
```
n.size <- 10
n.rep <- 5

Budget <- movies$Budget[!is.na(movies$Budget)] # remove NAs

sample5 <- lapply(1:5, function(i) sample(Budget, size = n.size))
sample.mean5 <- lapply(sample5, function(x) mean(x))
sample.mean5 <- unlist(sample.mean5)

mydata <- data.frame(x = sample.mean5)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = sample.mean5)) +
  geom_dotplot(dotsize=0.9, stackratio=0.9, binwidth=1) +
  ggtitle("A single sample mean") +  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  theme(plot.title = element_text(hjust = 0.5))
```



(d). Generate 1000 random samples of size $n = 10$ and plot the sample means.

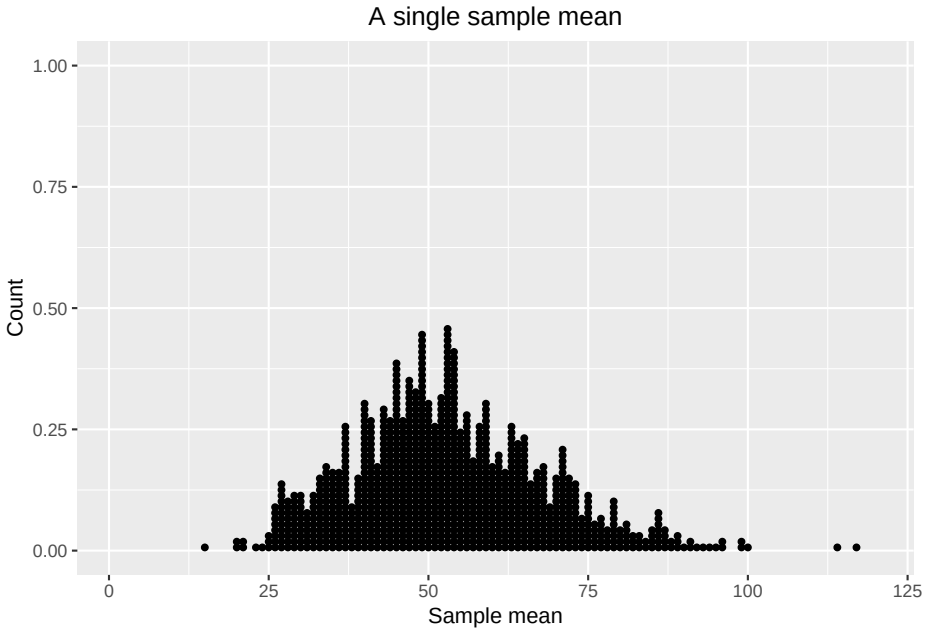
```
# Generate 1000 samples
n.size <- 10
n.rep <- 1000

Budget <- movies$Budget[!is.na(movies$Budget)] # remove NAs

sample1000 <- lapply(1:n.rep, function(i) sample(Budget, size = n.size))
sample.mean1000 <- lapply(sample1000, function(x) mean(x))
sample.mean1000 <- unlist(sample.mean1000)

mydata <- data.frame(x = sample.mean1000)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = sample.mean1000)) +
  geom_dotplot(dotsize=1, method = "histodot", stackratio=0.9, binwidth=1) +
  ggtitle("A single sample mean") +  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  theme(plot.title = element_text(hjust = 0.5))
```



Question: What does each dot represent?

Click for answer

Answer: A sample mean budget from a sample of n=10

Question: What is the shape of your sampling distribution?

Click for answer

Answer: Slightly right skewed.

Question: Where is your distribution centered?

Click for answer

Answer: About $53 million, which is the population mean budget.

```
mean(movies$Budget, na.rm = TRUE)
```

```
[1] 53.48134
```

Question: The distribution should be centered at the population mean. Verify that the distribution is centered around the population mean, $\mu = 53.48$.

Click for answer

Answer: It is very close to the population mean.

```
# r-code
mean(sample.mean1000)
```

```
[1] 53.12677
```

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

Click for answer

Answer: About 15 million.

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 is 14.80 million quite cose to our previous informed guess.

```
# r-code
sd(sample.mean1000)
```

```
[1] 15.0577
```

(e). Repeat part(d) with sample size 50 instead of 10. Generate 1000 samples.

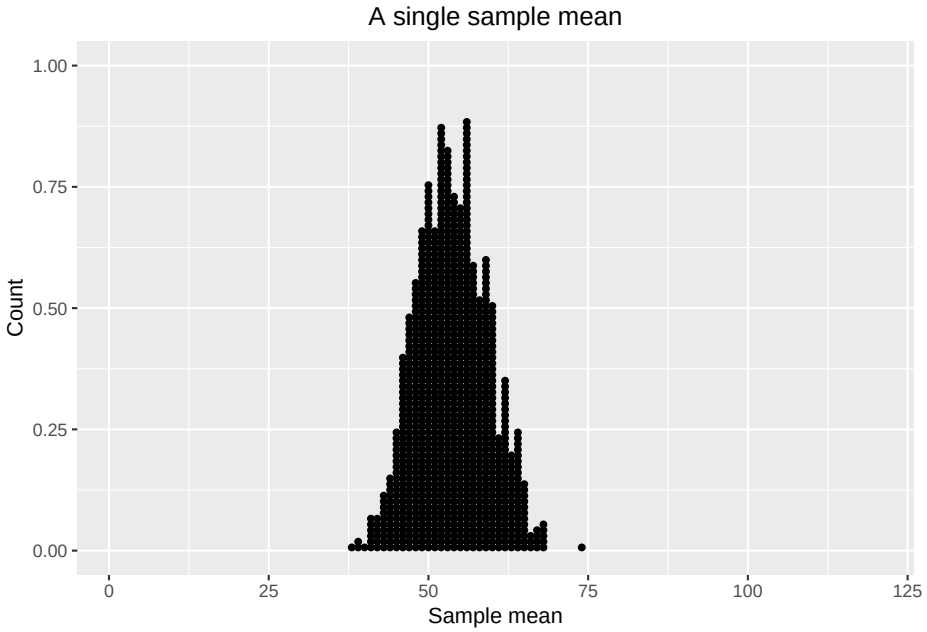
```
# Generate 1000 samples
n.size <- 50
n.rep <- 1000

Budget <- movies$Budget[!is.na(movies$Budget)] # remove NAs

sample1000 <- lapply(1:n.rep, function(i) sample(Budget, size = n.size))
sample.mean1000 <- lapply(sample1000, function(x) mean(x))
sample.mean1000 <- unlist(sample.mean1000)

mydata <- data.frame(x = sample.mean1000)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = sample.mean1000)) +
  geom_dotplot(dotsize=1, method = "histodot", stackratio=0.9, binwidth=1) +
  ggtitle("A single sample mean") +  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(plot.title = element_text(hjust = 0.5))
```



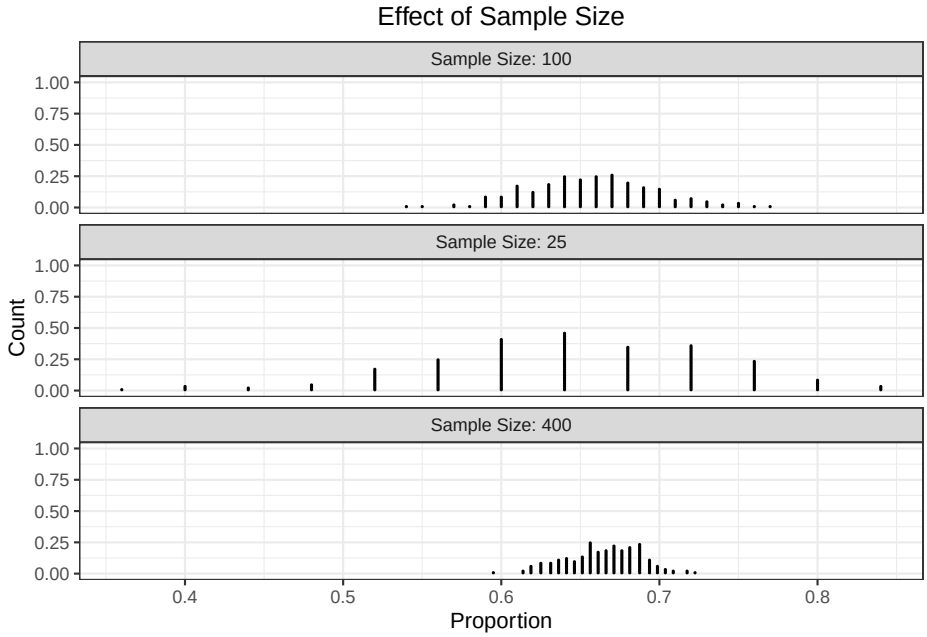
Question: Is this sampling distribution more or less symmetric compared to the distribution when $n = 10$?

Click for answer

Answer: The distribution is more symmetric with n=50 than when n=10.

7.2.3 Example 5: Effect of sample size

Let's investigate the effect of sample size in the sampling distribution using the same setting as in Exercise 1 with $p = 0.66$. The following are three sampling distributions corresponding to different sample sizes.



Question: What happens if we increase the sample size?

Click for answer

Answer: When we increase the sample size, the variability of the sampling distribution becomes smaller.

Question: Estimate the standard error of each and verify your answer to the previous question.

Click for answer

Answer: The standard errors are

```
sd(data.size.25$x)
```

```
[1] 0.09011439
```

```
sd(data.size.100$x)
```

```
[1] 0.04093137
```

```
sd(data.size.400$x)
```

```
[1] 0.02311007
```

As the sample size increases, the variability as measured by the standard error of the sampling distribution does indeed decrease.

7.2.4 Example 6: Bootstrap Sampling

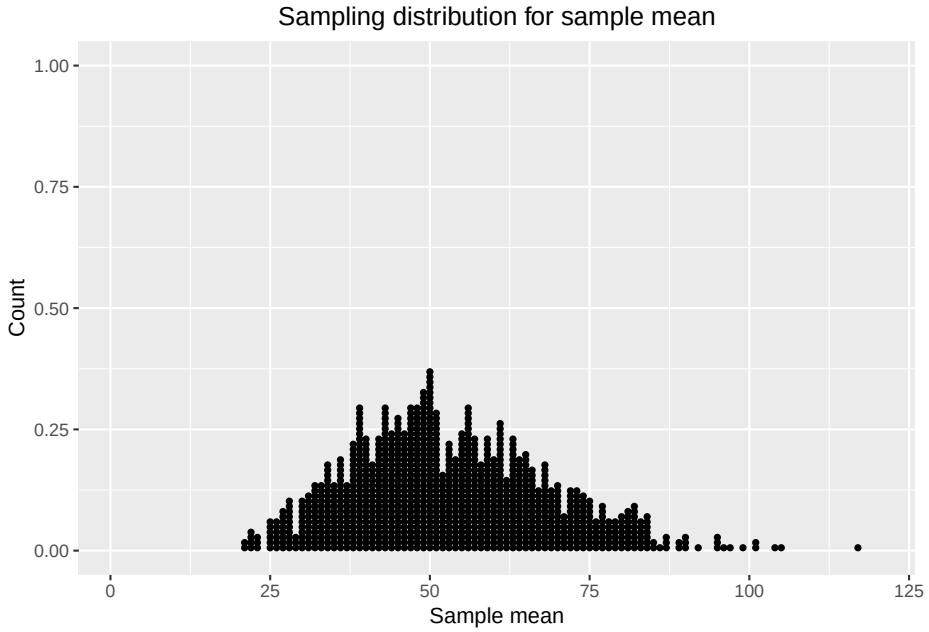
```
# Movies Example Again!
Budget <- movies$Budget[!is.na(movies$Budget)]

# Bootstrap samples
n.size <- 10
boot.sample1 <- sample(Budget, 10, replace = TRUE) # sampling with replacement

n.rep <- 1000
boot.sample1000 <- lapply(1:n.rep, function(i) sample(Budget, 10, replace = TRUE))
boot.samplemean1000 <- lapply(boot.sample1000, function(x) mean(x))
boot.samplemean1000 <- unlist(boot.samplemean1000)
```

```
# Plot the bootstrap distribution
mydata <- data.frame(x = boot.samplemean1000)

# Plot a dot plot of the sample proportion
ggplot(mydata, aes(x = boot.samplemean1000)) +
  geom_dotplot(dotsize=0.9, stackratio=0.9, binwidth=1, method = "histodot") +
  ggtitle("Sampling distribution for sample mean") +  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  theme(plot.title = element_text(hjust = 0.5))
```



(a). Compare the center/spread/shape of the bootstrap distribution to the distribution computed in Ex. 4 (d). Answer all the questions in Ex. 4(d).

Click for answer

Answer: The shape/center and variability of this bootstrap distribution is very similar to that of Ex 4 (d)

```
mean(mydata$x)
```

```
[1] 53.23545
```

```
sd(mydata$x)
```

```
[1] 15.41347
```

Chapter 8

Class Activity 8

8.1 Your Turn 1

8.1.1 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

(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
```

(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.

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

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

```
 [1] 147  23  93 185  52  85  23 185  23  93
```

(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).

8.1.2 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

(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 $s = 13.798$.

(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.

(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!

(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.

(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.

(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.

8.1.3 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".

(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 $\hat{p} = 0.59$.

(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)

(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).

(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.

(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.

(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%.

8.1.4 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".

(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 $\hat{p}_{Dem} - \hat{p}_{Rep} = 0.79 - 0.38 = 0.41$

(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.

(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).

(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.

(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.

(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.

8.1.5 Example 5: Statkey Body Temperature

Is normal body temperature really 98.6° F? A sample of body temperature for 50 healthy individuals was taken. Find this dataset in StatKey under "Confidence Interval for a Mean."

(a). What is the sample mean? What is the sample standard deviation? Use correct notation for each

Click for answer

Answer: $\bar{x} = 98.26$ and $s = 0.765$.

(b). Generate a bootstrap distribution, using at least 1000 simulated statistics. What is the standard error?

Click for answer

Answer: SE ≈ 0.108. Answers will vary slightly with different simulations (see output below).

(c) Use the standard error to find a 95% confidence interval. Show your work. Is 98.6 in the interval?

Click for answer

Answer:

$$\bar{x} \pm 2 * SE$$
$$98.26 \pm 2(0.108)$$
$$(98.04, 98.48)$$

We see that 98.6 is not on the interval.

8.1.6 Example 6. Bootstrap in R using Hollywood 2011 dataset!

We'll look at sampling movies from the population of 134 Hollywood movies made in 2011 and measuring their budget (millions of dollars). Construct a bootstrap sampling distribution for budgets (in millions of dollars) of all movies to come out of Hollywood in 2011, using samples of size n = 50.

