STA130H1S – Winter 2024

Week 5 Practice Problems

N. Moon and J. Speagle and Christopher Li

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

IMPORTANT - SETUP

```r
# Instructions for how students should define tests
STUDENT_NUMBER <- 1010057028;  # replace 130 by your real student number
```

Question 1: Estimating the Speed of Light

Background: In the late 1800s, physics theories assumed that all waves must move through some substance (a "medium"). Just as waves on a pond travel across, well, water, and sound waves travel through (mostly invisible) air, physicists thought that light must also travel through an invisible medium. They called this hypothesized medium the "luminiferous aether". This idea implied that you should be able to measure differences in the speed of light depending on how fast you are moving through this luminiferous aether in a particular direction. Trying to measure this effect, and therefore proving the existence of this luminiferous aether, became one of the main experimental goals in the 1800s.

The biggest issue with measuring this distance is that the speed of light is huge (roughly 300,000 km/s!) related to most speeds we can achieve on Earth. However, it turns out that we can exploit the fact that Earth orbits around the Sun at a speed of roughly 30 km/s to derive a natural experiment. This implies that light travelling "with" the flow of the aether will get a boost in speed from the Earth's orbit around the Sun, while light moving in the opposite direction ("against" the flow) will be slightly slowed down. Light moving perpendicular to the flow would be unaffected.

American physicists Albert A. Michelson and Edward W. Morley exploited this fact to try and precisely measure the speed of light in various directions in order to measure this expected difference. The bulk of their experiments took place between April and July in 1887 at what is now Case Western Reserve University in Cleveland, Ohio. Their failure to detect any shift at all became one of the foundational results that Albert Einstein used to justify his theory of Special Relativity.

In this question, we will be working with some earlier data from Michelson taken in 1879 (8 years before the famous experiment with Morley). This data is one of the default datasets loaded into R as `morley`.

```r
# load in and rename the data
michelson <- tibble(morley)
```

(a) Using the `help(morley)` function, please list the name of the variable that records the speed of light measurements and what units they are measured in.

```r
# Replace NULL below by your answer (it should be a character, e.g., "name")
Q1a_varname <- "Speed"

# Replace NULL below by your answer from the following list:
# "mph", "km/s", "m/s", "km/h"
Q1a_units <- "km/s"
```

(b) The current accepted estimated for the speed of light is around the constant value of $c = 299792$ km/s. Write down a hypothesis test below for whether the true speed of light v is equal to this value or is different from this value and comment on whether this is a one-sided or two-sided hypothesis test.

$$H_0 : v = c, H_\alpha : v! = c$$

(c) Our **test statistic** $\hat{v}$ will be the mean of the speed of light measurements taken from the Michelson data. In other words,

$$\hat{v} = \text{mean}(\{x_1, x_2, ..., x_n\}) = \frac{1}{n}\sum_{i=1}^{n} x_i \, .$$

What is the value of the test statistic for the Michelson data? What about the assumed parameter value under the null hypothesis (in the same units/scale as the Michelson data)?

Note: You can access individual columns using `tibble$varname` syntax. Remember to also check the units and/or exact values of the Michelson data, in case you need to multiply or add any constants.

```r
# Replace NULL below by your answer for the test statistic
Q1c_vhat <- mean(michelson[["Speed"]])

# Replace NULL below by your answer for the parameter value under the null
Q1c_v <- 792
```

(d) Below is R code that simulates $N = 10,000$ trials for the mean velocity that you might measure in a random sample of $n = 100$ observations **under the null hypothesis**, assuming a measurement error of $\sigma = 80$ km/s.

