STA130H1S – Fall 2022

Problem Set 4

() and STA130 Professors

Instructions

Complete the exercises of **Part 2** in this .Rmd file and submit your .Rmd and .pdf output through Quercus on Oct 6 by 5:00 p.m. ET.

Complete the exercises of **Part 3** in this .Rmd file and submit your .Rmd and .pdf output through Quercus on Oct 13 by 5:00 p.m. ET.

```
library(tidyverse)
```

Part 1: OPTIONAL Warm Up if Needed.

Complete these guided questions if you need some additional help getting started with hypothesis testing before moving on to **Part 2**, or if you want some additional practice with this hypothesis testing. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

Question 1: Warm Up with Biased Coin Flipping

Approximately 23% of the general population use the social media platform Twitter. Suppose that the Department of Statistical Sciences (DoSS) is conducting a study to see if this percentage is the same among their undergraduate students (that is, all students in an undergraduate DoSS statistics program). Suppose $n = 400$ students in statistics programs are randomly selected and asked whether or not they use Twitter. Suppose that 103 of these 400 students respond that they use Twitter.

(a) What is the NULL hypotheses H_0 in terms of p? What is H_1 in terms of H_0? In a simple sentence without H_0 and p notation, what is the claim of the NULL hypothesis?

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Set `set.seed(11)` and use the `sample()` function to simulate the number of students who use Twitter in a random sample of 400 DoSS students under the assumption that the prevalence of Twitter usage is the same among DoSS students as it is in the general population. How many Twitter users did you have in your simulated sample of 400 students?

```
set.seed(11) # REQUIRED so the random sample is reproducible!
# Code your answer here
```

```
sample(c("Head","Tail"), size=10, replace=TRUE)
```

Hints

```
##  [1] "Tail" "Tail" "Tail" "Head" "Tail" "Head" "Head" "Tail" "Tail" "Tail"
```

```r
# will do the same thing as:
sample(c("Head","Tail"), size=10, prob=c(0.5, 0.5), replace=TRUE)
```

```
##  [1] "Tail" "Tail" "Head" "Head" "Head" "Head" "Tail" "Tail" "Tail" "Tail"
```

```r
# Even though the exact counts of "Head" and "Tail" differ each time you
# run this code, if you simulate enough coin flips (by increasing
# the value of 'size', you'll get approximately the same proportion
# of "Head" and "Tail" outcomes)

# To modify the code to make Tails much more likely than Heads,
# we could change the probs:
sample(c("Head","Tail"), size=10, prob=c(0.1, 0.9), replace=TRUE)
```

```
##  [1] "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail"
```

(c) Use `geom_bar()` to visualize the number of Twitter users versus non-Twitter users with a bar plot. How does this simulated proportion compare to the general population rate of 23% and to the 103 of 400 sampled DoSS students?

```r
# Code your answer here
```

```r
# You can make a vector a column of a `tibble` like this
tibble(flips = c("Head", "Tail", "Tail"))
```