```
# import dataset
movies <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/HollywoodMovies20
```

(a) Generate 1 sample of size 50 with replacement from the `Budget` variable. If there are any `NA` values, they should be removed first.

```
# remove the NA values
Budget <- movies$Budget[!is.na(movies$Budget)]
```

```
# Bootstrap samples
n.size <- 50
boot.sample1 <- sample(Budget, size = n.size, replace = TRUE) # sampling with replacement
```

(c) Generate 1000 samples of size 50 with replacement from the redefined `Budget` variable in part (a). There are many methods to do this. We will use `lapply` function to do this simulation faster. Using `lapply` we can apply functions to a list or vector.

```
n.rep <- 1000
# replicate the sampling with replacement 1000 times
boot.sample1000 <- lapply(1:n.rep, function(x) sample(Budget, size = n.size, replace =

# Calculate the mean of each resample
boot.samplemean1000 <- lapply(boot.sample1000, function(x) mean(x))

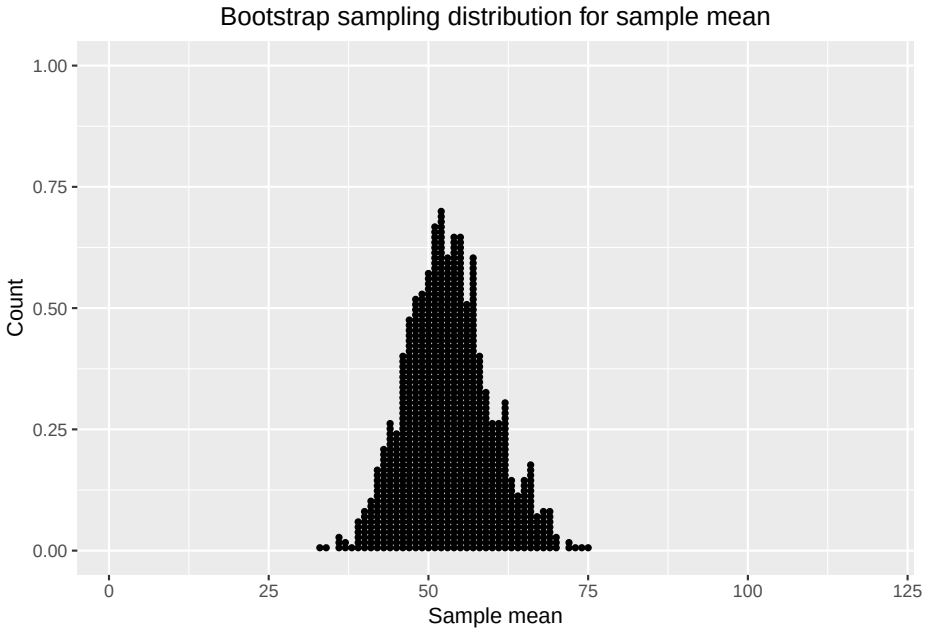
# Transform the list back to a vector for further computations
boot.samplemean1000 <- unlist(boot.samplemean1000)
```

(d) Make a dotplot of the 1000 sample means calculated in part (c). The function to do this in `ggplot2` is `geom_dotplot`. There are two methods for binning the data values. `dotdensity` is the default option for dot-density binning and `histodot` is for fixed bin width like a histogram.

```
# Plot the bootstrap distribution

boot.samples <- data.frame(samples = boot.samplemean1000) # define a data frame

# Plot a dot plot of the sample proportion
ggplot(boot.samples, aes(x = samples)) +
  geom_dotplot(dotsize=0.9, stackratio=0.9, binwidth=1, method = "histodot") +
  xlab("Sample mean") + ylab("Count")+
  scale_x_continuous(limits = c(1,120))+
  ggtitle("Bootstrap sampling distribution for sample mean") +
  theme(plot.title = element_text(hjust = 0.5))
```



8.1.7 Example 7: 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.

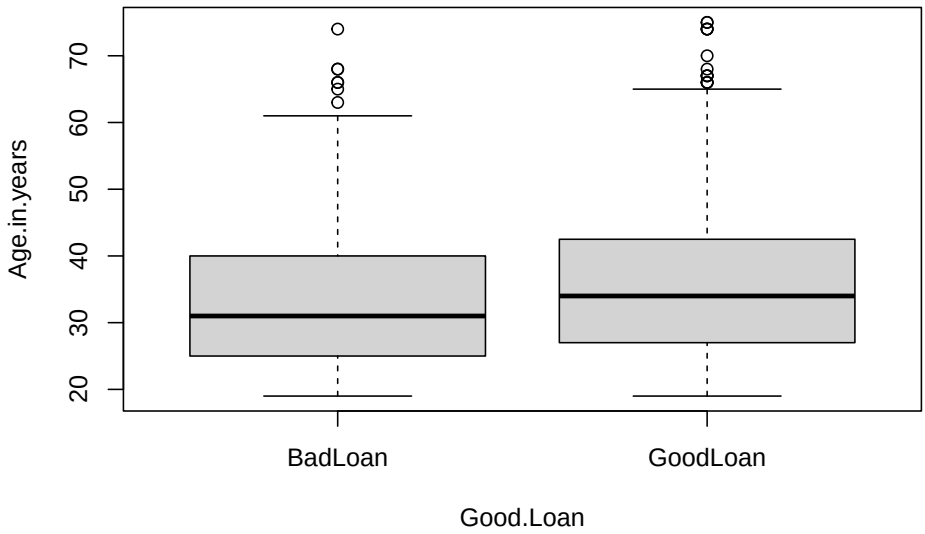
```
credit <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/CreditData.csv")
table(credit$Good.Loan)
```

```
BadLoan GoodLoan
    300      700
```

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




```
tapply(credit$Age.in.years, credit$Good.Loan, mean)
```

```
 BadLoan GoodLoan
33.96333 36.22429
```

- 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.

(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:

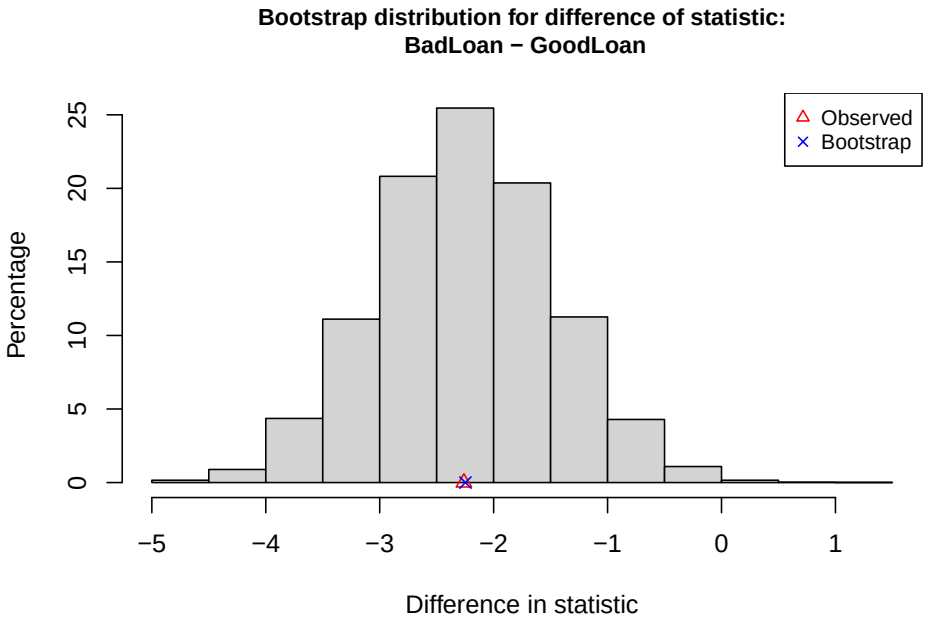
```
library(CarletonStats)
boot(Age.in.years ~ Good.Loan, data=credit)
```

```
    ** Bootstrap interval for difference of statistic

 Observed difference of statistic:  BadLoan - GoodLoan =  -2.26095
 Mean of bootstrap distribution: -2.24781
 Standard error of bootstrap distribution: 0.77531

 Bootstrap percentile interval
     2.5%      97.5%
-3.7553214 -0.7066548

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



- 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
```

(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.

8.1.8 Example 8 : 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).

(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
```

To make shorter names describing these two outcomes, we can use the `levels` command on the factor variable `Telephone`. Here we see what the original levels are for this variable:

```
credit$Telephone <- factor(credit$Telephone)
levels(credit$Telephone)
```

```
[1] "none"
[2] "yes, registered under the customers name"
```

This shows us the (vector) of two names. We can assign new, shorter names to this variable:

```
levels(credit$Telephone) <- c("no", "yes")
table(credit$Telephone)
```

```
 no yes
596 404
```

Now we have the same data, just coded with different names.

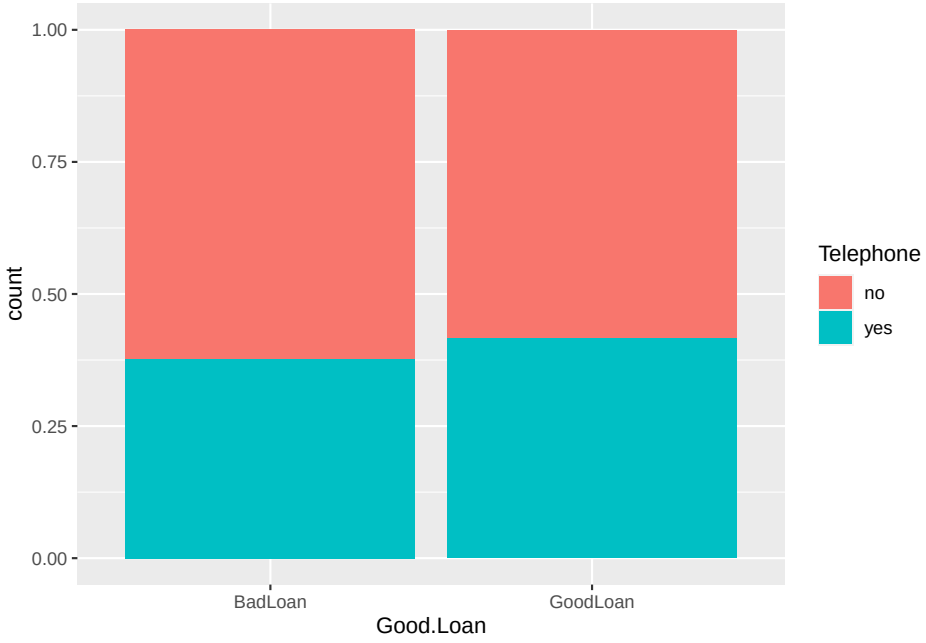
(b). Phone rate by default type

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
```

(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, registered under the customers name",
head(credit[,c("Telephone", "Telephone_binary")])
```

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

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
1000
```

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

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

```
[1] 0
```

```
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!

(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.

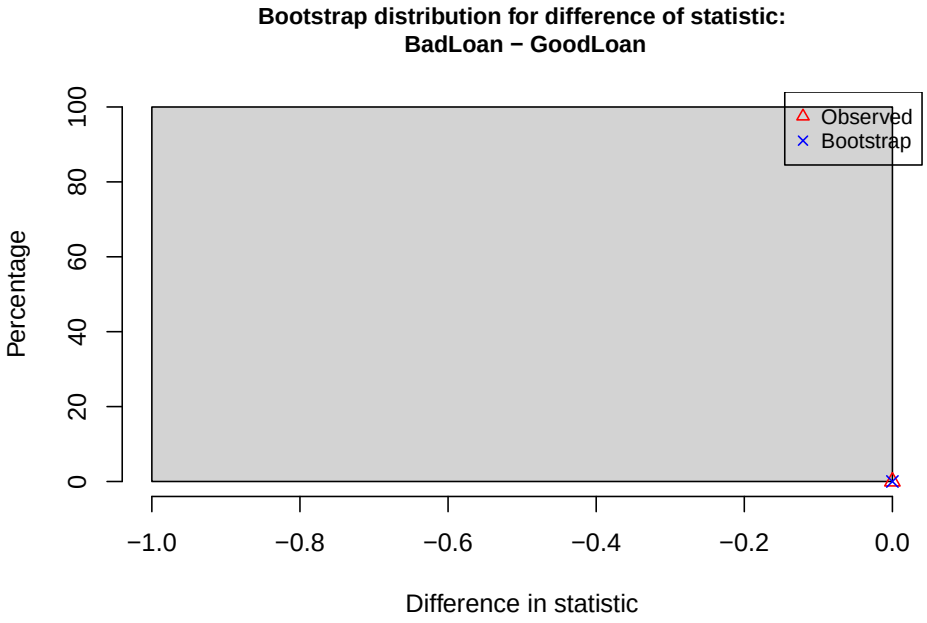
```
boot(Telephone_binary ~ Good.Loan, data=credit)
```

```
   ** Bootstrap interval for difference of statistic

Observed difference of statistic:  BadLoan - GoodLoan =  0
Mean of bootstrap distribution: 0
Standard error of bootstrap distribution: 0

Bootstrap percentile interval
2.5% 97.5%
   0     0

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



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
```

(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 compared to the percentage 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 9

Class Activity 9

9.0.1 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.

9.0.2 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.

9.0.3 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.

9.0.4 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 10

Class Activity 10

10.0.1 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.

10.0.2 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!)

10.0.3 Example 3: Guess the inference method!

For each question below, indicate whether it is best assessed with a confidence interval, hypothesis test, or whether statistical inference is not needed to answer it. If inference is used, define the population parameter(s) of interest and indicate your hypotheses if a test is appropriate. Use correct notation.

(a). What proportion of US adults support gun control?

Click for answer

Answer: Estimating a population parameter with a confidence interval: p: proportion of US adults who support gun control

(b). Does the proportion of people who support gun control differ between males and females?

Click for answer

Answer: Testing the claim of a difference:

$$p_f : \text{proportion of females who support gun control}$$
$$p_m : \text{proportion of males who support gun control}$$

$$H_0 : p_f = p_m$$
$$H_a : p_f \neq p_m$$

(c). What proportion of this class supports gun control?

Click for answer

Answer:

Neither, just collect data on the entire population (class) to answer this question.

(d). How much more do men earn, on average, compared to women in the US?

Click for answer

Answer: How much means an estimate: confidence interval for $\mu_m - \mu_f$

$$\mu_f : \text{mean income (yearly) of US females}$$
$$\mu_m : \text{mean income (yearly) of US males}$$

(e). What proportion of Minnesota voters in the 2012 election voted for President Obama?

