STA130H1F – Winter 2021

Week 5 Problem Set

S.Caetano & N. Moon Yixing Xu

Instructions

How do I hand in these problems for the 11:59 a.m. ET, February 11 deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/550922) by 11:59 a.m. ET, on February 11. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[Question 1]

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?

The null hypothesis is that the defendant is innocent.

The alternative hypothesis is that the defendant is not innocent.

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

Rejecting the null hypothesis would mean that there is convincing evidence to conclude that the null hypothesis that the defendant is innocent is wrong and that we can accept the alternative hypothesis that the defendant is not innocent.

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

Failing to reject the null hypothesis would mean that there is not enough evidence to reject the hypothesis that the defendant is innocent, and thus there is also not enough convincing evidence for the alternative hypothesis and we cannot accept the alternative hypothesis.

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

A type II error would be failing to reject the null hypothesis that the defendant is innocent when the alternative hypothesis that the defendant is guilty or not innocent is true.

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

A type I error would be concluding that the defendant is not innocent when the defendant is innocent.

[Question 2]

There have been many questions regarding whether or not usage of social media increases anxiety levels. A study was conducted to study the relationship between social media usage and student anxiety. The following scores are measures of anxiety levels of students on a Monday. Note that higher scores indicate higher anxiety. Moreover, if a student used social media more than 1 hour per day then their usage was categorized as "High".

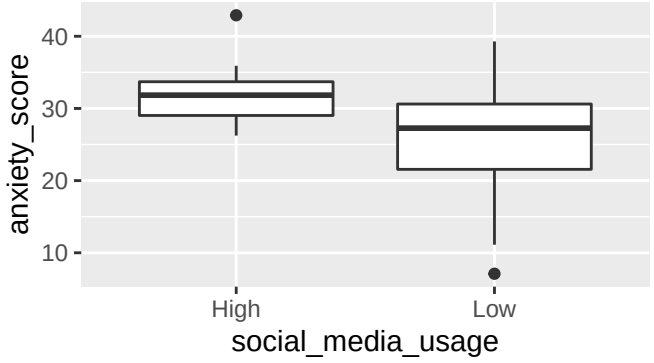
```
social_media_usage <- c(rep("Low", 30), rep("High", 16));
anxiety_score <- c(24.64, 39.29, 16.32, 32.83, 28.02,
                   33.31, 20.60, 21.13, 26.69, 28.90,
                   26.43, 24.23, 7.10,  32.86, 21.06,
                   28.89, 28.71, 31.73, 30.02, 21.96,
                   25.49, 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",...
## $ anxiety_score      <dbl> 24.64, 39.29, 16.32, 32.83, 28.02, 33.31, 20.60,...
```

(a) Construct boxplots of `anxiety_score` for each type of test. Write 2-3 sentences comparing the distributions of anxiety scores for the two types of test.

The median anxiety score for those who reported high social media usage is higher than the median anxiety score for those with low social media usage. Also, the variability or range of anxiety scores for those who reported low social media usages is larger/wider than the distribution of the scores for those with high social media usage.

```
anxiety_data %>% ggplot(aes(x = social_media_usage, y = anxiety_score)) +
  geom_boxplot()
```



(b) 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?

(i) State the hypotheses you are testing (be sure to define any parameters you refer to).

The null hypothesis is there is no median difference in anxiety levels between high and low social media usages

The alternative hypothesis is there is a median difference in anxiety levels between high and low social media usages.

We will use a significance level of 0.05 to assess this data.