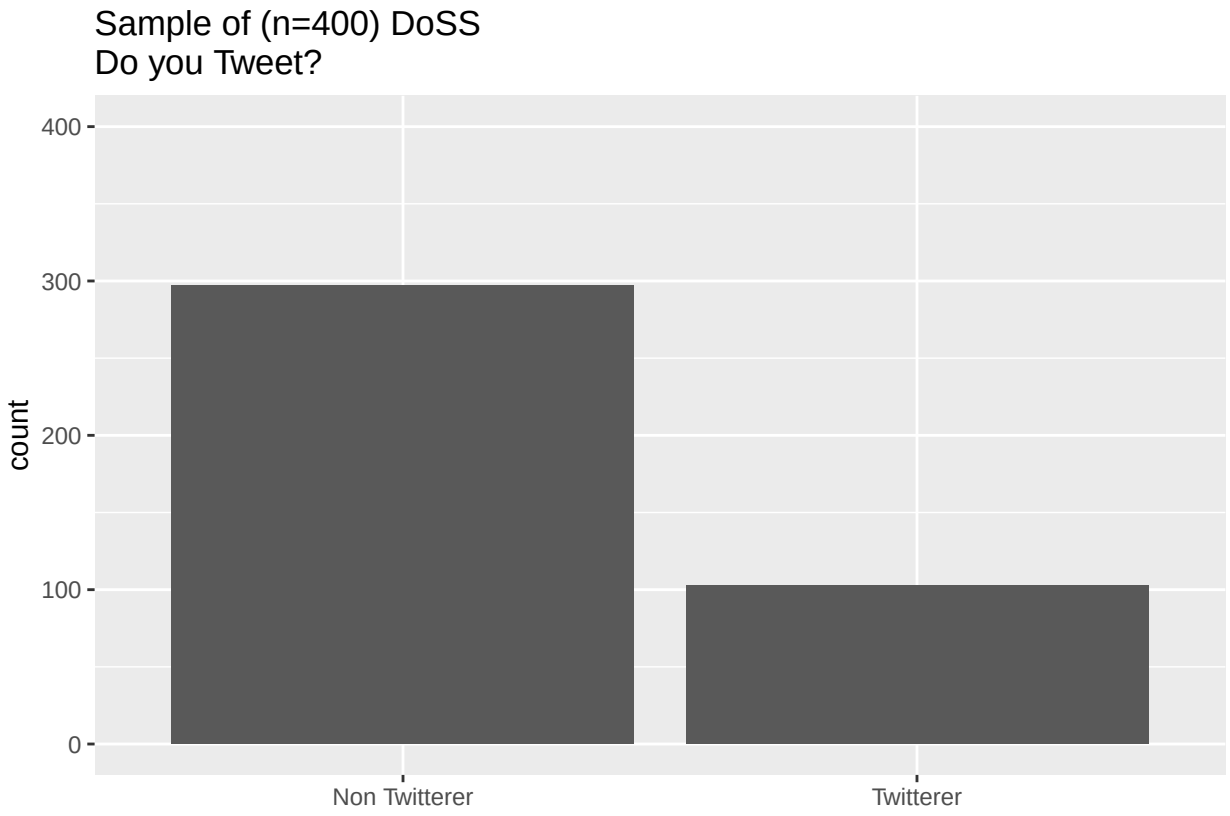
Hints

```
## # A tibble: 3 x 1
##   flips
##   <chr>
## 1 Head
## 2 Tail
## 3 Tail
```

(d) How is the `geom_bar()` function different than the `geom_col()` function below?

REPLACE THIS TEXT WITH YOUR ANSWER

```r
ggplot(data=NULL, aes(x=c("Twitterer","Non Twitterer"), y=c(103,400-103))) +
  geom_col() + lims(y=c(0,400)) +
  labs(title="Sample of (n=400) DoSS\nDo you Tweet?", y="count", x="")
```



(e) Simulate the sampling distribution of the test statistic under the assumption that the prevalence of Twitter usage among DoSS students matches that of the general population. Set the seed to the last 2 digits of your student number, use a simulation of size 1000, make a plot of the simulated sampling distribution, and describe the distribution in a few sentences.

- If you don't set `set.seed()` the simulation will be different each time its run.
- Your knit won't be reproducible and won't align with your interpretations and conclusions.

```
# Clearly label your figure with `labs(x="A primary title\n and a second line")`
# Code your answer here
```

(e) What is the definition of a p-value?

REPLACE THIS TEXT WITH YOUR ANSWER

(f) What is the p-value of the hypothesis test based on the sampling distribution above?

```
# Use the `abs()` function to reflect the "as or more extreme" aspect of the p-value
# Code your answer here
```

(g) At the $alpha = 0.05$ **significance level, what is your conclusion about this hypothesis test based on the p-value computed above?**

REPLACE THIS TEXT WITH YOUR ANSWER

(h) Which of the following statements is correct regarding the p-value above?

(A) The probability that the proportion of DoSS students who use Twitter matches the general population.

(B) The probability that the proportion of DoSS students who use Twitter does not match the general population.

(C) The probability of obtaining a number of students who use Twitter in a sample of 400 students at least as extreme as the result in this study.

(D) The probability of obtaining a number of students who use Twitter in a sample of 400 students at least as extreme as the result in this study, if the prevelance of Twitter usage among all DoSS students matches the general population.

(i) What happens to the p-value if you change the seed value in `set.seed()`? What happens to the p-value if you change the size of the simulation?