Click for answer

Answer: Neither, this number is computed from voting records.

(f). Is a higher rate of cricket chirping associated with higher summer night temps?

Click for answer

Answer:

Test a claim about an association:

$$H_0 : \rho = 0 \text{ or } \beta_1 = 0$$

$$H_a : \rho > 0 \text{ or } \beta_1 > 0$$

$$\rho : \text{correlatio nbetween temp and chirp rate}$$

$$\beta_1 : \text{change in temp for a unit increase in chirp rate}$$

10.0.4 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.

10.0.5 Example 2 Revisited:

(a). If the results of the test comparing sleep and caffeine for memory are statistically significant, what does that mean in terms of sleep, caffeine, and memory?

Click for answer

Answer: We can state that there is a difference between sleep and caffeine in their effectiveness at helping word recall.

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

Click for answer

Answer: It means the sample results are inconclusive and we can't tell if there is a difference. Results might just be random chance.

Chapter 11

Class Activity 11

Midterm Review !!

Chapter 12

Class Activity 12

Midterm !!

Chapter 13

Class Activity 13

13.1 Example 1: 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. The null and alternative hypotheses for testing whether people have ESP are $H_0 : p = 0.2$ and $H_A : p > 0.2$ where p is the true proportion correct guesses. To test these hypotheses, we try this $n = 10$ times and get 3 correct guesses.

13.1.0.1 (a) Explain how to generate a randomization distribution for $\hat{p}$, the sample proportion of correct guesses, that is consistent with $H_0 : p = 0.2$.

Click for answer

Answer: To mimic one random guess, assuming no ESP, we could take 4 black cards and 1 red card and randomly select one card. The red card would be a correct guess. Repeat this a total of 10 times and compute the sample proportion of correct guesses. Plot the sample proportion on a dotplot and repeat lots more times.

13.1.0.2 (b) Navigate to the Statkey website.

Select the **Test for Single Proportion** option under **Randomization Hypothesis Tests**. Click **Edit Data** and enter a **count** of 3 and **sample size** of 10. Then select the **Null Hypothesis** proportion **p** and change its value to **0.20**.

- **Generate 1 Sample** from this null randomization distribution. How many correct guesses were obtained in this sample? Repeat this a couple of times.
- **Generate 1000 Samples** a couple of times. How unusual is getting at least 3 correct guesses in 10 tries?

13.1.0.3 (c) Compute the randomization p-value

Select the **Right Tail** button at the top of the plot. Change the x-axis value to $\hat{p} = 0.3$. What is the p-value: the proportion of resampled $\hat{p}$ values are 0.30 or above?

Click for answer

Answer: The p-value is about 31%

13.1.0.4 (d) Interpret + Conclusion

Interpret the p-value. Does the p-value support the alternative hypothesis (do you think 3 correct out of 10 tries are statistically significant results) or is it inconclusive? Explain.

Click for answer

Answer: In about 31% of all samples with 10 attempts, we would get at least 3 correct guesses, just by chance, if ESP does not exist. It is inconclusive; the data we observed is not that unusual if ESP does not exist.

13.2 Example 2: Which P-value shows more evidence?

Using the randomization distribution below to test $H_0 : \rho = 0$ vs. $H_A : \rho > 0$.

13.2.0.1 (a) Match the p-value and sample statistic

Match the sample correlation and p-values given below, shading the area on the randomization distribution that corresponds to each sample correlation/p-value combo.

- Sample correlations: $r = 0.1, r = 0.3, r = 0.5$

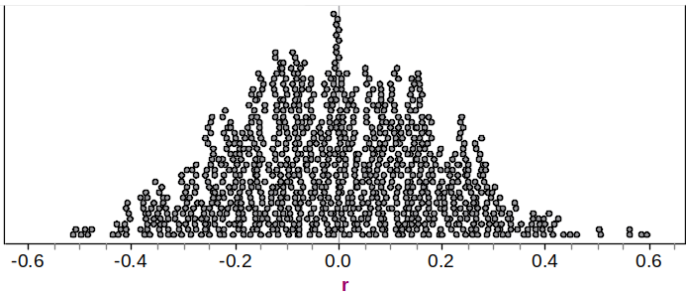


Figure 13.1: Example 2

- P-values: $0.005, 0.15, 0.35$

Click for answer

Answer:: $r = 0.1$ and $p - value = 0.35$; $r = 0.3$ and $p - value = 0.15$; $r = 0.5$ and $p - value = 0.005$;

13.2.0.2 (b) Which sample correlation/p-value combo shows the most evidence for the alternative hypothesis?

Click for answer

Answer: The smaller the p-value and further the sample correlation is from 0, the stronger the evidence

13.3 Example 3: 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.

13.3.0.1 (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

13.3.0.2 (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?

13.3.0.3 (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?

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 $2(0.022) = 0.044$.

13.3.0.4 (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

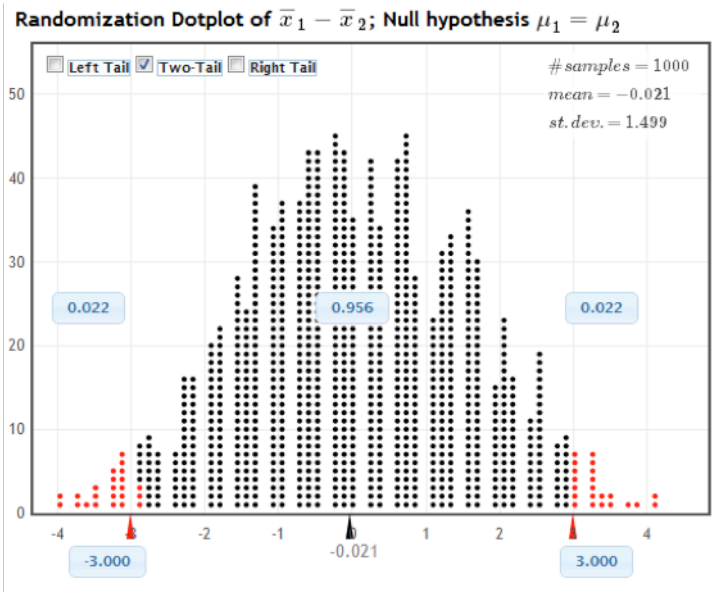


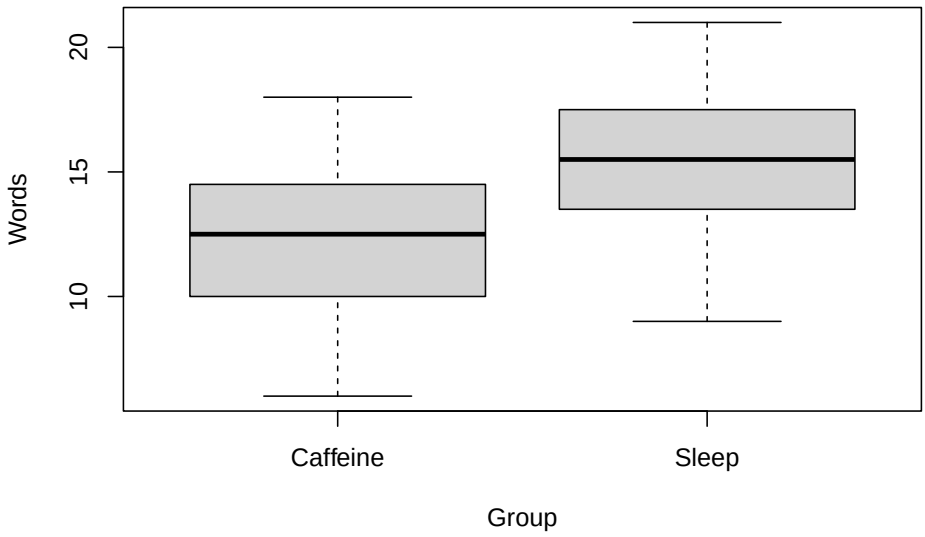
Figure 13.2: Example 3

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.

13.3.0.5 (e) Redo in Rstudio

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

```
wordData <- read.csv("http://math.carleton.edu/Stats215/Textbook/SleepCaffeine.csv")
boxplot(Words ~ Group , data=wordData)
```



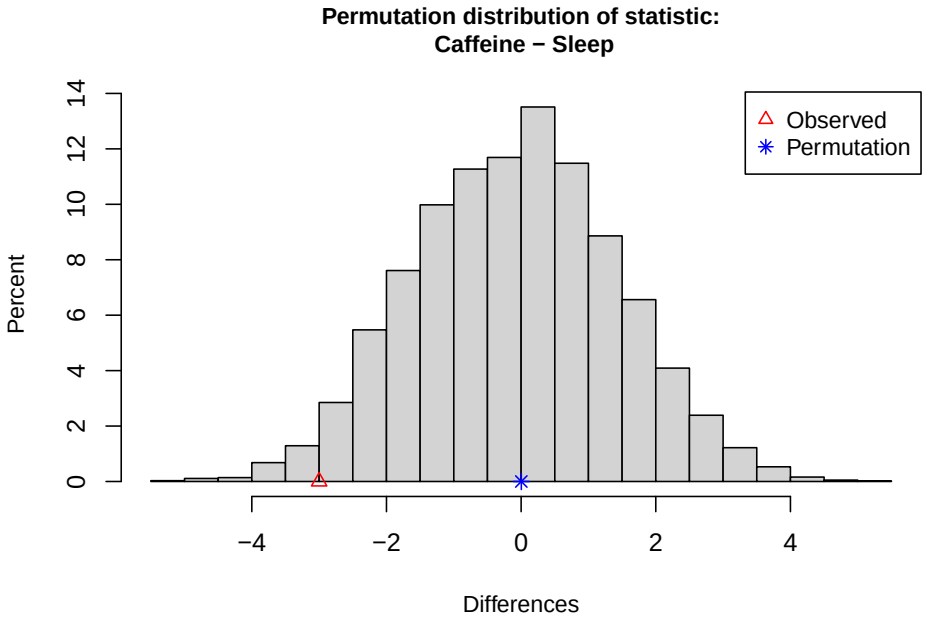

```
tapply(wordData$Words, wordData$Group, summary)
```

```
$Caffeine
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   10.00   12.50   12.25   14.25   18.00

$Sleep
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   13.75   15.50   15.25   17.25   21.00
```

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: 0.00277
Standard error of permutation distribution: 1.50432
P-value:  0.0452

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

- 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.

13.4 Example 4: 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:

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

13.4.0.1 (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).

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

```
  X        Institution  Res NonRes  Diff
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
```

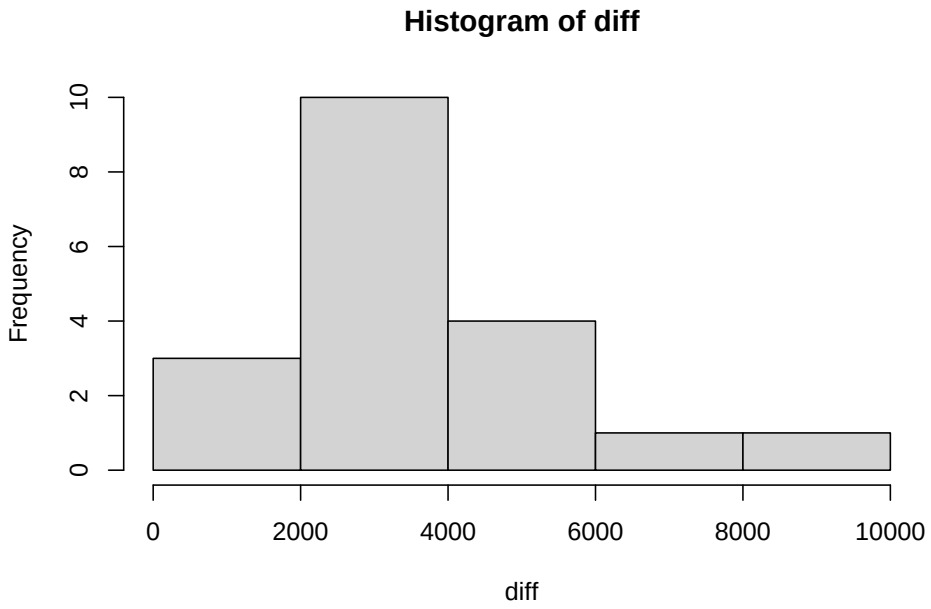
13.4.0.2 (b) Permutation test for paired data

