Week 5 Breakout activity

Prof. Caetano and [your name and group number here]

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3
```

```
gratitude <- read_csv("gratitude.csv")
```

Questions 1 to 7 inclusive are about the same investigation. For these questions, recall that in the description of the data for this week we are told that, for the adjustment to university life score, **scores over 144 are considered to indicate successful adjustment while scores of 144 or less indicate difficulties with adjusting**.

Questions 8 to 11 are about a different investigation, using the `gratititude_v2.csv` dataset.

Question 1

Adjust the code below to create a new variable called `adjusted` that has the value TRUE if the student has a score greater than 144 and FALSE if the student has score of 144 or less. Then create an object called `group_props` using `summarise()` to find out what proportion of students have "successfully adjusted to university life" in each group.

```
gratitude_new <- gratitude %>%
  mutate(adjusted =
           case_when(adjust_college > 144 ~ TRUE,
                     adjust_college <= 144 ~ FALSE)) # this line needed to be completed

group_props <- gratitude_new %>%
  group_by(treatment) %>%
  summarise(prop_adj = mean(adjusted))

group_props
```

```
## # A tibble: 2 x 2
##   treatment prop_adj
## * <chr>        <dbl>
## 1 control      0.129
## 2 treatment    0.467
```

What percentage of students who were part of the gratitude journaling treatment group successfully adjusted to university life?

A. 13%

B. 30%

C. 47%

D. 100%

Question 2

Suppose one of your peers is very skeptical about "this whole gratitude nonsense". They think that there will be no real difference between the treatment and control groups with regard to the proportion who successfully adjust to university life.

Which of the following statements is TRUE?

A. Your peer's opinion is like the null hypothesis for a two sample hypothesis test.

B. Your peer's opinion is like the alternative hypothesis for a two sample hypothesis test.

C. Your peer's opinion is like the null hypothesis for a hypothesis test on one proportion.

D. Your peer's opinion is like the alternative hypothesis for a hypothesis test on one proportion.

Question 3

Suppose you want to test whether there is a difference between the proportions of students in the treatment and control groups that successfully adjust to university (i.e. have a score of more than 144 on the adjust to university life scale).

Which of the following would be the correct null hypothesis for this test?

A. $H_0 : \bar{x}_{treatment} - \bar{x}_{control} = 144$

B. $H_0 : p_{treatment} - p_{control} = 0$

C. $H_0 : p_{adjust.to.univeristy} = 0$

D. $H_0 : \mu_{treatment} - \mu_{control} = 0$

Question 4

Which of the following would be the BEST alternative hypothesis for this test?

A. The average adjustment to university score for first-year Turkish college students who practice gratitude journaling for three weeks is not the same as the average adjustment to university score for first-year Turkish college students who did not practice gratitude journaling for three weeks.

B. In our sample, the proportion of students who successfully adjust to university life in the treatment group is not the same as the proportion in the control group.

C. It is not the case that the proportion of first-year Turkish college students who successfully adjust to university life among those who practice gratitude journaling for three weeks is the same as the proportion of first-year Turkish college students who successfully adjust to university life among those who did not practice gratitude journaling for three weeks.

D. The proportion of first-year Turkish college students who practice gratitude journaling for three weeks is the same as the proportion of first-year Turkish college students who did not practice gratitude journaling for three weeks.

Question 5

Update the code below to calculate the test statistic for this investigation. NOTE: This code relies on the code in Question 1 being complete.

```
group_props
```

```
## # A tibble: 2 x 2
##   treatment prop_adj
## * <chr>        <dbl>
## 1 control      0.129
## 2 treatment    0.467
```

```
test_stat <- group_props %>%
  summarise(test_stat = diff(prop_adj)) # this line needed to be completed
test_stat
```