REPLACE THIS TEXT WITH YOUR ANSWER

Part 2: One Sample Hypothesis Testing

DUE THURSDAY Oct 6 by 5 p.m. ET

Question 2: Scottish Medicine

A Scottish woman noticed that her husband's scent changed. Six years later he was diagnosed with Parkinson's disease. His wife joined a Parkinson's charity and noticed that odour from other people. She mentioned this to researchers who decided to test her abilities. They recruited 6 people with Parkinson's disease and 6 people without the disease. Each of the recruits wore a t-shirt for a day, and the woman was asked to smell the t-shirts (in random order) and determine which shirts were worn by someone with Parkinson's disease. She was correct for 12 of the 12 t-shirts! You can read about this here.

(a) Without conducting a simulation, describe what you would expect the sampling distribution of the proportion of correct guesses about the 12 shirts to look like if someone was just guessing.

(b) Carry out a simulation and a hypothesis test of the woman being a lucky guesser as opposed to having some ability to identify Parkinson's disease by smell given that she correctly classified 12 of 12 t-shirts.

```
set.seed(0) # Set the random seed to the last two digits of your student number
N <- 1000 # Change this to 10000 if a finer simulation resolution is required
# Code your answers here
```

(c) Actually, initially the woman correctly identified all 6 people who had been diagnosed with Parkinson's but incorrectly identified one of the others as having Parkinson's. It was only eight months later that the final individual was diagnosed with the disease. What is the p-value when only 11 of 12 were known to be correct?

```
# Code your answers here
```

(d) Are you able to get the p-value for the test using the initial data (i.e., 11 correct instead of 12 correct) without running a new simulation?

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Is the conclusion of the hypthesis test the same for an observed test statistic of 11/12 as for an observed test statistic of 12/12?

REPLACE THIS TEXT WITH YOUR ANSWER

Question 3: Fisher's Tea Experiment

There is an interesting account of the British statistician Ronald Fisher at a tea party in the 1920s. One of the other guests was algae scientist Dr. Muriel Bristol, who refused a cup of tea from Fisher because he put milk in first BEFORE pouring the tea. Bristol was convinced she could taste the difference, and much preferred the taste of tea where the milk was poured in afterwards. Fisher didn't think that there could be a difference and proposed a hypothesis test to examine the situation.

Fisher made 8 cups of tea, 4 with milk in first and 4 with tea in first, and gave them to Dr. Bristol without her seeing how they were made and she would say if she thought the tea or the milk was poured first. As it turned out, Dr. Bristol correctly identified if the tea or milk was poured first for all 8 of the cups. Fisher, being a skeptical statistician wanted to test if this could be happening by chance with Bristol just randomly guessing (or whether it seemed more likely that Bristol was not guessing).