*Note: **You must include the `set.seed(STUDENT_NUMBER)` line in your code and run it every time you run the cell below to ensure the sample is fully reproducible. Because each of you will have a different student number, each student will have slightly different answers.**

```r
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)     # Make sure that this is your student number
```

```
## [1] 1010057028
```

```r
                            # If it isn't correct, go to the SETUP section and
                            # store the correct value of your student number in
                            # the STUDENT_NUMBER variable (and run the code chunk)
                            # then come back to this question
                            # Make sure you don't share screenshots of your
                            # solution that share your student number with others
set.seed(STUDENT_NUMBER)  # REQUIRED so the random sample is reproducible!

# setup
n_trials <- 10000  # number of simulations/trials
n_sample <- 100  # number of observations in random sample
mu <- 792  # in units of km/s
sigma <- 80  # in units of km/s

# simulate!
vhat_simulations <- rnorm(n_trials, mean = mu, sd = sigma)
for (i in 1:n_trials){
  vel_sim <- rnorm(n_sample, mean = mu, sd = sigma)
  vhat_sim <- mean(vel_sim)
  vhat_simulations[i] <- vhat_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

What is the value of the 100th simulated test statistic $\hat{v}_{\text{sim}}$?

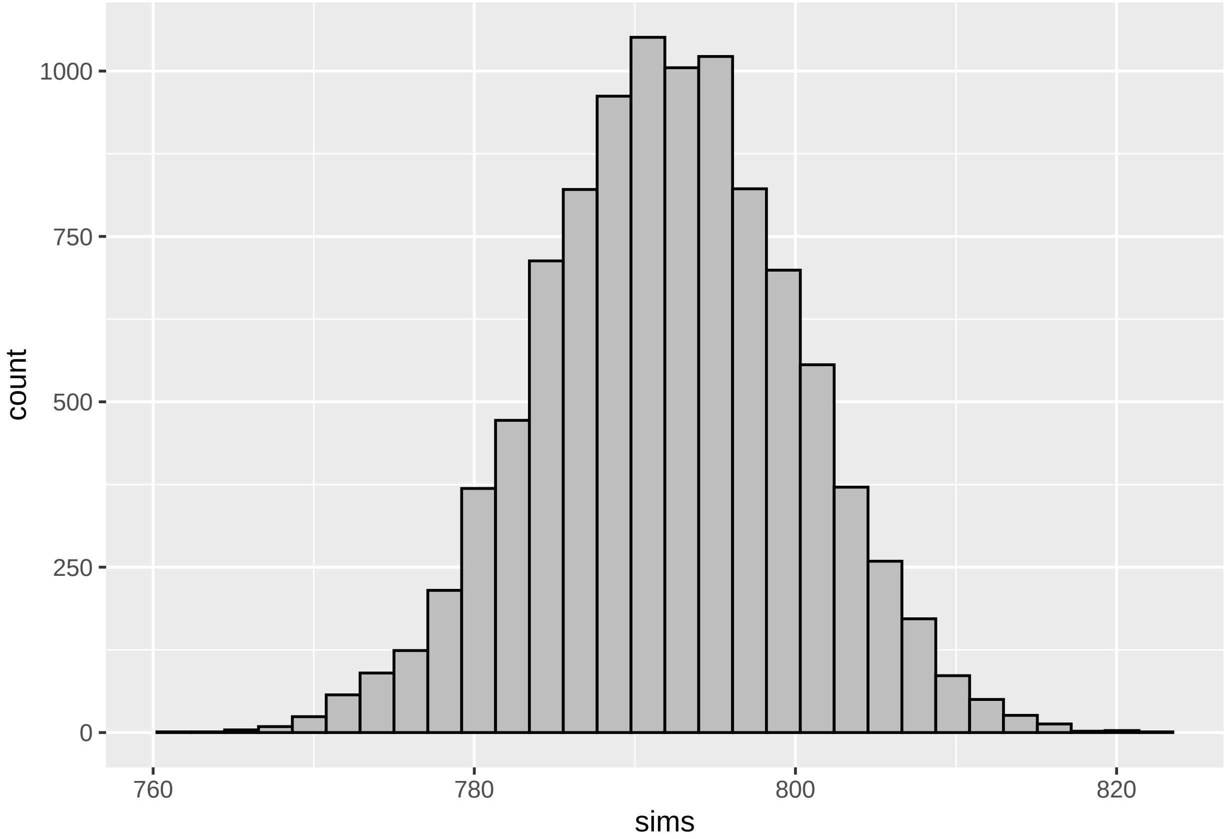
Hint: You can access the i-th element of a vector using `x[i]` syntax.