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
```

```
hist(diff)
```



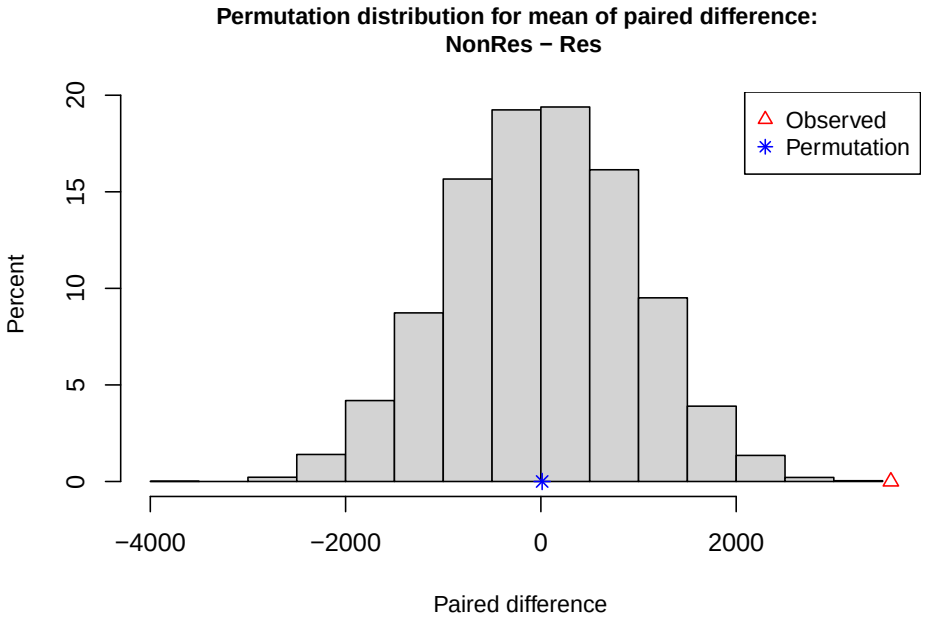
- What is the average difference in tuition costs?

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: 12.87129
Standard error of permutation distribution: 945.1306
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).

13.5 Example 5: Evaluating Drugs to Fight Cocaine Addition

In a randomized experiment on treating cocaine addiction, 48 cocaine addicts who were trying to quit were randomly assigned to take either desipramine (a new drug), or Lithium (an existing drug). The response variable is whether or not the person relapsed (which means the person was unable to break out of the cycle of addiction and returned to using cocaine.) **We are testing to see if desipramine is better than lithium at treating cocaine addiction.** The results are shown in the two-way table.

\	Relapse	No Relapse	total
Desipramine	10	14	24
Lithium	18	6	24

13.5.0.1 (a) Using p_D for the true proportion of desipramine users who relapse and p_L for the true proportion of lithium users who relapse, write the null and alternative hypotheses.

Click for answer

Answer: $H_0 : p_D - p_L = 0$ vs. $H_A : p_D - p_L < 0$

13.5.0.2 (b) Compute the appropriate sample statistic needed to assess the hypotheses above.

Click for answer

Answer: We see that $\hat{p}_D = \frac{10}{24} = 0.417$ and $\hat{p}_L = \frac{18}{24} = 0.75$ so we have $\hat{p}_D - \hat{p}_L = 0.417 - 0.75 = -0.333$. Be sure to compute the **difference** since we need one number (observed difference) to test the hypotheses, not two separate numbers. You could also compute the difference as $L - D$ and get $+0.333$.

13.5.0.3 (c) How might we compute a randomization sample for this data?

Click for answer

Answer: Since drug doesn't matter, we combine all 48 patients together and see that 28 relapsed and 20 didn't. To see what happens by random chance, we randomly divide them into two groups and compute the difference in proportions of relapses between the two groups. The difference in proportions is the statistic.

13.5.0.4 (d) Navigate to the Statkey website.

Select the **Test for Difference in Proportions** option under **Randomization Hypothesis Tests**. Click **Edit Data** and let Group 1 be "Desipramine" and 2 be "Lithium", enter relapse **counts** of 10 and 18 and **sample sizes** of 24. Check that the null hypothesis matches yours in (a). Generate a couple thousand samples. Describe the resulting distribution. Where is it centered?

Click for answer

Answer: The resulting distribution, shown in Figure 1, will be bell-shaped and centered at the value from the null hypothesis, which is zero.

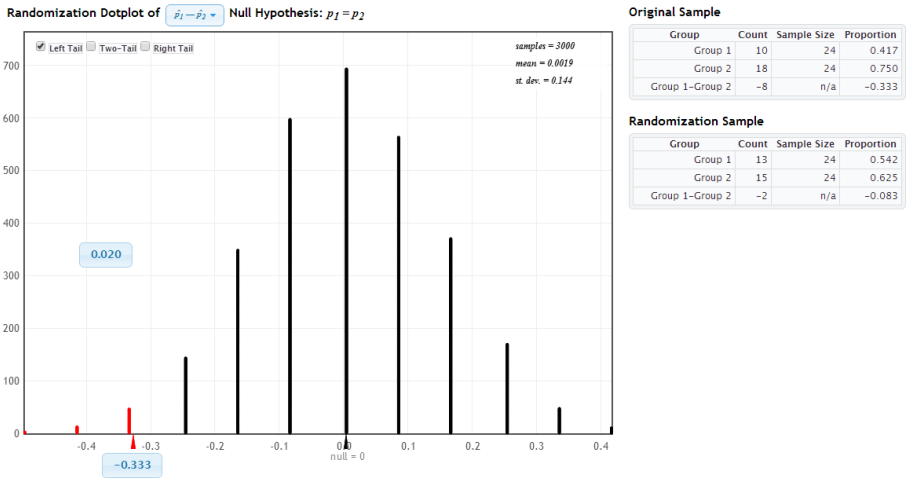


Figure 13.3: Example 1d

13.5.0.5 (e) Compute and interpret the p-value for this test.

Click for answer

Answer: This is a left-tail test when computing the difference as D - L, and we see on StatKey that the p-value (proportion of randomization samples with a difference -.333 or smaller) is about 2 (Figure 1). About 2% of the time we would see at least 33% fewer relapse cases using despramine than lithium just due to chance if there was no difference in the relapse rates of the two treatments.

Note the two key features of this "in context" interpretation of 2%: it assumes that the null is true (no treatment difference) and it uses the observed statistic (data) used to compute the p-value (rate of despramine relapse is .33 below the rate of lithium).

13.5.0.6 (f) Make a formal decision (reject or not) using a 5% significance level, then restate your conclusion in context for the problem (do not use words like "reject" or "hypothesis").

Click for answer

Answer: We reject the null hypothesis since the p-value of 2% is less than 5%. We can conclude that despramine is better at helping people kick the cocaine habit. Note the "in context" conclusion: Just state your conclusion in english, no need to talk about the value of the p-value or "just by chance."

13.5.0.7 (h) Use Statkey to compute and interpret a 95% bootstrap confidence interval for the difference in the relapse proportion for the two treatments. Explain how this CI agrees with your test conclusion in (f).

Click for answer

Answer: I am 95% confident that the relapse rate for despramine will be between 8.3 to 58.3 percent less than the relapse rate for lithium. This completely agrees with the test conclusion that despramine is a better treatment for cocaine addiction. (Figure 2 shows the bootstrap distribution that is centered at the sample difference of -0.333.)

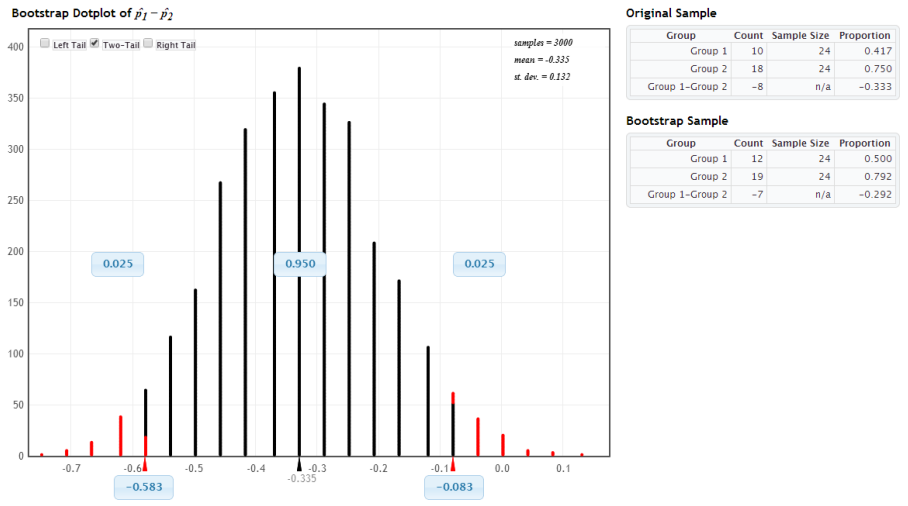


Figure 13.4: Example 1h

Chapter 14

Class Activity 14

14.1 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)
```

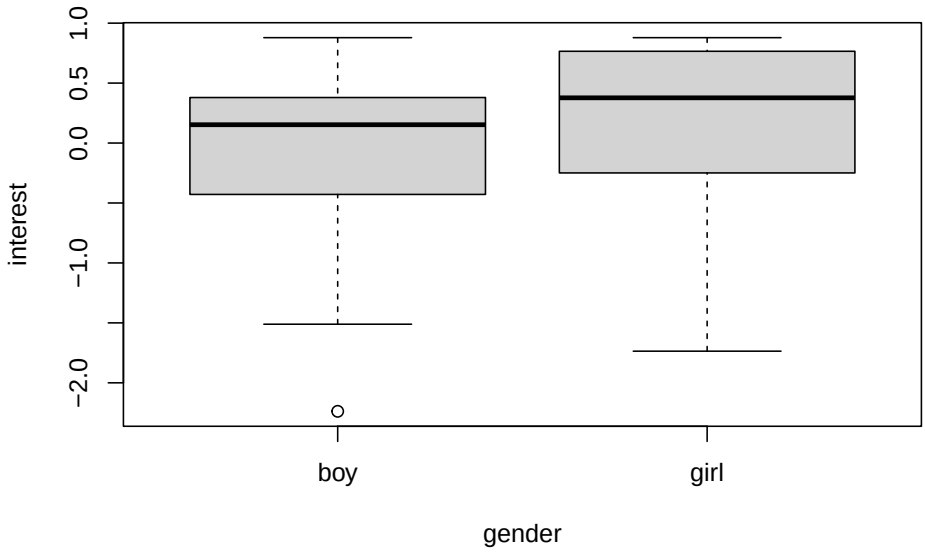
```
    study subj gender   age    interest race     race2
1 Study 4   65   girl age 6  0.37953534    5     white
2 Study 4   66   girl age 6 -0.78071539    5     white
3 Study 4   67   girl age 6 -0.47631654    5     white
4 Study 4   68   girl age 6 -0.07234632    5     white
5 Study 4   69    boy 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
```

14.1.0.1 (a) Interest in 5 year olds - test

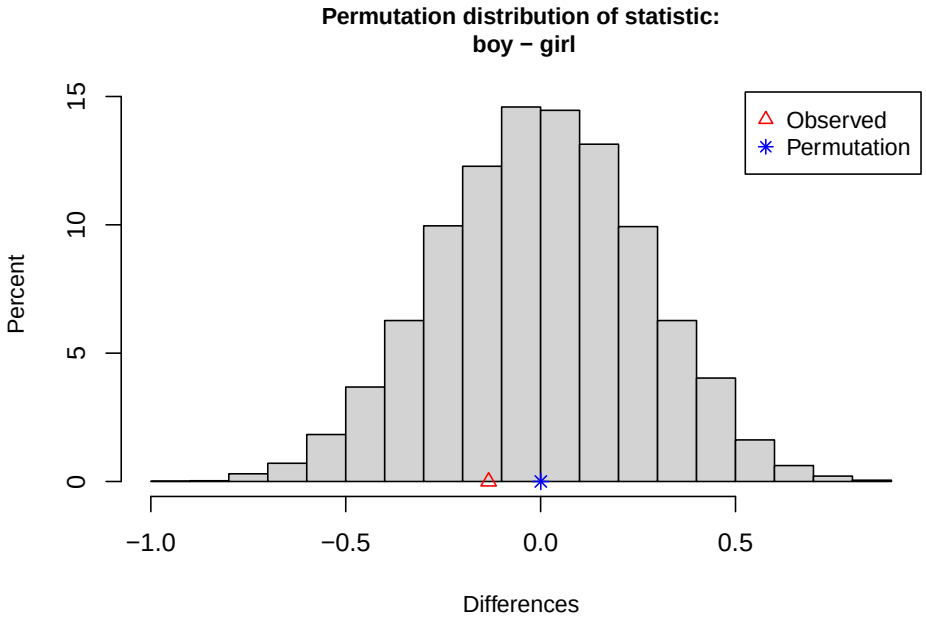
Recall the comparison of mean interest level in 5 year old boys and girls. Generate the randomization distribution for this test:

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

```
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: 0.00049
Standard error of permutation distribution: 0.26175
P-value:  0.606

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

- 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
```

14.1.0.2 (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?