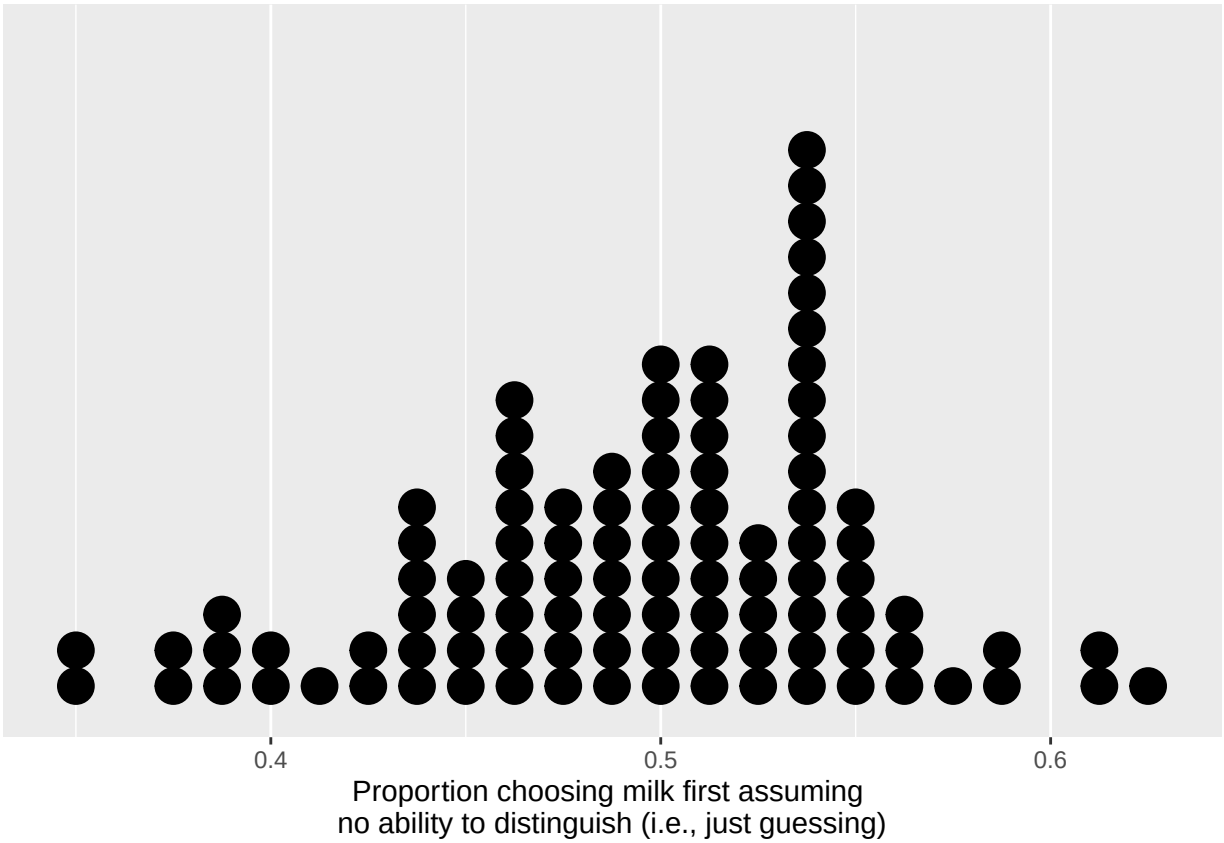
Suppose you run an experiment like this with students in STA130. You get a random sample of 80 STA130 students to each taste one British-style cup of tea and tell you whether they think the milk or tea was poured first. 49 students correctly state which was poured first. Go through the steps to test whether students are just guessing or not.

(a) What is the NULL hypotheses H_0 in terms of p? What is H_1 in terms of H_0? In a simple sentence without H_0 and p notation, what is the claim of the NULL hypothesis?

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Conduct a hypothesis test on the basis of the following simulated sampling distribution of the test statistic assuming the NULL hypothesis is true. For simplicity, this distribution shows the results of only 100 simulations, but in practice this likely wouldn't provide very good p-value resolution refinement.

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



- What does each single dot in the plot represent?

REPLACE THIS TEXT WITH YOUR ANSWER

- Based on this plot, what is your estimate of the p-value?

REPLACE THIS TEXT WITH YOUR ANSWER

- At the $alpha = 0.05$ significance level, what is your conclusion about this hypothesis test based on the p-value computed above?

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Suppose the analysis described in (b) is repeated but this time 1000 simulations are used to get a better estimate of the p-value, and the resulting p-value is 0.04. Do not conduct this simulation. At the $alpha = 0.05$ significance level, what is your conclusion about this hypothesis test for a p-value of 0.04?

REPLACE THIS TEXT WITH YOUR ANSWER

Question 4: OPTIONAL primer for potential TUT discussion

A criminal court considers two opposing claims about a defendant: they are either innocent or guilty. In the Canadian legal system, the role of the prosecutor is to present convincing evidence that the defendant is not innocent. Lawyers for the defendant attempt to argue that the evidence is *not convincing enough* to rule out that the defendant could be innocent. If there is not enough evidence to convict the defendant and they are set free, the judge generally does not deliver a verdict of "innocent", but rather of "not guilty".

(a) If we look at the criminal trial example in the hypothesis test framework, which would be the null hypothesis and which the alternative?

REPLACE THIS TEXT WITH YOUR ANSWER

(b) In the context of this problem, describe what rejecting the null hypothesis would mean.

REPLACE THIS TEXT WITH YOUR ANSWER

(c) In the context of this problem, describe what failing to reject the null hypothesis would mean.

REPLACE THIS TEXT WITH YOUR ANSWER

(d) In the context of this problem, describe what a type II error would be.

REPLACE THIS TEXT WITH YOUR ANSWER

(e) In the context of this problem, describe what a type I error would be.