```r
# Replace NULL below by your answer for the simulated test statistic value
Q1d_vhat_sim_100 <- vhat_simulations[100]
```

(e) [**OPTIONAL**] Use `geom_boxplot()` and `geom_histogram()` to plot the distribution of the simulated test statistics (i.e. the sampling distribution).

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

```r
tibble(sims = vhat_simulations) %>% ggplot(aes(x=sims)) +
  geom_histogram(color = "black",
         fill = "gray")
```



(f) Compute the p-value based on the null hypothesis H_0 using the values of $\hat{v}_{\text{sim}}$ you computed from the sampling distribution, the observed test statistic $\hat{v}$, and the assumed true value under the null hypothesis v.

Hint: Remember you can take advantage of "coercion" in R to compute the number of objects that satisfy a logical condition and `abs()` to compute the absolute value of a value/vector.

```r
# Replace NULL with your answer below
Q1f_pvalue <- sum(abs(vhat_simulations - 792) >= abs(Q1c_vhat - Q1c_v)) / n_trials
```

Based on your result, write 1-3 sentences on what this implies about how likely we would observe data similar to what Michelson collected assuming that the accepted modern value of the speed of light c is correct.

> There is a probability of 0 that we get the results we achieved through chance alone.

(g) **[OPTIONAL]** Can you think of at least one potential issue with either the null hypothesis, simulation procedure, or the Michelson data which may influence the validity of the calculation you performed above?

> The data Michelson observed may not be correct, since the instruments he had were not as sophisticated as the ones we have now.

(h) Let's assume that you discover there is a bias in the Michelson data that causes the reported velocities to be systematically overestimated (too large) by 40 km/s. What is the new p-value after you correct for this bias?

```
# Replace NULL with your answer below
Q1h_pvalue <- sum(abs(vhat_simulations - 792) >= abs(Q1c_vhat - 40 - Q1c_v)) / n_trials
```

Based on this new p-value, would you reject the null hypothesis in favour of the alternative given an α level (i.e. $p < \alpha$) of $\alpha = 0.01$?