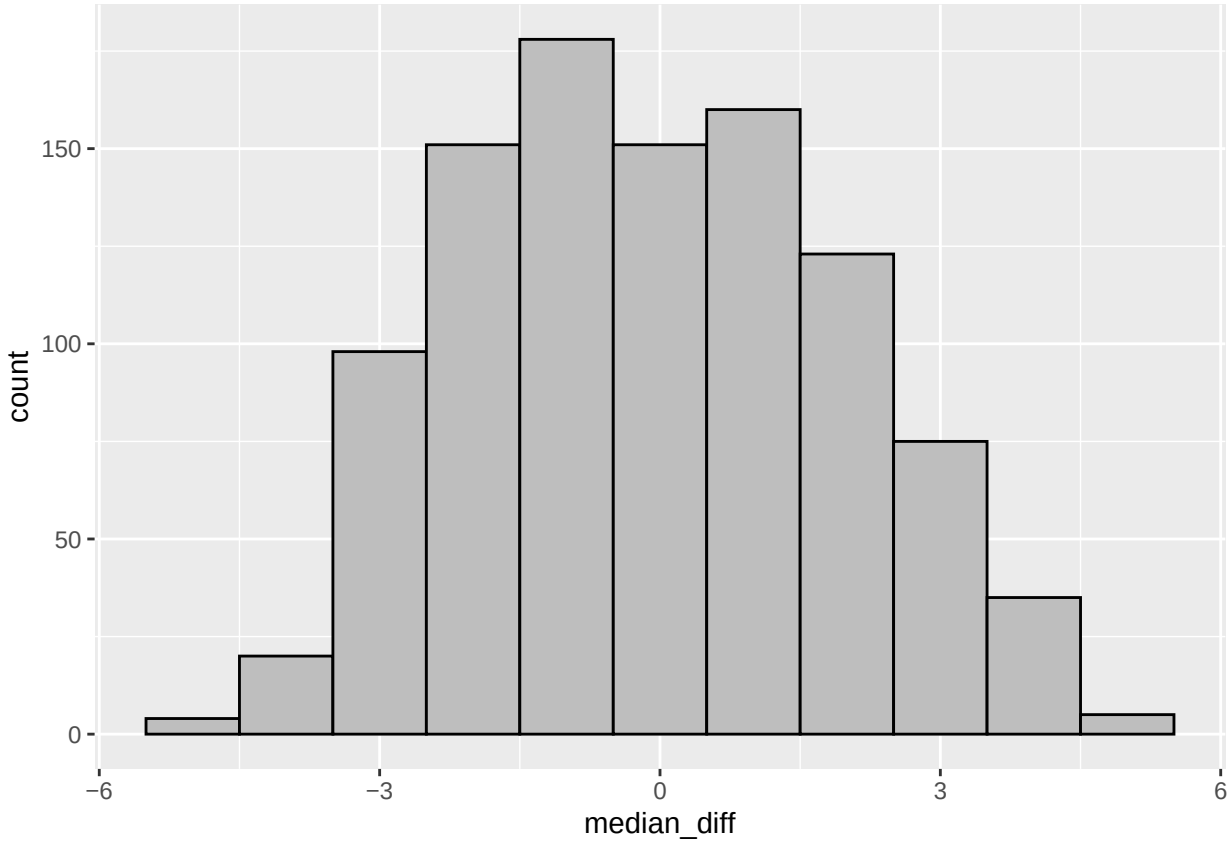
(ii) Look at the code below and write a few sentences explaining what the code inside the for loop is doing and why.

The code below is a predicted histogram of the median differences in anxiety levels between high and low social media usages that we might see in samples of size 46 there is a real difference.

```
# 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")
```




```
# Calculate p-value
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
```

(iii) Write a few sentences summarizing your conclusions. Be sure to interpret the p-value carefully and to clearly address the research question.

Since the p-value of 0.009 is less than our significance level of 0.05, we will reject our null hypothesis that there is not a median difference in anxiety levels and accept the alternative hypothesis that there is a median difference in anxiety levels between low and high social media usages.

[Question 3]

(Adapted from "Biostatistics for the Biological and Health Sciences") The table below presents data from a random sample of passengers sitting in the front seat of cars involved in car crashes. Researchers are interested in whether the fatality rates (i.e. 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 R command `rep`. This function creates a vector which replicates its first argument the number of times indicated by its second argument. For example, the `rep("hello", 5)` creates a vector with 5 elements, each of which is "hello". Run the code chunk below to load a tidy tibble called `data` which you'll use for the remainder of this question.

```
library(tidyverse)

data2 <- 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)))

glimpse(data2)
```

```
## Rows: 20,408
## Columns: 2
## $ group   <chr> "airbag", "airbag", "airbag", "airbag", "airbag", "airbag",...
## $ outcome <chr> "dead", "dead", "dead", "dead", "dead", "dead", "dead", "de...
```

(a) State appropriate hypotheses to compare the proportions of deaths in cars with and without airbags. Be sure to define any parameters you refer to in your hypotheses.

The null hypothesis is that there is no difference in death rates/proportions between passengers in cars with airbags and cars without airbags.