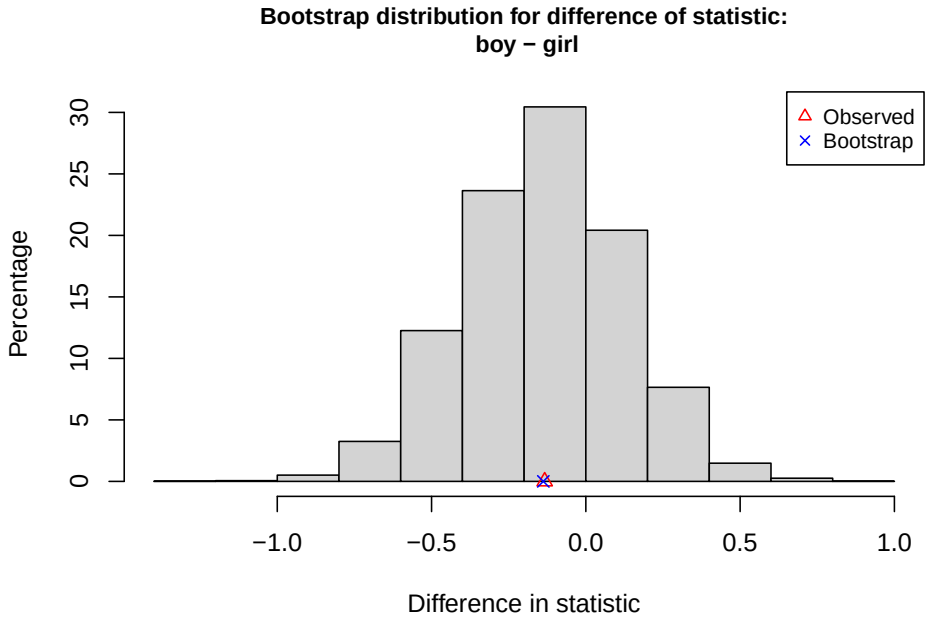
```
set.seed(7)
boot(interest ~ gender, data = study4age5)
```

```
    ** 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.

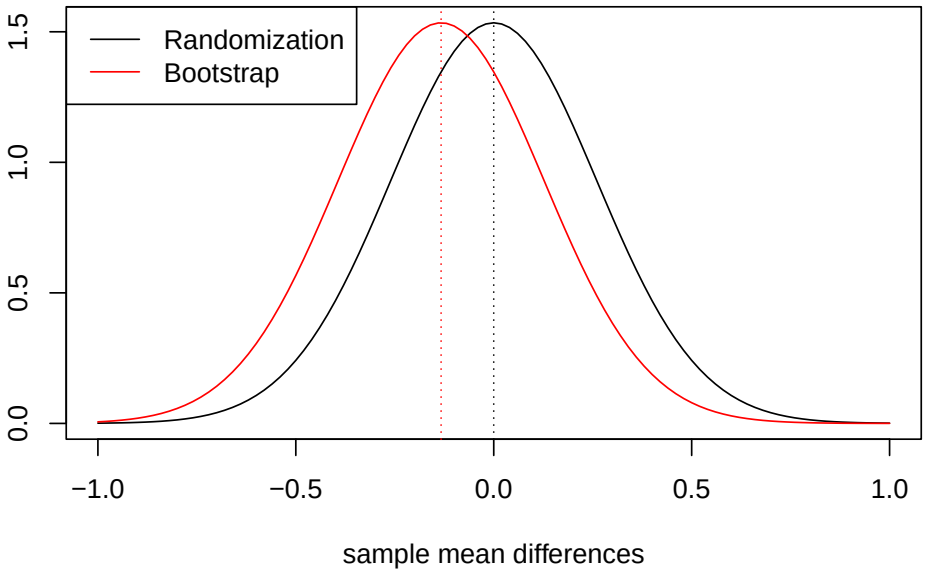
- What is the bootstrap SE? Is it similar to the randomization distribution SE?

Click for answer

Answer: SE is about 0.26, which is very similar to the randomization distribution SE.

- Sketch out both the bootstrap and randomization distributions. Make sure to accurately represent the center and variation of these distributions.

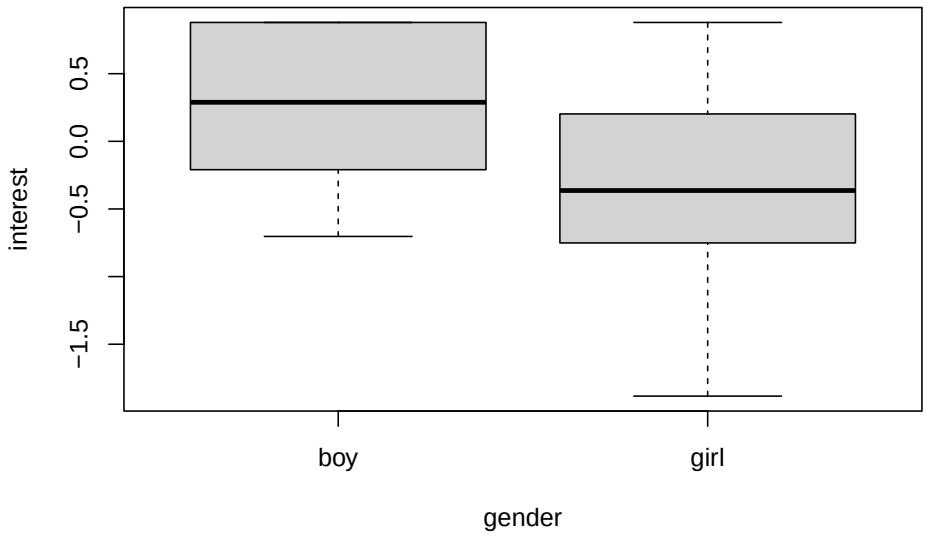
```
curve(dnorm(x,0,.26),from=-1,to=1, ylab="",xlab="sample mean differences")
abline(v=0, lty=3)
curve(dnorm(x,-0.133,.26),from=-1,to=1, add=T, col="red")
abline(v=-0.133, col="red", lty=3)
legend("topleft", col=c("black","red"), legend=c("Randomization","Bootstrap"), lty=1)
```



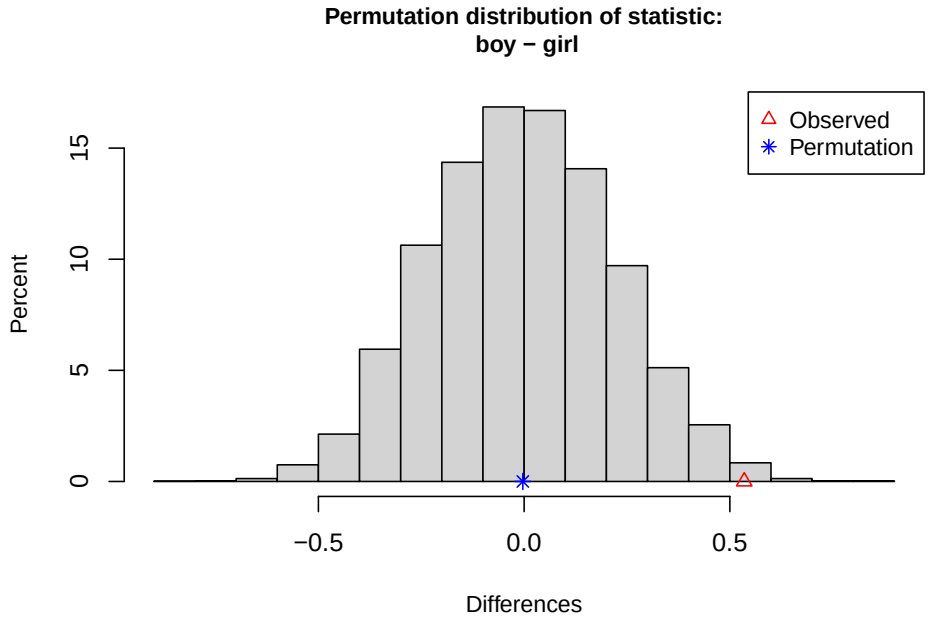
14.1.0.3 (c) Interest in 6 and 7 year olds - test

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
```

14.1.0.4 (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?

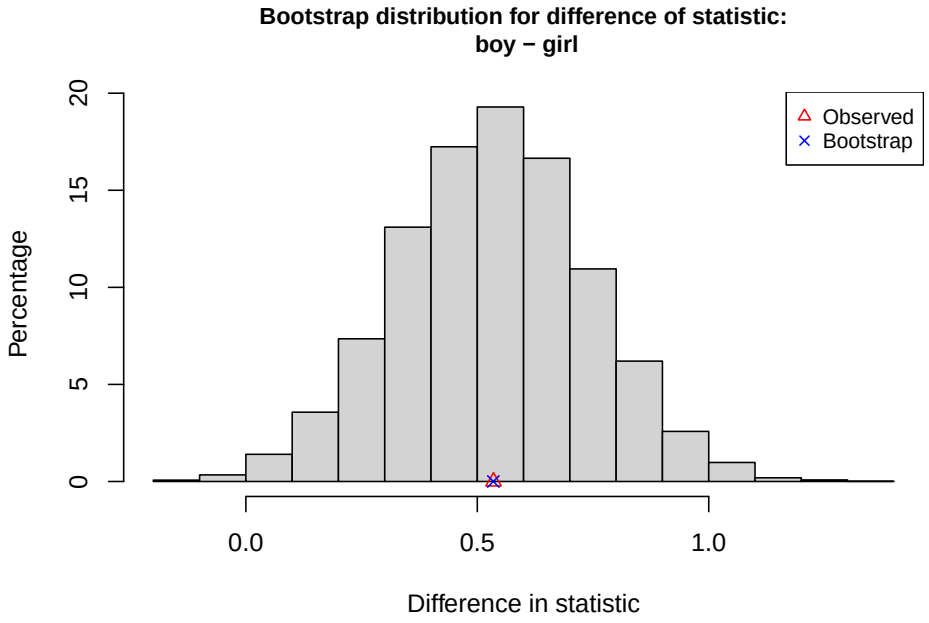
```
boot(interest ~ gender, data = study4age67)
```

```
    ** 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.

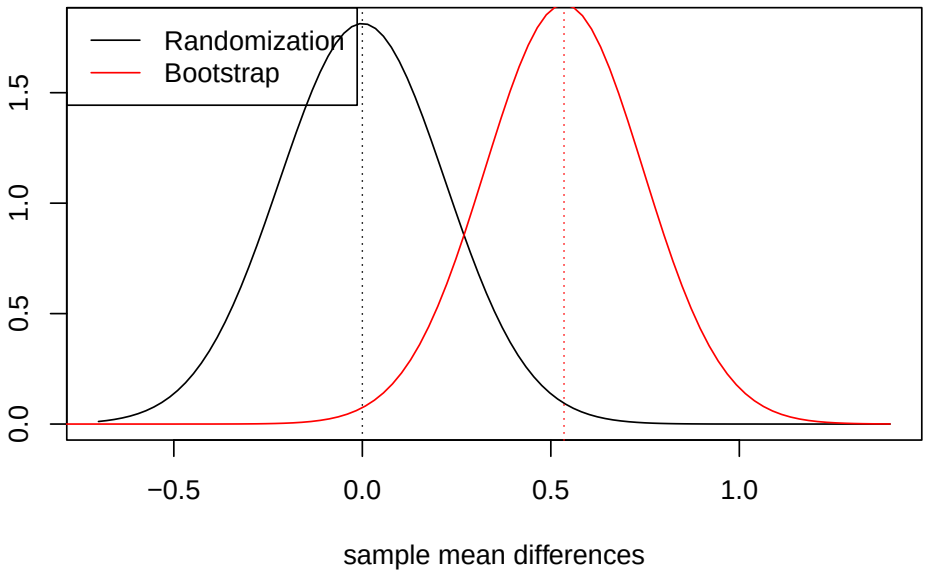
- What is the bootstrap SE? Is it similar to the randomization distribution SE?

  Click for answer

  Answer: SE is about 0.21, which is very similar to the randomization distribution SE.

- Sketch out both the bootstrap and randomization distributions. Make sure to accurately represent the center and variation of these distributions.

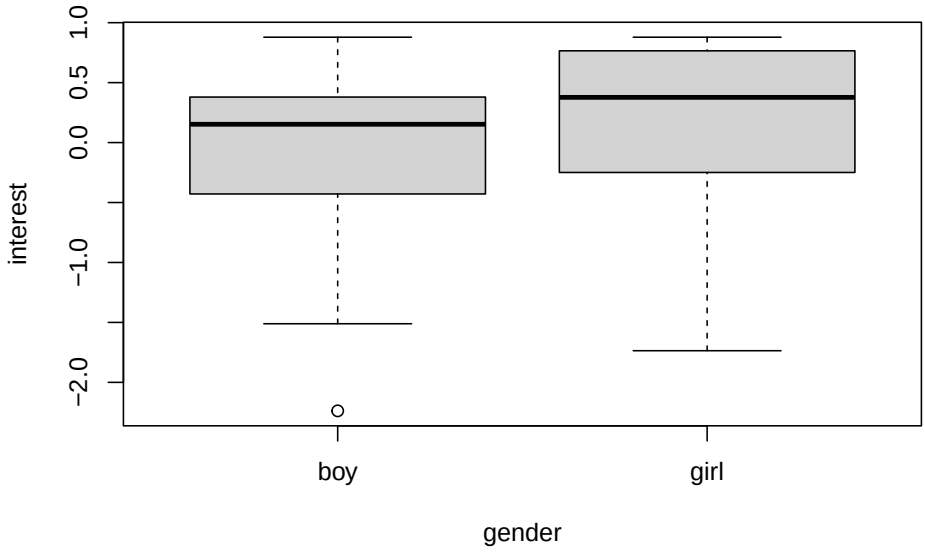
```
curve(dnorm(x,0,.22),from=-.7,to=1.4, ylab="",xlab="sample mean differences")
abline(v=0, lty=3)
curve(dnorm(x,0.535,.21),from=-1,to=1.4, add=T, col="red")
abline(v=0.535, col="red", lty=3)
legend("topleft", col=c("black","red"), legend=c("Randomization","Bootstrap"), lty=1)
```



14.1.0.5 (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)
```

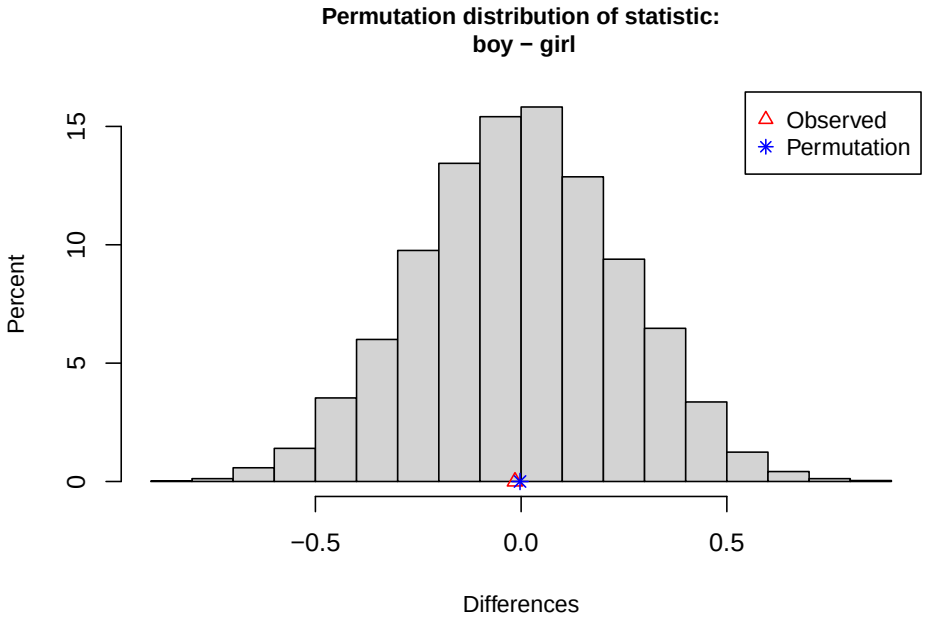



```
which(study4age5$interest < -2)
```

```
[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)
```

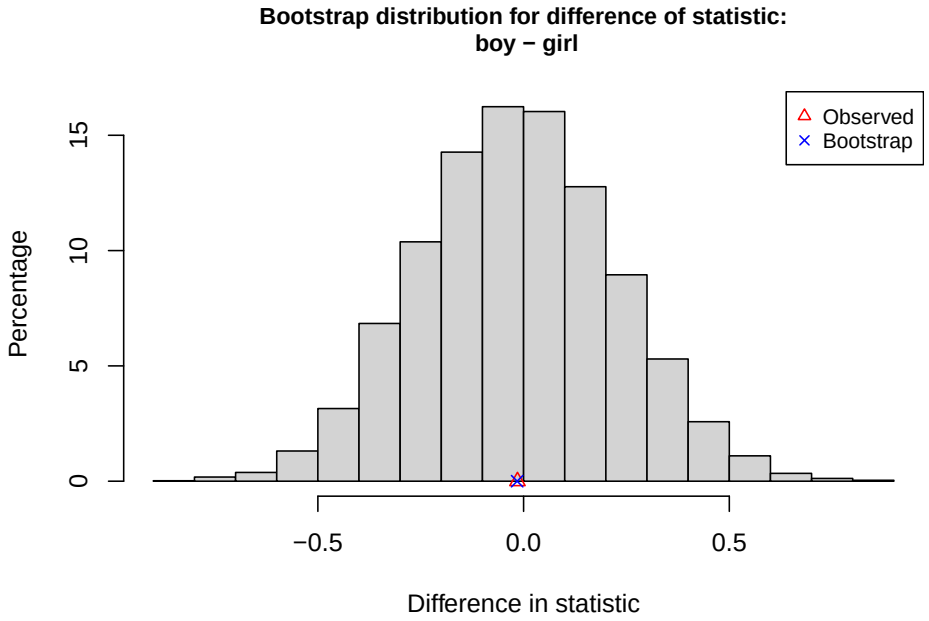
```
   ** Bootstrap interval for difference of statistic

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 15

Class Activity 15

15.1 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 $X \sim N(580, 70)$.

15.1.0.1 (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
```

15.1.0.2 (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
```

15.1.0.3 (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
```

15.1.0.4 (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
```

15.1.0.5 (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
```

15.2 Example 2: Standard Normal

The standard normal distribution has a mean of 0 and standard deviation of 1.