```
# Replace NULL with your answer below (either TRUE or FALSE)
Q1h_reject <- FALSE
```

Question 2: Observing the Luminiferous Aether

While the potential bias discussed in Q1h might at first appear alarming, for the purposes of Michelson's experiments this isn't actually too alarming. This is because Michelson was interested in measuring the *differences* in the speed of light in multiple orientations, which means that a constant offset would just cancel out. As such, let's now turn our attention to trying to see whether or not there was a discernable difference in Michelson's data.

(a) For the purposes of this problem set, let's focus our efforts on just two of the 5 orientations (experiments) Michelson ran. Please modify the original `michelson` data to contain only the $n = 40$ results from the first two experiments and store it as a new tibble in the variable below.

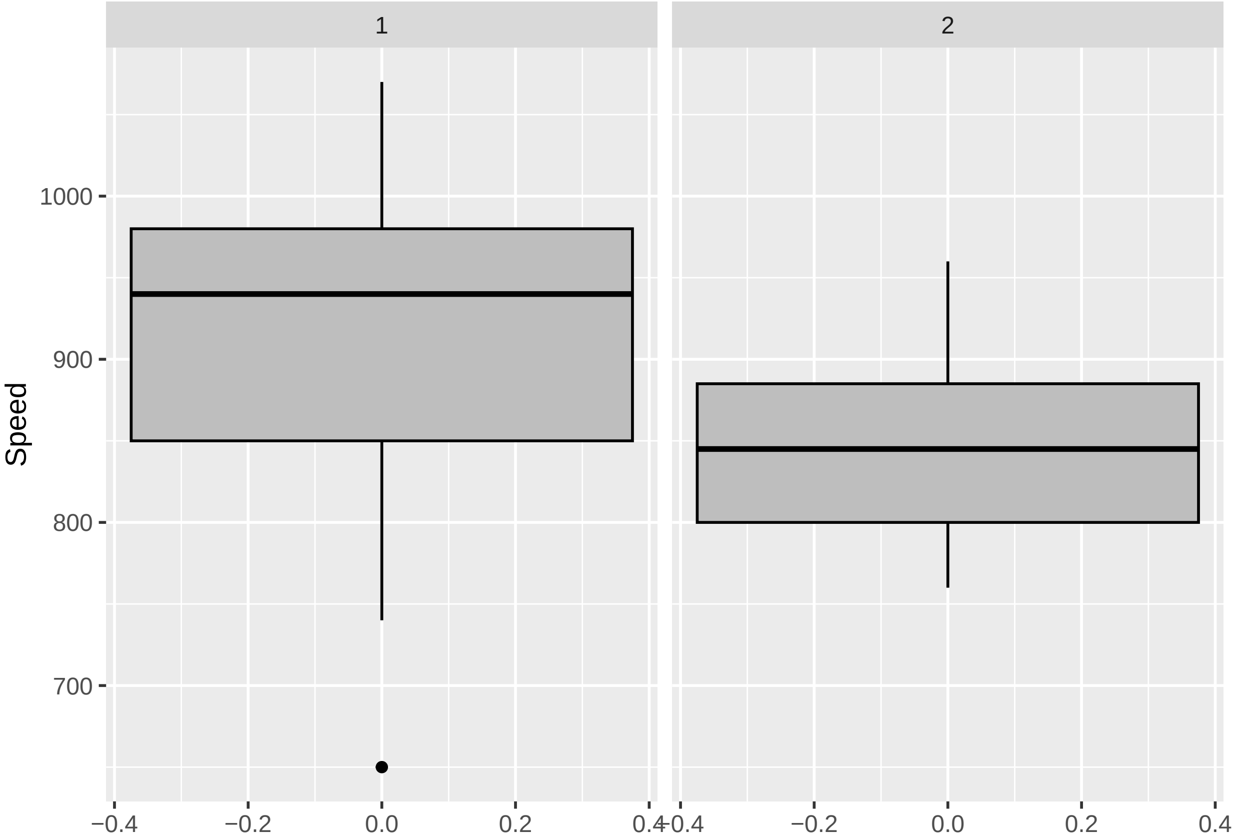
```
# Replace NULL with your answer below (should be a modified tibble)
Q2a_michelson_2samp <- michelson %>% filter(Expt == 1 | Expt == 2) %>% head(40)
```

(b) Write down a 2-sample hypothesis test where the null states that the speed of light in orientation 1 (v_1) and orientation 2 (v_2) are the same and the alternative states that they are different.

$$H_0 : v_1 = v_2, H_\alpha : v_1! = v_2$$

(c) [**OPTIONAL**] Plot the distribution of the speed of light measurements from the two experiments side-by-side using the plotting function of your choice combined with `facet_wrap()`.

```
# add your plots below
Q2a_michelson_2samp %>% ggplot(aes(x=Speed)) +
  geom_boxplot(color = "black",
           fill = "gray") +
facet_wrap(~Expt) +
  coord_flip()
```



Comment briefly on any noticeable features and how similar/different the distributions appear to be.

> These 2 distributions seem to have very different medians. In addition, experiment 1 seems to have a larger range than experiment 2, and it has an outlier at a speed of around 650.

(d) Our **test statistic** $\Delta\hat{v}$ will be the difference in the mean of the speed of light measurements taken from experiments 1 and 2. This is calculated below.