REPLACE THIS TEXT WITH YOUR ANSWER

Part 3: Two Sample Hypothesis Testing Required Questions

DUE THURSDAY Oct 13 by 5 p.m. ET

Question 5: Social Media and Anxiety

There have been many questions regarding whether or not usage of social media increases anxiety levels. For example, do TikTok and Facebook posts create an unattainable sense of life success and satisfaction? Does procrastinating by watching YouTube videos or reading Twitter posts contribute unnecessary stress from deadline pressure? A study was conducted to examine the relationship between social media usage and student anxiety. Students were asked to categorize their social media usage as "High" if it exceeded more than 2 hours per day, and then student anxiety levels where scored through as series of questions, with higher scores suggesting higher student anxiety.

```
# `The rep()` function was introduced above, and you can see what it does here
social_media_usage <- c(rep("Low", 30), rep("High", 16));
anxiety_score <- c(24.64, 39.29, 16.32, 32.83, 28.02,
                   33.31, 20.60, 21.13, 26.69, 28.90,
                   26.43, 24.23, 7.10,  32.86, 21.06,
                   28.89, 28.71, 31.73, 30.02, 21.96,
                   25.49, 38.81, 27.85, 30.29, 30.72,
                   21.43, 22.24, 11.12, 30.86, 19.92,
                   33.57, 34.09, 27.63, 31.26,
                   35.91, 26.68, 29.49, 35.32,
                   26.24, 32.34, 31.34, 33.53,
                   27.62, 42.91, 30.20, 32.54)
anxiety_data <- tibble(social_media_usage, anxiety_score)
glimpse(anxiety_data)
```

```
## Rows: 46
## Columns: 2
## $ social_media_usage <chr> "Low", "Low", "Low", "Low", "Low", "Low", "Low", "L~
## $ anxiety_score      <dbl> 24.64, 39.29, 16.32, 32.83, 28.02, 33.31, 20.60, 21~
```

(a) What is the NULL hypotheses H_0 in terms of $Median_{High}$ and $Median_{Low}$? In simple terms, what is the claim of the NULL hypothesis? What is H_1 in terms of H_0?

REPLACE THIS TEXT WITH YOUR ANSWER

Hint A formal NULL hypotheses that the means of two groups are the same would be $H_0 : \mu_{High} = \mu_{Low}$.

(b) Revisit your statements regarding the NULL hypotheses above with confounding in mind; namely, since social media usage is a self selecting process, perhaps social media users are already more anxious people on average regardless of their social media usage. If we make a determination about the NULL hypothesis are we actually addressing the question of "whether or not usage of social media increases anxiety levels"? Or are we just using a hypothesis test to examine if there is an observable difference between the two groups (regardless of its causes)?

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Construct boxplots of `anxiety_score` for the two levels of social media usage, and write 2-3 sentences describing and comparing the distributions of anxiety scores across the social media usage groups.

```
# Code your answers here
```

REPLACE THIS TEXT WITH YOUR ANSWER

(d) What do these data visually suggest regarding the claim that the *median* anxiety level is different for those who use social media in high frequency compared to those who use social media in lower frequency?

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Look at the code below and write a few sentences explaining what the code inside the for loop is doing and why.