15.2.0.1 (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.)

15.2.0.2 (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 16

Class Activity 16

16.1 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.

16.1.0.1 (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

$$H_0 : p = 0.5 \quad H_A : p > 0.5$$

16.1.0.2 (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

$$z = \frac{\text{statistic} - \text{null parameter}}{SE} = \frac{0.67 - 0.50}{0.021} = 8.10$$

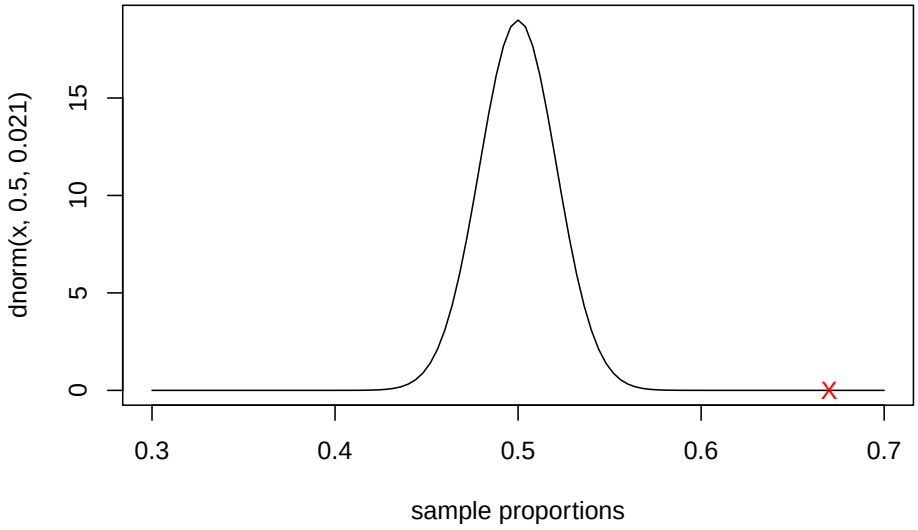
The observed proportion is 8.1 SEs above the hypothesized value of 0.5.

```
(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")
```



16.1.0.3 (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
```

16.1.0.4 (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.

16.1.0.5 (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
```

16.2 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.

16.2.0.1 (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

$$H_0 : p_f = p_m \quad H_A : p_f \neq p_m$$

16.2.0.2 (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:

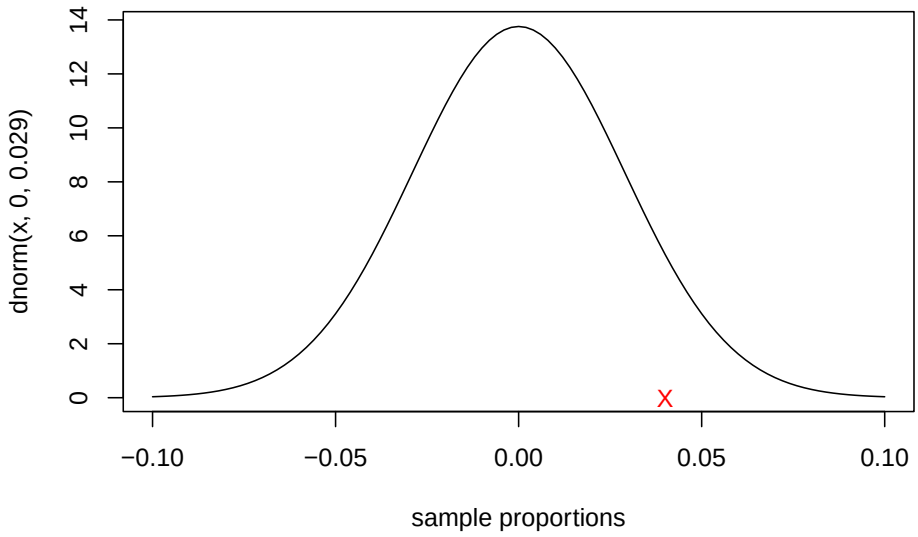
$$z = \frac{\text{statistic} - \text{null parameter}}{SE} = \frac{(0.71 - 0.67) - 0}{0.029} = 1.379$$

```
(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")
```



16.2.0.3 (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
```

16.2.0.4 (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 17

Class Activity 17

17.1 Example 1: Is the Economy a Top Priority?

A survey of 1,502 Americans in January 2012 found that 86% consider the economy a "top priority" for the president and congress. In the section 3.2 handout, we gave the standard error for this sample proportion as 0.01, then this SE was used to compute a confidence interval. Show how this SE was computed using the appropriate SE formula from chapter 6.

Click for answer

Answer: We have a sample proportion of $\hat{p} = 0.86$. The SE of the sample proportion for a confidence interval is given by:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.86(1-0.86)}{1502}} = 0.0089 \approx 0.01$$

17.2 Example 2: 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:

$$\text{Statistic } +/- Z^*SE$$

$$\begin{aligned}
&\hat{p} \pm z^* \cdot \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} \\
&0.64 \pm 1.96 \cdot \sqrt{\tfrac{0.64(1-0.64)}{500}} \\
&0.64 \pm 0.042 \\
&(0.598, 0.682)
\end{aligned}$$

(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.

17.3 Example 3: 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:

$$\begin{aligned}
0.02 &= z^* \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} \\
n &= \left(\tfrac{z^*}{0.02}\right)^2 \hat{p}(1-\hat{p}) \\
&= \left(\tfrac{1.96}{0.02}\right)^2 0.64(1-0.64) = 2212.76
\end{aligned}$$

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.

17.4 Example 4: 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 $\widehat{p} = \frac{108}{556} = 0.1942$ and the sample size is $n = 556$. The test statistic is:

$$z = \frac{\text{Statistic} - \text{Null}}{SE} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1942 - 0.1875}{\sqrt{\frac{0.1875(1-0.1875)}{556}}} = 0.405$$

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 18

Class Activity 18

18.1 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:

$$SE = \sqrt{\frac{\hat{p}_{1994}(1-\hat{p}_{1994})}{n_{1994}} + \frac{\hat{p}_{2015}(1-\hat{p}_{2015})}{n_{2015}}} = 0.0194 \approx 0.02$$

18.2 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

18.2.1 (a) Are the conditions met for using the normal distribution?

Click for answer

Answer: Yes (all cell counts at least 10)

18.2.2 (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 $\widehat{p} = \frac{58}{96} = 0.6042$.

18.2.3 (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))

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.

18.3 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. 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.

18.4 Example 4: 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.

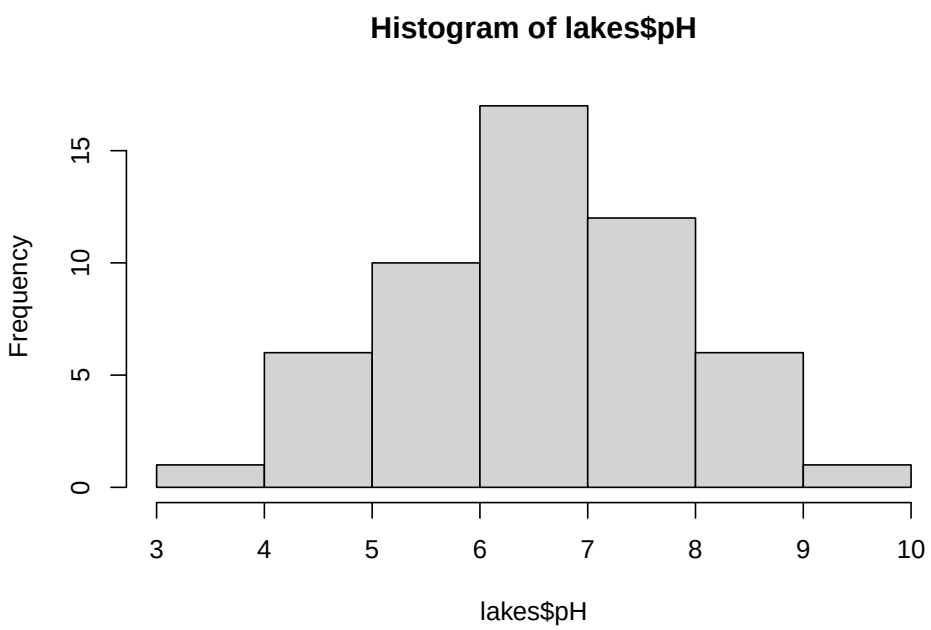
```
lakes <- read.csv("http://www.lock5stat.com/datasets1e/FloridaLakes.csv")
head(lakes)
```

```
  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
```

18.4.0.1 (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.

18.4.0.2 (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
```

18.4.0.3 (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
```

18.4.1 (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).

18.4.1.1 (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.

18.4.1.2 (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
```

18.5 Example 5: 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")
```

18.5.0.1 (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:

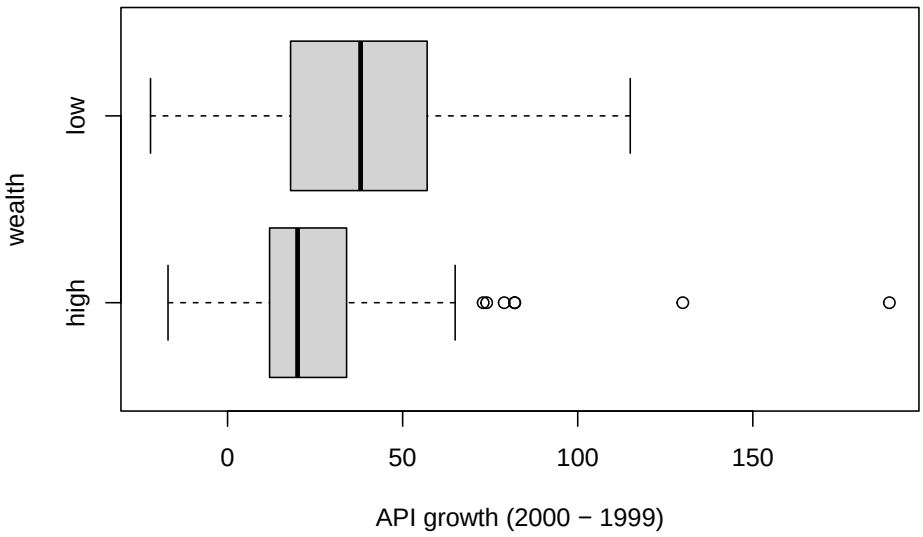
```
api$wealth <- ifelse(api$meals > 50, "low","high")
table(api$wealth)
```

```
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.

18.5.0.2 (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
```

18.5.0.3 (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
```

18.5.0.4 (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.

18.5.0.5 (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.

18.5.0.6 (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 no
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.

18.5.0.7 (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.

18.5.0.8 (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.

Chapter 19

Class Activity 19

19.1 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)
```

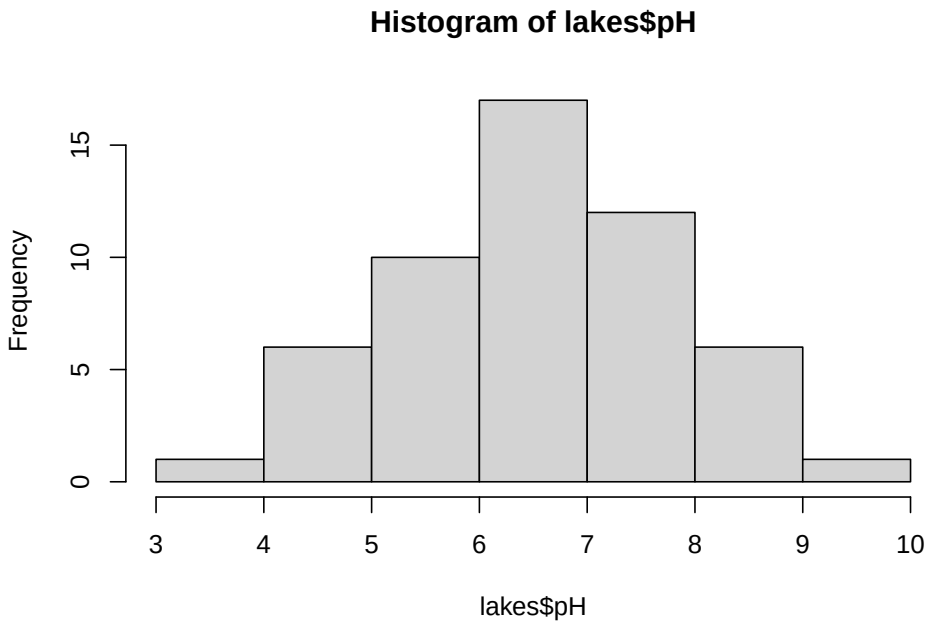
```
  ID         Lake Alkalinity  pH Calcium Chlorophyll
1  1    Alligator        5.9 6.1     3.0         0.7
2  2        Annie        3.5 5.1     1.9         3.2
3  3       Apopka      116.0 9.1    44.1       128.3
4  4 Blue Cypress       39.4 6.9    16.4         3.5
5  5        Brick        2.5 4.6     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
```

19.1.0.1 (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.

19.1.0.2 (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
```

19.1.0.3 (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
```

19.1.1 (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).

19.1.1.1 (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.

19.1.1.2 (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
```

19.2 Example 2: 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")
```

19.2.0.1 (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:

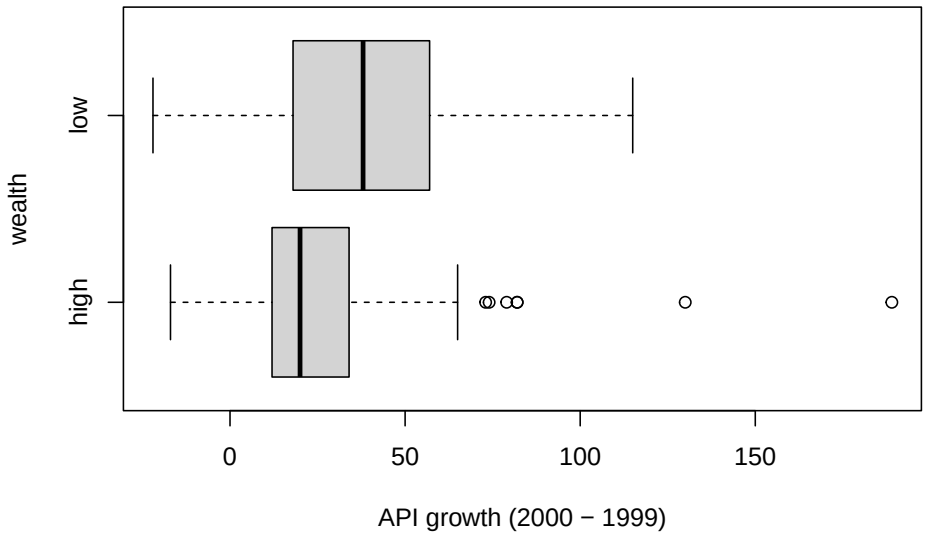
```
api$wealth <- ifelse(api$meals > 50, "low","high")
table(api$wealth)
```

```
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.