```
# Note: including the .groups="drop" option in summarise() will suppress
# a friendly warning that R normally prints out:
# "`summarise()` ungrouping output (override with `.groups` argument)".
delta_vhat <-
  michelson %>%
  filter(Expt <= 2) %>%
  group_by(Expt) %>%
  summarise(means = mean(Speed), .groups="drop") %>%
  summarise(value = diff(means)) %>%
  as.numeric()


print(delta_vhat)
```

```
## [1] -53
```

Below is R code that simulates $N = 1000$ values of the test statistic $\Delta\hat{v}_{\text{sim}}$ **under the null hypothesis** using a permutation test. In this test, we assume that our groups are identical under our null hypothesis. Mixing the two groups together, randomly generating new groups with the same sizes, and then recomputing our test statistic each time therefore should allow us to simulate values from the sampling distribution provided our sample size is large enough.

```r
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)     # Make sure that this is your student number
```

```
## [1] 1010057028
```

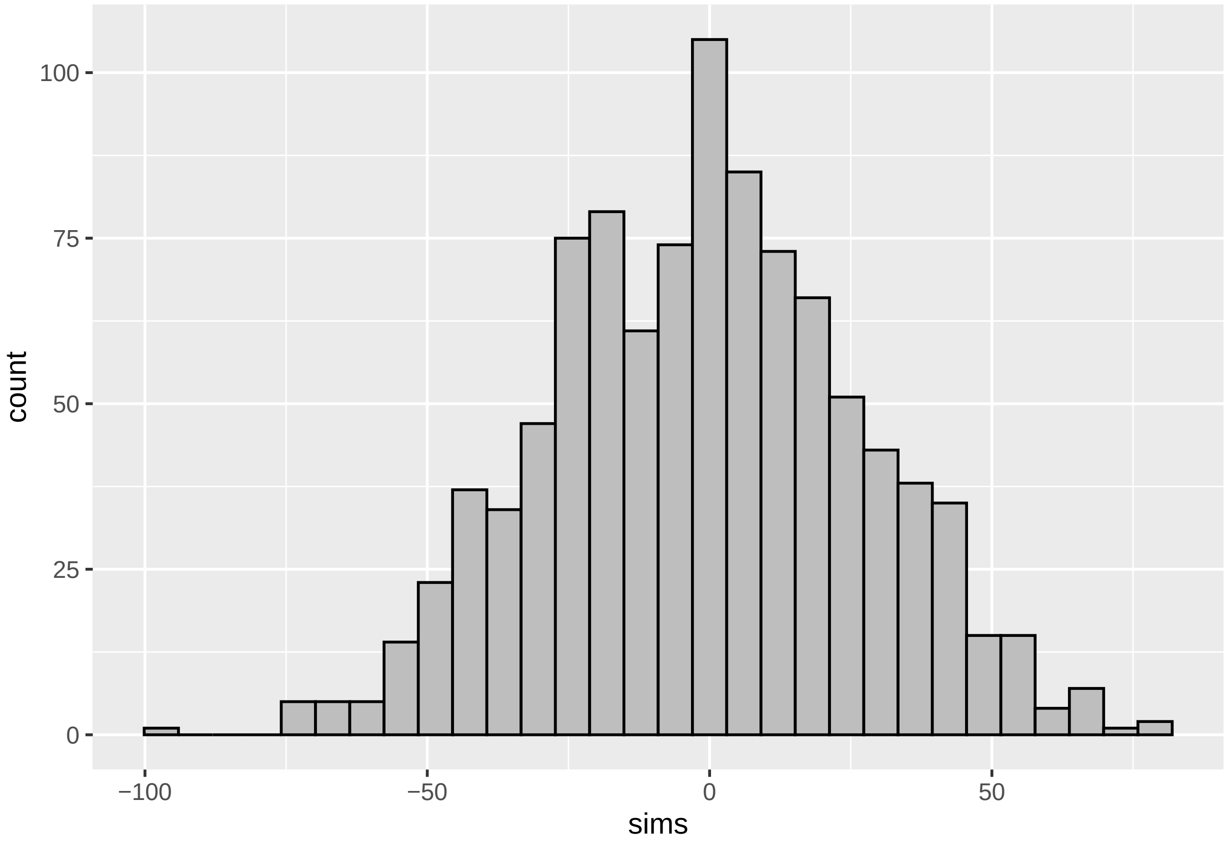
```r
                          # If it isn't correct, go to the SETUP section and
                          # store the correct value of your student number in
                          # the STUDENT_NUMBER variable (and run the code chunk)
                          # then come back to this question
                          # Make sure you don't share screenshots of your
                          # solution that share your student number with others
set.seed(STUDENT_NUMBER)  # REQUIRED so the random sample is reproducible!