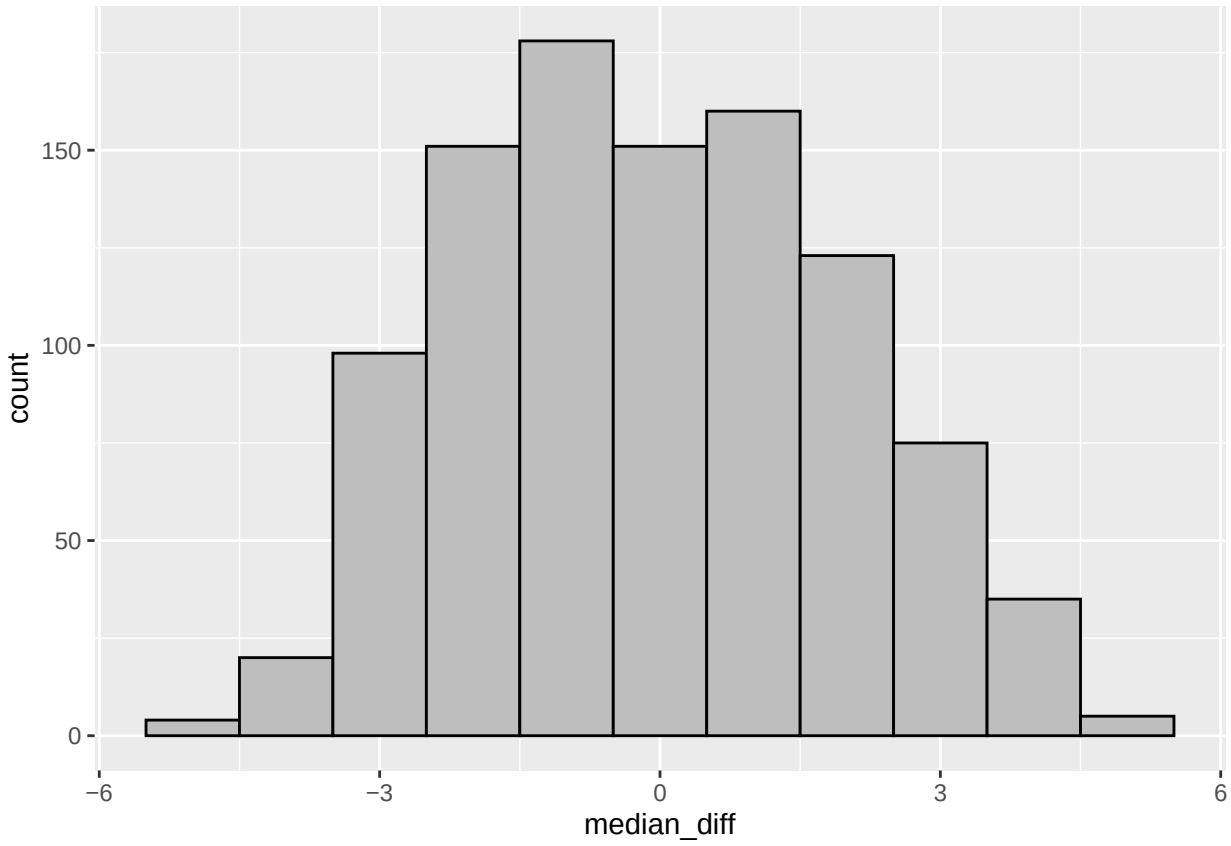
REPLACE THIS TEXT WITH YOUR ANSWER

```
# Note: including the .groups="drop" option in summarise() will suppress a friendly
# warning R prints otherwise "`summarise()` ungrouping output (override with
#`.groups` argument)".
  # Including the .groups="drop" option is optional, but you should include it if you
#don't want to see that warning.
test_stat <- anxiety_data %>% group_by(social_media_usage) %>%
  summarise(medians = median(anxiety_score), .groups="drop") %>%
  summarise(value = diff(medians))
test_stat <- as.numeric(test_stat)
test_stat
```

```
## [1] -4.57
```

```
set.seed(523)
repetitions <- 1000;
simulated_values <- rep(NA, repetitions)

for(i in 1:repetitions){
  simdata <- anxiety_data %>% mutate(social_media_usage = sample(social_media_usage))

  sim_value <- simdata %>% group_by(social_media_usage) %>%
```

```
    summarise(medians = median(anxiety_score), .groups="drop") %>%
    summarise(value = diff(medians))

  simulated_values[i] <- as.numeric(sim_value)
}

sim <- tibble(median_diff = simulated_values)

sim %>% ggplot(aes(x=median_diff)) + geom_histogram(binwidth=1, color="black", fill="gray")
```




```
num_more_extreme <- sim %>% filter(abs(median_diff) >= abs(test_stat)) %>% summarise(n())
p_value <- as.numeric(num_more_extreme / repetitions)
p_value
```

```
## [1] 0.009
```

(f) Summarize the NULL hypothesis and then, at the $alpha = 0.05$ significance level, state your conclusion about the hypothesis test of the NULL hypothesis based on the p-value computed above.

REPLACE THIS TEXT WITH YOUR ANSWER

(g) Do these data support the claim that the *median* anxiety level is different for those who use social media in high frequency compared to those who use social media in lower frequency? How about the claim that "usage of social media increases anxiety levels"?

REPLACE THIS TEXT WITH YOUR ANSWER

Question 6: Airbags

The table below is adapted from "Biostatistics for the Biological and Health Sciences" and presents data from a random sample of passengers sitting in the front seat of cars involved in car crashes. Based on this data we'd like to make a determination as to whether or not death rates differ for passengers in cars with airbags and passengers in cars without airbags.