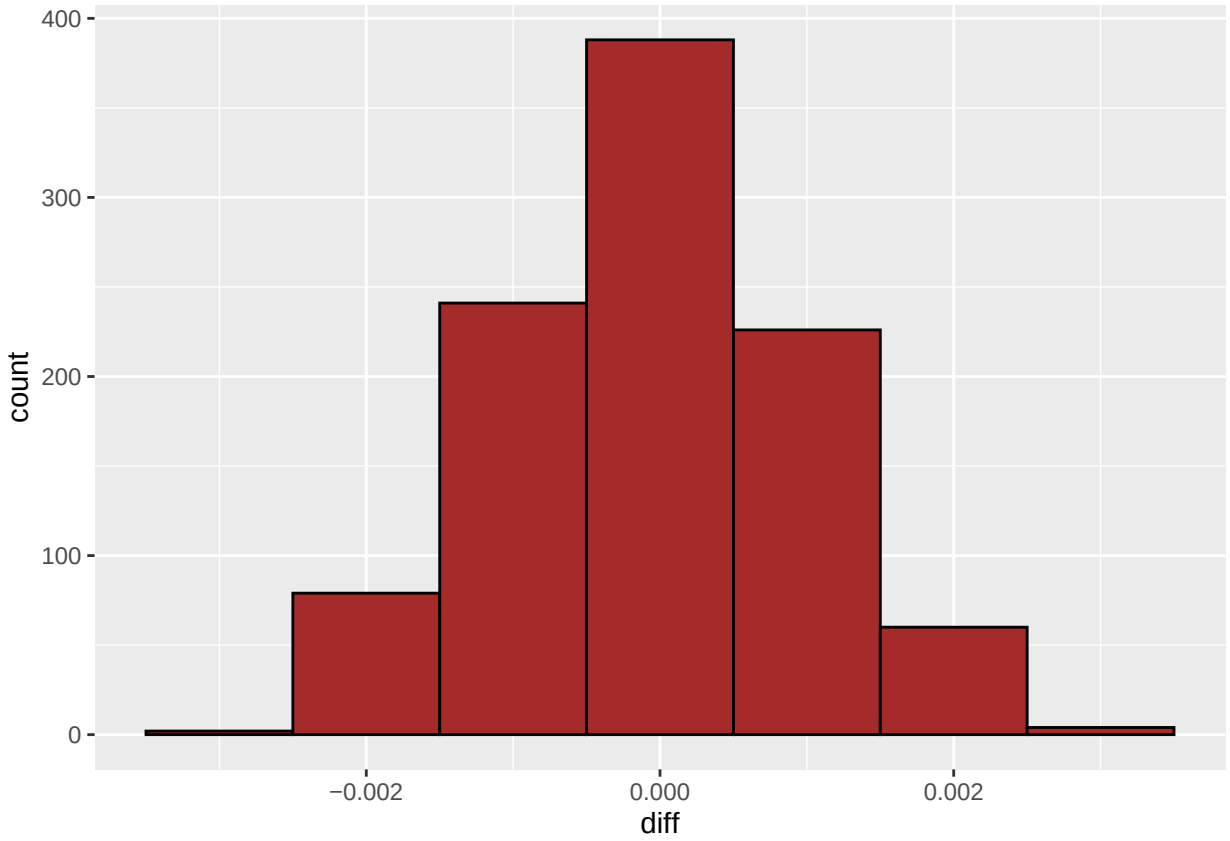
The alternative hypothesis is that there is a difference in death rates/proportions between passengers in cars with airbags and cars without airbags.

(b) Carry out a hypothesis test for the hypotheses stated in part (a).

```
set.seed(103) # Replace the number in the parentheses with the 1st, 3rd, and 5th
# digits in your student number.

repetitions <- 1000
sim_values <- rep(NA, repetitions) #replicates numeric values
#for a specific number of times
```

```
test_stat2 <- data2 %>% group_by(group) %>%
  summarise(n = n(), deaths = sum(outcome == "dead"),
```

```
            prop = (deaths/n))%>%
  summarise(val = diff(prop))
test_stat2 <- as.numeric(test_stat2)

for (i in 1:repetitions){
  simdata2 <- data2 %>% mutate(group = sample(group))

  sim_value1 <- simdata2 %>%
    group_by(group) %>%
    summarise(n = n(), deaths = sum(outcome == "dead"),
              prop = (deaths/n))%>%
    summarise (val = diff(prop))

  sim_values[i] <- as.numeric(sim_value1)
}

sim2 <- tibble(diff = sim_values)
sim2 %>% ggplot(aes(x= diff)) +
  geom_histogram(binwidth = 0.001, color = "black", fill = "brown")
```




```
extremes <- sim2 %>%
  filter(abs(diff) >= abs(test_stat2)) %>%
  summarise(n())
p_val <- as.numeric(extremes/repetitions)
p_val
```

```
## [1] 0.041
```