# setup
n_trials <- 1000  # number of permutations

# simulate!
delta_vhat_simulations <- rep(NA, n_trials)
for(i in 1:n_trials){
  # perform a random permutation
  simdata <-
    michelson %>%
    filter(Expt <= 2) %>%
    mutate(Expt = sample(Expt, replace=FALSE))
  # compute the simulated test statistic
  delta_vhat_sim <-
    simdata %>%
    group_by(Expt) %>%
    summarise(means = mean(Speed), .groups="drop") %>%
    summarise(value = diff(means)) %>%
    as.numeric()
  # store the simulated value
  delta_vhat_simulations[i] <- delta_vhat_sim
}
### DO NOT CHANGE THE CODE ABOVE ####
```

Based on our simulations, how many simulations give values of $\Delta\hat{v}_{\text{sim}} > 0$? How about < 0?

```r
# Replace NULL with your answer below
Q2d_pos <- sum(delta_vhat_simulations > 0)
Q2d_neg <- sum(delta_vhat_simulations < 0)
```

(e) [OPTIONAL] Plot the distribution of the simulated test statistics (i.e. the sampling distribution) for the permutation test computed above.

```r
# add your plots below
tibble(sims = delta_vhat_simulations) %>% ggplot(aes(x=sims)) +
  geom_histogram(color = "black",
           fill = "gray")
```



(f) Compute the p-value based on the null hypothesis H_0 using the values of $\Delta\hat{v}_{\text{sim}}$ you computed from the sampling distribution, the observed test statistic $\Delta\hat{v}$, and the assumed true value under the null hypothesis $\Delta v = v_1 - v_2$.

```
# Replace NULL with your answer below
Q2f_pvalue <- sum(abs(delta_vhat_simulations) >= abs(-53)) / n_trials
```

Based on your result, write 1-3 sentences on what this implies about how likely we would observe data similar to what Michelson collected assuming that the speed of light is the same in each direction.

There is a probability of 0.055 that we get a difference of 53 or higher by pure chance alone.

Based on this p-value, would you reject the null hypothesis in favour of the alternative given an α level (i.e. $p < \alpha$) of $\alpha = 0.05$?

```
# Replace NULL with your answer below (either TRUE or FALSE)
Q2f_reject <- FALSE
```

(g) **[OPTIONAL]** Can you think of at least one potential issue with either the null hypothesis, simulation procedure, or the Michelson data which may influence the validity of the calculation you performed above?

We only have 20 values of each, when idealy we should have 30 for the large counts condition.

[OPTIONAL] Question 3: Social Media

Note: This entire question is optional.

There have been many questions regarding whether or not usage of social media increases anxiety levels.

For example, does regular viewing of Instagram or WeChat posts create an unattainable sense of life success and satisfaction? Does procrastinating by watching YouTube or TikTok videos or reading Reddit threads contribute unnecessary stress from deadline pressure?

To try and answer some of these questions, 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 were scored through using series of questions. Higher scores on the questionnaire suggest higher student anxiety.

```
# define our initial dataset
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)

# preview our data
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) Assume that each group can be characterized by the **median** of their set of anxiety scores. Let’s call the median anxiety scores A_{high} among the “high” social media usage group and A_{low} among the “low” social media usage group. Write down the null and *1-sided* alternative hypotheses H_0 and H_1 in math notation. Then, restate the claims of the null/alternative hypotheses in simpler language.