19.2.0.2 (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
```

19.2.0.3 (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
```

19.2.0.4 (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 not equal to
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.

19.2.0.5 (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.

19.2.0.6 (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.

19.2.0.7 (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.

19.2.0.8 (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.

19.3 Example 3: 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
```

19.3.0.1 (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.

19.3.0.2 (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.

$$H_0 : \mu = 0 \text{ vs. } H_A : \mu > 0$$

19.3.0.3 (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.

19.3.0.4 (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).

19.3.0.5 (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.

19.3.0.6 (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 20

Class Activity 20

20.1 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:

$H_0 : p_{FP} = 0.1, \quad p_{NFP} = 0.9$ 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

20.2 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 21

Class Activity 21

21.1 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.

21.1.0.1 (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.

21.1.0.2 (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
counts <- table(survey$religiousness, survey$comfortness)
counts
```

```
                      almost always sometimes rarely
  not religious                 103        76     15
  religious not active           39        41     19
  religious active               18        24     15
```

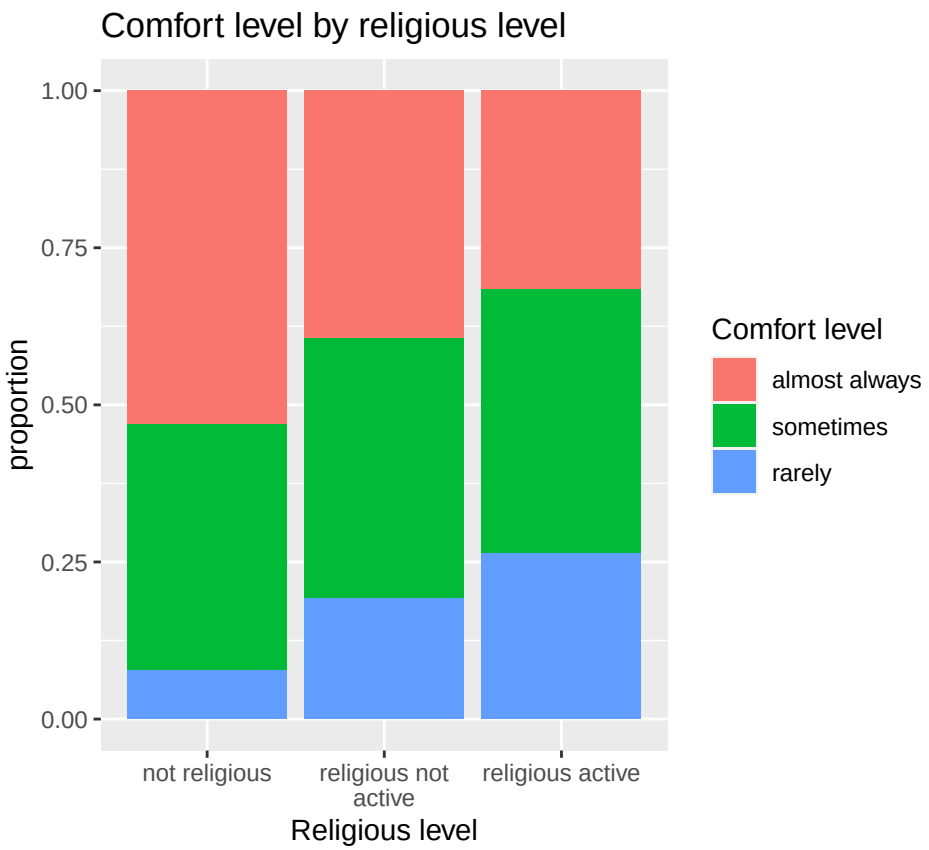
```
sum(counts)  # number of respondents
```

```
[1] 350
```

```
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
```

```
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%).

21.1.0.3 (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
```

21.1.0.4 (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.

$$\frac{(103 - 88.6857143)^2}{88.6857143} = 2.31$$

21.1.0.5 (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.

21.1.0.6 (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 $= 7 \times 10^{-4}$).

21.1.0.7 (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.

21.1.0.8 (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 = 0.001999
```

The p-value is slightly different, but your conclusion should be the same.

21.1.0.9 (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),3)
```

```
[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.

21.2 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

	Arrested	Not Arrested	Total
Total	51	72	123

Is a statistically significant difference between the rate of arrest (or no arrest) in the two treatment groups.

21.2.0.1 (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.

21.2.0.2 (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).

21.2.0.3 (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.

21.2.0.4 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
```

21.3 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:

21.3.1 Step 1:

H_0 : exercise and living arrangements independent of each other

H_A : exercise and living arrangements dependent of each other

21.3.2 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.

21.3.3 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
```

```
[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
```

21.3.4 Step 4:

The p-value can also be calculated as

```
1 - pchisq(60.43, df = 6)
```

```
[1] 3.680733e-11
```

21.3.5 Step 5:

There is significant evidence of an association between the exercise and living arrangements ($\chi^2 = 60.43$, df=6, p-value ≈ 0).

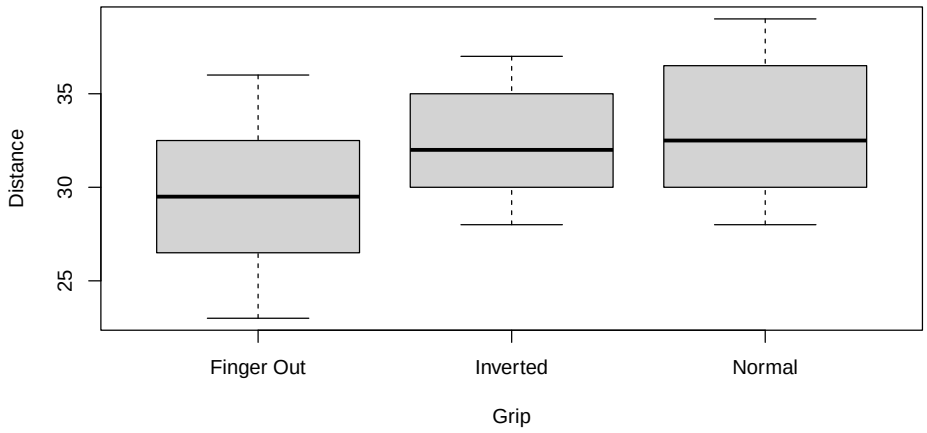
Chapter 22

Class Activity 22

22.1 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**.

22.1.0.1 (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.

22.1.0.2 (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.

22.1.0.3 (c) Checking assumptions

Can we trust the p-value obtained above using the F distribution?

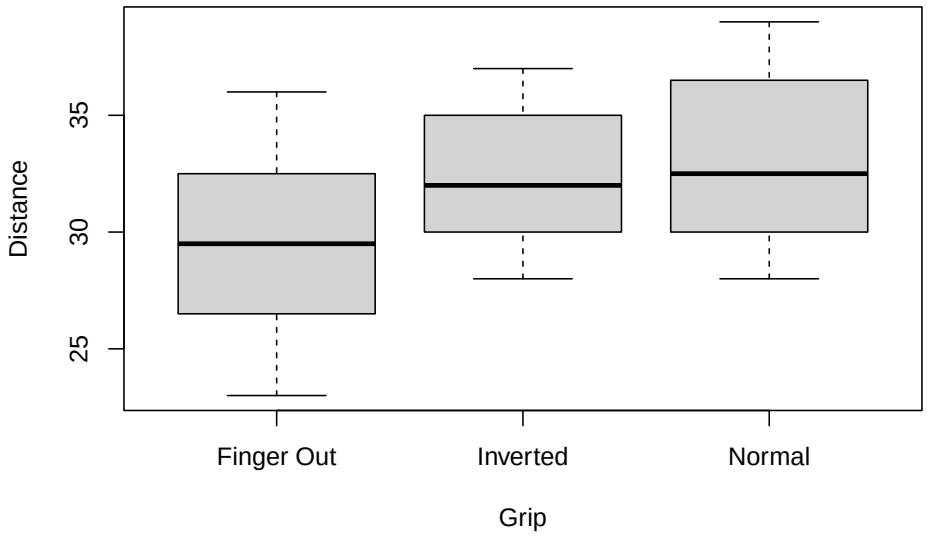
```
table(frisbee$Grip)  # check n's
```

```
Finger Out   Inverted     Normal
         8          8          8
```

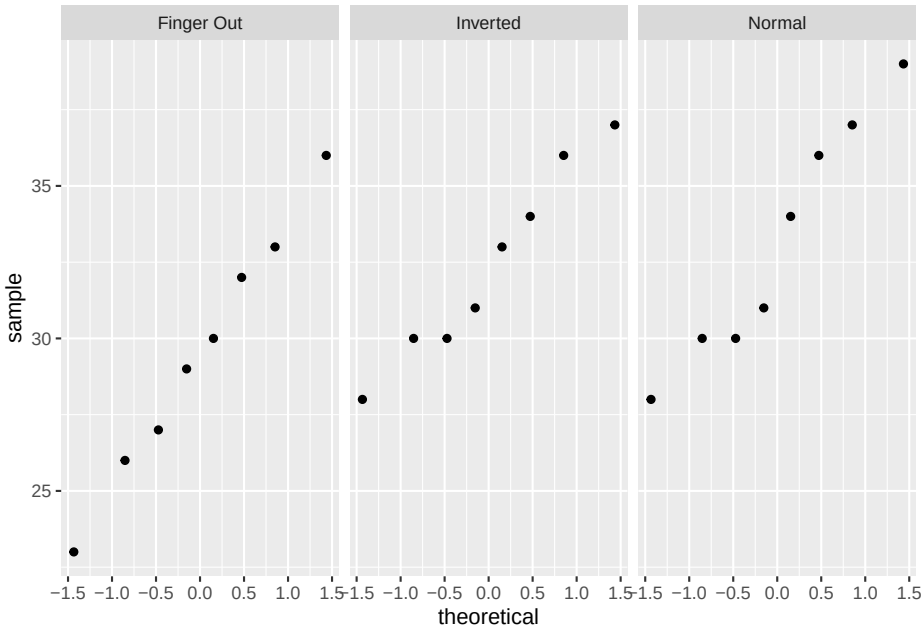
```
tapply(frisbee$Distance, frisbee$Grip, sd)  # similar SD's?
```

```
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.

22.2 Example 2: Comparing % religious guess by religion

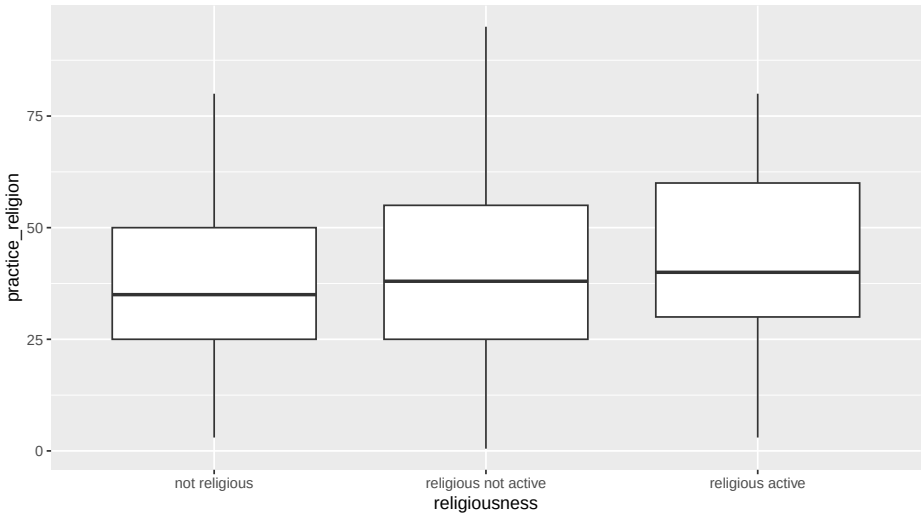
One of the class survey questions asked respondents to give their best guess at the percentage of students at Carleton who practice a religion. Here we can compare responses to this question by the religiousness of the respondent:

```
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
```

22.2.0.1 (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.

22.2.0.2 (b) Check assumptions

Can use trust the results from a one-way ANOVA test?