(c) Based on your answer in part (b), would you reject the null hypothesis at the 0.1 significance level?

Since I got a p value of 0.049, which is smaller than the significance level of 0.1, I will have to reject the null hypothesis.

(d) Based on your answer in part (c), what kind of error did you possibly make?

I could possibly be making a type I error, where I reject the null hypothesis even though the null hypothesis is true.

[Question 4]

In class we've talked about two kinds of hypothesis tests. In the first kind (week 4) we talked about how to test whether a proportion is equal to a specific value, with hypotheses of the form: $H_0 : p = p_0$ vs $H_A : p \neq p_0$. In this week's class (week 5), we talked about how to test if there is a difference between two groups (e.g. a difference in the means of two groups, the medians of two groups, or proportions of two groups). A test for the difference between the means of two groups takes the form: $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$.

For each of the following scenarios, state appropriate hypotheses H_0 and H_A. 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.

$H_0 : p - p_0 = 0$

The null hypothesis is that there is no difference in 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.

$H_A : p - p_0 \neq 0$

The alternative hypothesis is that there is a difference in 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.

, where p is the proportion of individuals who exercised at least 100 minutes per week and live in condos, and p_0 is the proportion of individuals who exercised at least 100 minutes per week and did not live in condos.

(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.

$H_0 : p - p_0 = 0$

The null hypothesis is that there is no difference between the proportion of babies who are born prematurely when the baby's mother smoked while pregnant and when a baby's mother did not smoke while pregnant.

$H_A : p - p_0 \neq 0$

The alternative hypothesis is that there is a difference between the proportion of babies who are born prematurely when the baby's mother smoked while pregnant and when a baby's mother did not smoke while pregnant.

, where p is the proportion of babies born prematurely when the baby's mother smoked while she was pregnant, and p_0 is proportion of babies born prematurely when the baby's mother did not smoke while she was pregnant.

(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.

$H_0 : \mu_1 - \mu_2 = 0$

The null hypothesis is that there is no difference between the average amount spent on Nintendo products within 48 hours of seeing an ad vs not seeing an ad.

$H_A : \mu_1 - \mu_2 \neq 0$

The alternative hypothesis is that there is a difference between the average amount spent on Nintendo products within 48 hours of seeing an ad vs not seeing an ad.

, where μ_1 is the average amount of spending on Nintendo products for a user within 48 hours of seeing an ad and μ_2 is the average amount of spending on Nintendo products within 48 hours if the user did not see an ad.

(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.

$H_0 : m_1 - m_2 = 0$

The null hypothesis is that there is no difference between the median salaries of graduates from the statistics and graduates of mathematics programs.

$H_A : m_1 - m_2 \neq 0$

The alternative hypothesis is that there is a difference between the median salaries of graduates from the statistics and graduates of mathematics programs.

, where m_1 is the median salaries of graduates from statistics programs and m_2 is median salaries of graduates from mathematics programs.

Part 2

For this week, you can complete the required task as a written assignment, or you can submit an oral response (i.e. just sound or sound and video). While it is optional this time, there will be another problem set part #2 that you will be required to complete an oral response, and therefore it is a good idea to try and practice it now!

Write or say a short summary of ONE of the following studies. A summary *must* include the following components:

1. Context of the problem (i.e. introduction)
2. Summary of the methods. E.g. State hypotheses; define the test statistic; etc.
3. Summary of the findings (i.e. the main results)
4. Conclusion (e.g. main finding(s), significance, etc.)
5. Limitations (*optional for now*, but good practice and required for the final project; e.g. study design issues etc.)