Note: Depending on your choice of alternative hypothesis, your answers below might involve either 1-sided or 2-sided tests. Please make sure all your calculations remain consistent with your initial choice.

$$H_0 : a_l = a_h, H_a : a_l < a_h$$

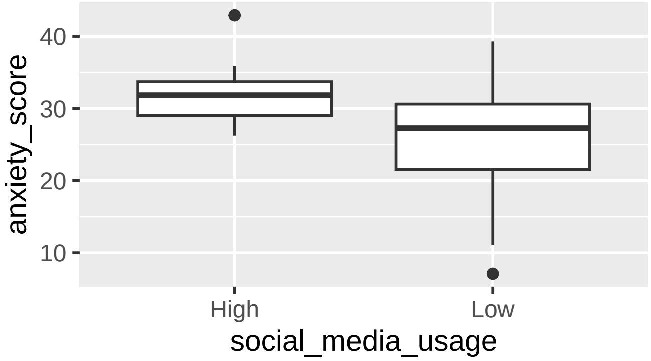
(b) Revisit your statements regarding the null hypotheses above with **confounding variables** in mind. In particular, consider that social media usage may be a self selecting process, with social media users potentially already more anxious people on average regardless of their social media usage. If we make a determination about the null hypothesis, are we actually addressing the initial 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)?

Note: This distinction is often referred to using the cautionary phrase “correlation does not imply causation”.

We are more looking at whether or not there’s an observable difference between these two groups.

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

```
# Code your answers here
anxiety_data %>%
  ggplot() +
  aes(x=social_media_usage, y=anxiety_score) +
  geom_boxplot()
```



The median anxiety level of low social media users is lower than the median anxiety level of higher social media usage. In addition, the range of low social media users is vastly higher than the range of high social media users.

What do these data visually suggest regarding the claim that the **median** anxiety level is different for those who use social media more than 2 hours per day compared to those who use social media less than 2 hours per day?

It is lower.

(d) Compute the difference in the medians between the two samples (i.e. the 2-sample test statistic).

```
# code your answer below

# low - high
median_diff <- anxiety_data %>% group_by(social_media_usage) %>% summarize(medians=median(anxiety_score
```

(e) Compute the sampling distribution using $N = 1000$ simulations/permutations.

```
### DO NOT CHANGE THE CODE BELOW ####
print(STUDENT_NUMBER)     # Make sure that this is your student number
```

```
## [1] 1010057028
```

```
                          # If it isn't correct, go to the SETUP section and
                          # store the correct value of your student number in
                          # the STUDENT_NUMBER variable (and run the code chunk)
                          # then come back to this question
                          # Make sure you don't share screenshots of your
                          # solution that share your student number with others
set.seed(STUDENT_NUMBER)  # REQUIRED so the random sample is reproducible!
### DO NOT CHANGE THE CODE ABOVE ####

# code your answer below
n_trials <- 1000  # number of permutations

# simulate!
delta_ahat_simulations <- rep(NA, n_trials)
```

```
for(i in 1:n_trials){
  # perform a random permutation
  simdata <-
    anxiety_data %>%
    mutate(social_media_usage = sample(social_media_usage, replace=FALSE))

  # compute the simulated test statistic
  delta_ahat_sim <-
    simdata %>% group_by(social_media_usage) %>% summarize(medians=median(anxiety_score), .groups="drop
  # store the simulated value
  delta_ahat_simulations[i] <- delta_ahat_sim
}
```

(f) Compute the p-value based on the null hypothesis H_0.

```
# Replace NULL with your answer below
Q3f_pvalue <- sum(delta_ahat_simulations <= median_diff) / n_trials
```

Based on this p-value, would you reject the null hypothesis in favour of the alternative given an α level (i.e. $p < \alpha$) of $\alpha = 0.05$?

```
# Replace NULL with your answer below (either TRUE or FALSE)
Q3f_reject <- TRUE
```

(g) Do these data support the claim that the **median** anxiety level is different for those who use social media more than 2 hours per day compared to those who use social media less than 2 hours per day? How about the claim that "usage of social media increases anxiety levels"?

Yes it does say that there is a difference, however we can't make any claims about causality.