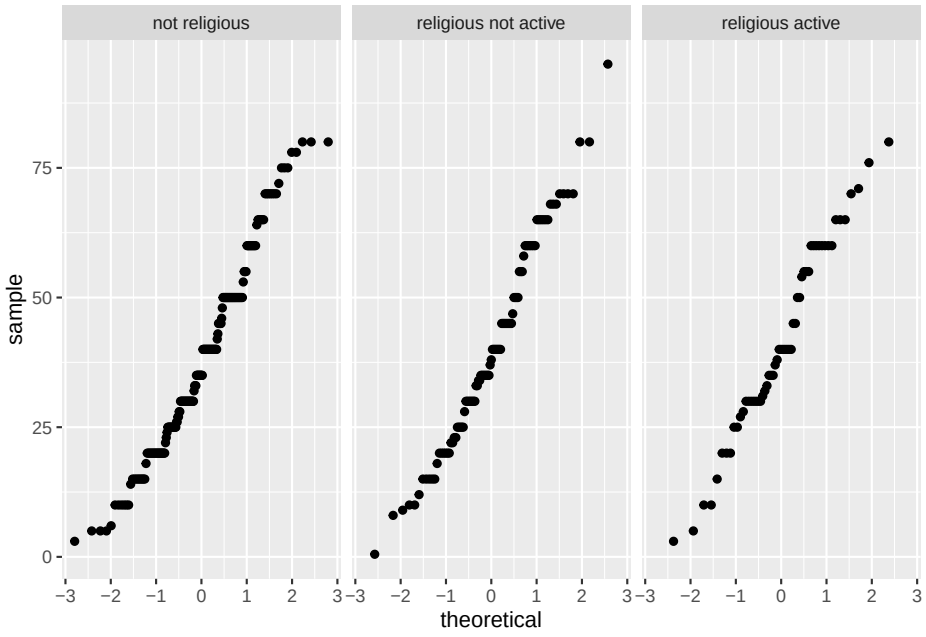
```
table(survey$religiousness)
```

```
       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%).

22.2.0.3 (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.

22.2.0.4 (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 23

Class Activity 23

23.1 Example 1: Cuckoo Eggs

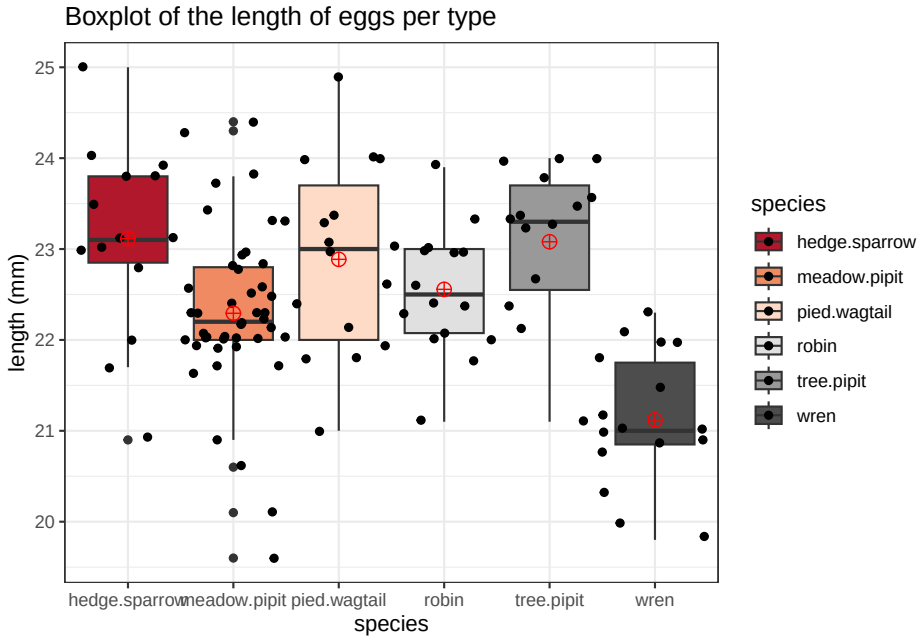
The common cuckoo does not build its own nest: it prefers to lay its eggs in another birds' nest. It is known, since 1892, that the type of cuckoo bird eggs are different between different locations. In a study from 1940, it was shown that cuckoos return to the same nesting area each year, and that they 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.

23.1.1 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?

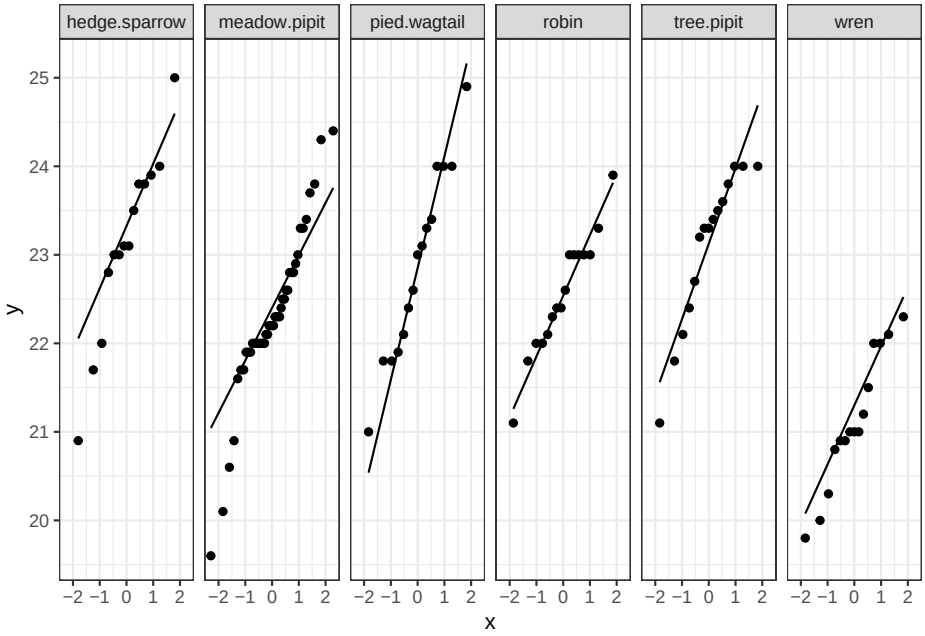


23.1.2 (1b) Formally verify that the assumptions are valid by using the outputs given.

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
```

23.1.3 (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.

23.1.4 (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
```

$$22.556 \pm 1.981 * \frac{\sqrt{0.8194}}{\sqrt{16}}$$
$$= (22.108, 23.005)$$

23.1.5 (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
```

$$(22.556 - 21.120) \pm 1.981 \cdot \sqrt{0.8194\left(\frac{1}{16} + \frac{1}{15}\right)}$$

$= (0.792, 2.081)$

23.1.6 (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
```

$$(22.887 - 22.556) \pm 1.981 \cdot \sqrt{0.8194\left(\frac{1}{15} + \frac{1}{16}\right)}$$
$$= (-0.314, 0.975)$$

23.1.7 (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:

$$H_0 : \mu_i = \mu_j \text{ vs. } H_a : \mu_i \neq \mu_j$$

23.2 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

23.2.0.1 (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.

23.2.0.2 (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.

23.2.0.3 (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
```

23.2.0.4 (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
```

23.2.0.5 (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.

23.2.0.6 (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 :

$$(21.34 - 13.16) \pm (2.201))\sqrt{132.97\left(\frac{1}{5} + \frac{1}{5}\right)} = 1.63, 14.73$$

```
(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 1.63 to 14.3 mcg/g higher than the true mean concentration at site 3.

```
(21.34 - 13.16)
```

```
[1] 8.18
```

```
qt(.975, 11)
```

```
[1] 2.200985
```

```
sqrt(12.088*(1/5+1/5))
```

```
[1] 2.198909
```

```
(21.34 - 13.16) -  qt(.975, 5-1)*sqrt(12.088*(1/5+1/5))
```

```
[1] 2.07485
```

```
(21.34 - 13.16) +  qt(.975, 5-1)*sqrt(12.088*(1/5+1/5))
```

```
[1] 14.28515
```

Chapter 24

(PART*) Basics R

Chapter 25

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.

25.1 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.

25.2 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.)

25.3 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.

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

- Download the latest version of R:
  - Windows: http://cran.r-project.org/bin/windows/base/
  - Mac: http://cran.r-project.org/bin/macosx/
- Download the free Rstudio desktop version (Windows or Mac): https://www.rstudio.com/products/rstudio/download/

Use the default download and install options for each.

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

If you want to compile R Markdown to .pdf files, you also need a LaTeX distribution (Note: this is not necessary if you choose to compile as a Word document.) Click instructions for Windows or instructions for Mac, depending on your operating system to complete the installation.

25.6 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.

25.7 Instructions

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.

Extra notes:

- 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.

25.8 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.

Chapter 26

R Markdown

This is a R Markdown 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.

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
```

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

You can embed Latex equations in-line, $\frac{1}{n}\sum_{i=1}^{n} x_i$ or in a new line as

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

Embed an R code chunk:

Use

```
```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  
```

26.1 Including Plots

You can also embed plots. See Figure 26.1 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
)
```

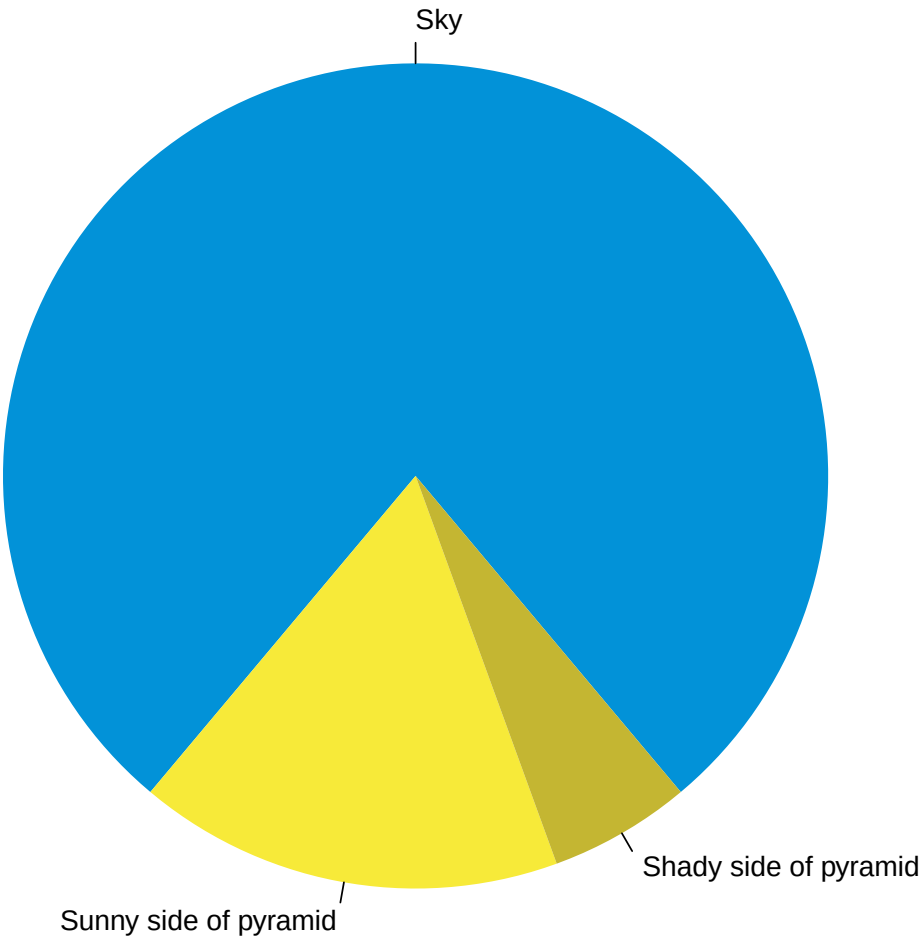


Figure 26.1: A fancy pie chart.

(Credit: Yihui Xie)

26.2 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
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

26.3 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 27

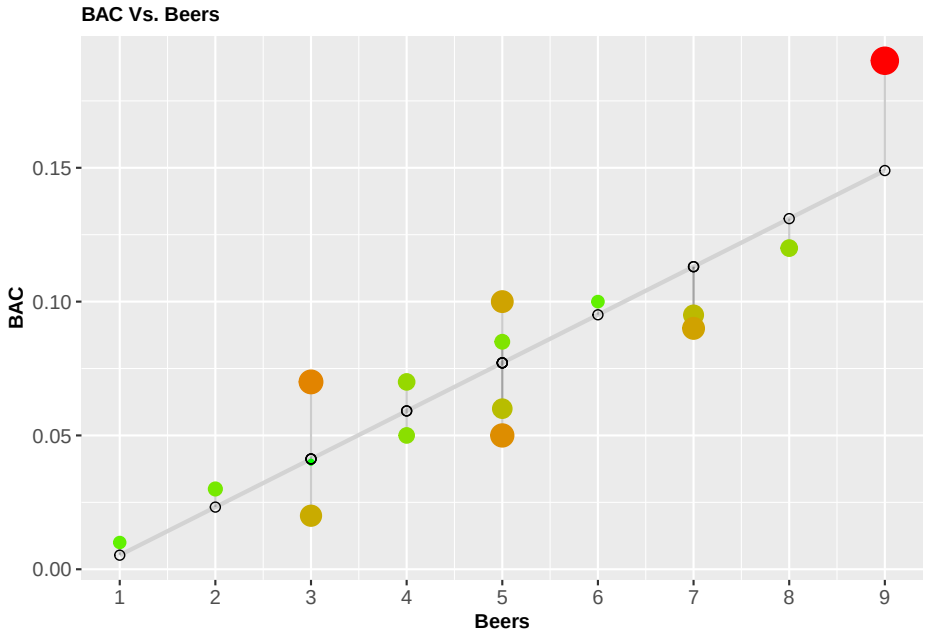
Helpful R codes

27.1 Residual Plots in ggplot2

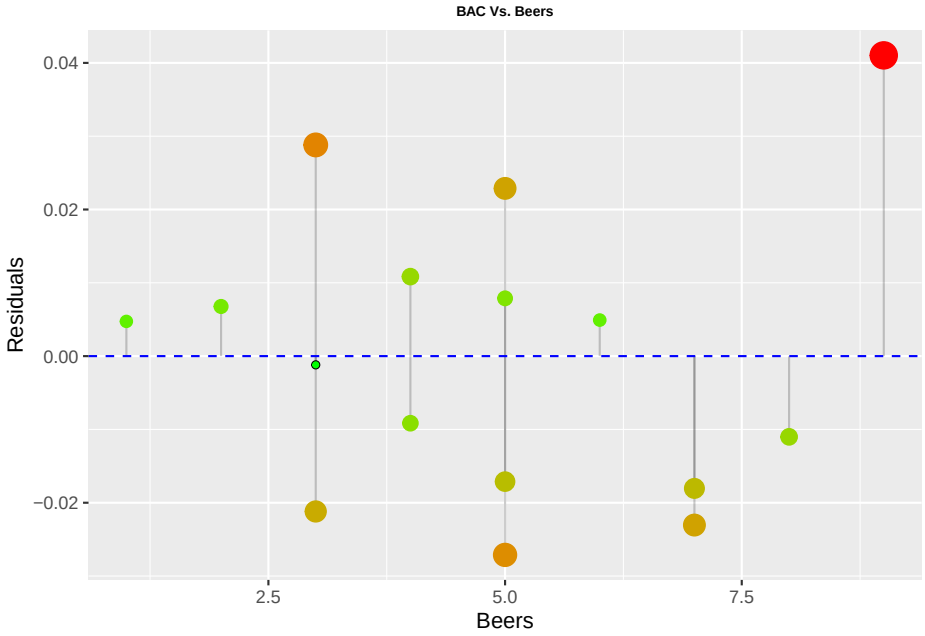
```
# residual size plot
library(ggplot2)
bac <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/BAC.csv")

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'))
```



27.2 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')
```