(a) A health survey asked 200 individuals aged 20-45 living in Toronto to report the number minutes they exercised last week. Researchers were interested in determining whether the average duration of exercise differed between people who consume alcohol and those who do not consume alcohol. Assume the researchers who conducted this study found that people who drank alcohol exercised, on average, 20 minutes per week. In contrast, people who did not drink alcohol exercised 40 minutes per week, on average. The researchers reported a p-value of 0.249.

(b) A study was conducted to examine whether the sex of a baby is related to whether the baby's mother smoked while she was pregnant. The researchers obtained a sample of 10,000 births from a birth registry of all children born in Ontario in 2018. The researchers found that based on this sample 4% of mothers reported smoking during pregnancy and 52% of babies born to mothers who smoked were male. In contrast, 51% of babies born to mothers who did not smoke were male. The researchers reported a p-value of 0.50.

(c) Based on results from a survey of graduates from the University of Toronto, we would like to compare the median salaries of graduates of statistics programs and graduates of computer science programs. 1,000 recent graduates who completed their bachelor's degree in the last five years were included in the study; 80% of the respondents were female and 20% were male. Among statistics graduates, the median reported income was $76,000. Among computer science graduates, the median reported income was $84,000. The researchers reported a p-value of 0.014.

(d) A team of researchers were interested in understanding millennials' views regarding housing affordability in Toronto. The team interviewed 850 millennials currently living in Toronto. 84% reported that they felt housing prices were unaffordable in the city. Suppose the researchers were interested in testing whether this proportion was different from a study published last year, which found that 92% of millennials reported that housing costs were unaffordable. The researchers reported a p-value of 0.023.

This study is a two-sample hypothesis test that compares the proportion of millennials in Toronto who believe housing prices are unaffordable this year vs last year. The null hypothesis is that there is no difference in the proportion of millennials in Toronto who believe housing prices are unaffordable this year vs last year. The alternative hypothesis is that there is a difference in the proportion of millennials in Toronto who believe housing prices are unaffordable this year vs last year.

The test statistic is the difference of 8% or 0.08 from the samples in the studies conducted by the team of researchers this year and last year. Last year, 92% of Toronto millennials found housing unaffordable,

as opposed to 84% this year. What the researchers concluded was that, since their p-value of 0.023 is less than the significance level of 0.05, there was convincing evidence to reject the null hypothesis and accept the alternative hypothesis that there indeed is a difference in the proportions of Toronto millennials who find housing unaffordable since last year.

However, it is reasonable to acknowledge that there is a possibility of a Type I error, because we may have rejected the null hypothesis when we should have accepted it. Also, we do not know much about the sample size of last year's study, nor do we know where, in Toronto, the team of researchers conducted their study last year and this year. The area or neighborhoods in Toronto in which these studies were conducted could play a major role in the difference in proportions.

(e) Suppose a drug company was interested in testing a new weight-loss drug. They enrolled 20,000 participants and assigned 10,000 to take their new drug, SlimX, and 10,000 to take a placebo. The researchers found that over 2 months, participants who took SlimX lost, on average, 5 lbs. In comparison, the control group lost 4.5 lbs during the same time. The researchers reported a p-value of <0.0001.

Some things to keep in mind

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).
- Be specific. A good principle when responding to a prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt).
- You *cannot* directly copy the sentences from the explanation provided. You also *cannot* use quotation marks in your response. You *must* rewrite or restate these responses in your own words (i.e. paraphrase).

For those who choose to do an oral submission

- Provide an oral response (either voice clip or video) that is no more than 4 minutes in length.
- If you choose to make a video or voice clip for the assignment, do not feel the need to do tons of 'takes'. Rather, you can repeat yourself if you make a mistake, or feel you are unclear. This is not meant to be an additional burden, but rather to provide you with the opportunity to practice your oral communication skills
- You might be wondering how can I record this? One way to do this would be to schedule a Zoom meeting and record yourself in it. You can record the video to the cloud, or even directly on your computer! There will be many file types, including a video version, and one that is just a voice recording.
- Depending on the size of your file, you may be able to upload the file directly to Quercus in the submission box. If you run into issues, you can also upload the video onto a streaming platform (e.g. YouTube or MyMedia). This will provide you with a URL that you can then submit for your assignment. Paste the URL in your answers to this problem set.
- If you are looking for more ideas of how to record yourself for this assignment, or run into issues on how to upload your assignment, please post to Piazza

Vocabulary

- Type I and II error
- Comparing two population means/proportions
- One- and two-sample hypothesis test