	Airbag available	No airbag available
Passenger Fatalities	45	62
Total number of Passengers	10,541	9,867

The code below creates a tidy data frame for this problem using the `rep()` function.

```
data <- tibble(group = c(rep("airbag",10541), rep("no_airbag",9867)),
               outcome = c(rep("dead",45), rep("alive",10541-45),
                           rep("dead",62), rep("alive",9867-62)))
```

(a) What is the NULL hypotheses H_0 in terms of p_{airbag} and $p_{no-airbag}$? What is H_1 in terms of H_0? In a simple sentence without H_0 and p_{airbag} and $p_{no-airbag}$ notation, what is the claim of the NULL hypothesis?

REPLACE THIS TEXT WITH YOUR ANSWER

(b) Simulate the the sampling distribution of the test statistic under the assumption that the NULL hypothesis state above is `TRUE`.

```
set.seed(523) # Replace the seed with the 1st, 3rd, and 5th digits or your student number.
# Code your answers here
```

```
# space for scratch work if needed
```

(c) At the $alpha = 0.10$ significance level, what is your conclusion about this hypothesis test based on the p-value computed above?

REPLACE THIS TEXT WITH YOUR ANSWER

(d) Based on your conclusion above, what kind of error could you have made?

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Does your conclusion support the claim that "airbags save lives"? Or does it seem reasonable to believe that there could be some sort of confounding (like in Question 5) by which people who choose to drive in cars without airbags are just more likely on average do die if they're in a car crash irrespective of any safety benefit of airbags?

REPLACE THIS TEXT WITH YOUR ANSWER

Question 7: OPTIONAL practice specifying NULL Hypotheses

We've covered two kinds of hypothesis tests.

- In the first version we hypothesize a proportion in a population (the parameter p) and test that value using the proportion observed in the sample (the sample average $\bar{x}$, also sometimes notated as $\hat{p}$). These *ONE sample hypotheses tests* have a NULL hypothesis of the form $H_0 : p = p_0$ (and of course the ALTERNATIVE hypothesis then takes the form $H_1 : H_0$ is `FALSE`). Since this one-sample framework

works for any average (not just proportions), another form of this test that is often encountered is $H_0 : \mu = \mu_0$ where μ is the mean of the population (corresponding to the sample average $\bar{x}$).

- In the second version we hypothesize a relationship between two populations, such as that both populations have the same mean or median (or proportion or standard deviation, etc.). These *TWO sample hypotheses tests* have a NULL hypothesis of the form $H_0 : \mu_1 = \mu_2$ or $H_0 : p_1 = p_2$ (or $H_0 : Median_1 = Median_2$ or $H_0 : \sigma_1 = \sigma_2$, etc.) and of course the ALTERNATIVE hypothesis is still $H_1 : H_0$ is `FALSE`.

For each of the following scenarios, state appropriate hypotheses H_0 and H_1. For each scenario, also state in simple terms what the claim of the NULL hypothesis is.

Be sure to carefully define any parameters you refer to.

(a) A health survey asked individuals to report the number of times they exercised each week. Researchers were interested in determining if the proportion of individuals who exercised at least 100 minutes per week differed between people who live in the condos vs people who do not live in condos.

REPLACE THIS TEXT WITH YOUR ANSWER

(b) A study was conducted to examine whether a baby is born prematurely/early (i.e., before their due date) to whether or not the baby's mother smoked while she was pregnant.

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Nintendo is interested in whether or not their online advertisements are working. They record whether or not a user had seen an ad on a given day and their amount of spending on Nintendo products in the next 48 hours. They are interested in determining if there is an association between whether or not the user say an ad and their expenditures.

REPLACE THIS TEXT WITH YOUR ANSWER

(d) Based on results from a survey of graduates from the University of Toronto, we would like to compare the median salaries of graduates from the statistics and graduates of mathematics programs.

REPLACE THIS TEXT WITH YOUR ANSWER

Part 4: OPTIONAL but STRONGLY Recommended for Practice

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

Question 8

Complete this One Sample Hypothesis Testing Practice Quiz using this Rmd file

Question 9

Complete this Two Sample Hypothesis Testing Practice Quiz using this Rmd file

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***