```
## # A tibble: 1 x 1
##   test_stat
##       <dbl>
## 1     0.338
```

```
test_stat_num <- as.numeric(test_stat$test_stat)
test_stat_num
```

```
## [1] 0.3376344
```

What is the test statistic for this investigation?

A. 13%

B. 0.338

C. 0.467

D. 0

Question 6

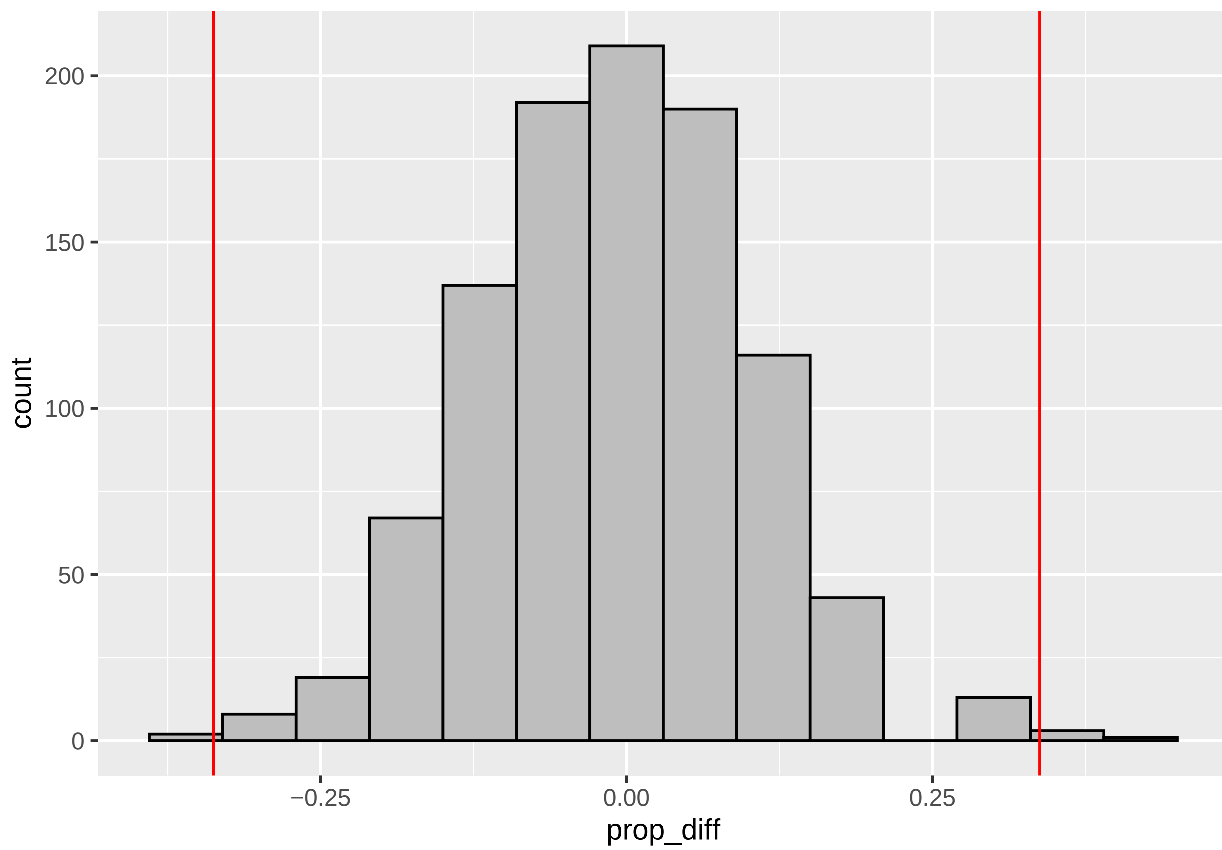
```r
set.seed(42)
repetitions <- 1000
simulated_values <- rep(NA, repetitions)
test_stat <- test_stat_num # this line uses the work from Q5

for(i in 1:repetitions){
  simdata <- gratitude_new %>%
    mutate(treatment = sample(treatment)) %>%
    group_by(treatment) %>%
    summarise(prop = mean(adjusted)) %>%
    summarise(value = diff(prop))

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

sim <- tibble(prop_diff = simulated_values)

sim %>% ggplot(aes(x=prop_diff)) +
  geom_histogram(binwidth = 0.06, color="black", fill="gray") +
  geom_vline(xintercept = abs(test_stat), colour = "red") +
  geom_vline(xintercept = -abs(test_stat), colour = "red")
```



Based on the plot alone, which of the following would you estimate the p-value to be closest to?

A. 0 (exactly)

B. 0.005

C. 0.10

D. 0.30

```
## Exact answer
sim %>%
  filter(abs(prop_diff) >= abs(test_stat_num)) %>%
  summarise(pval = n()/repetitions)
```

```
## # A tibble: 1 x 1
##    pval
##   <dbl>
## 1 0.006
```

Question 7

Based on the p-value from Question 6, what kind of error might we be making, assuming we are using an α of 0.01?

A. Type I

B. Type II

C. Selection bias

D. Coercion error

Question 8

An additional question was asked of each student before the experiment began. It was "What is your favourite number between 0 and 100?" and recorded next to their other measures in the dataset `gratitude_v2.csv`. We would now like to investigate *median* favourite numbers between the control and treatment groups.

Which ONE of the following statements about this new investigation is FALSE?

A. Assuming this is a well-conducted experiment, any differences between the median favourite numbers for the two groups should be due to chance acting alone.

B. As this variable was measured for each student before the experiment was conducted, any difference between the group medians is just due to the random allocation of students to treatment groups and not due to the treatments themselves.

C. Assuming this is a well-conducted experiment, differences between the median favourite numbers for the treatment and control groups should be due to one of two reasons: chance acting alone or chance and something else acting to produce a difference.

D. We would expect to get a large p-value at the end of this investigation.

Question 9

Choose the correct hypotheses to investigate if there is a difference in median favourite numbers between the treatment and control groups.

A. $H_0 : median_{treatment} = median_{control}$ **and** $H_1 : median_{treatment} \neq median_{control}$

B. $H_0 : \tilde{x}_{treatment} = \tilde{x}_{control}$ and $H_1 : \tilde{x}_{treatment} \neq \tilde{x}_{control}$

C. $H_0 : sample.median_{treatment} = sample.median_{control}$ and $H_1 : sample.median_{treatment} \neq sample.median_{control}$

D. $H_0 : median_{treatment} - median_{control} = 0$ and $H_1 : median_{treatment} - median_{control} > 0$

Question 10

Calculate the test statistic for this investigation into whether there is a difference in median favourite numbers between the treatment and control groups.

To do this, replace `sum` in the code below with the correct function.

```
gratitude_v2 <- read_csv("gratitude_v2.csv")

# Calculate your test statistic
test_stat_num <- gratitude_v2 %>%
  group_by(treatment) %>%
  summarise(medians = median(fav_num)) %>% # this line needed to be updated
  summarise(diff_med = diff(medians)) %>%
  as.numeric()

test_stat_num
```

```
## [1] -11
```

What is the test statistic for this investigation?

A. -97

B. -11

C. -1.46

D. 0.00

Question 11

You don't need to make any changes to the code below, but you may find it useful as example code for your problem set, etc. It does rely on your completion of the previous question.

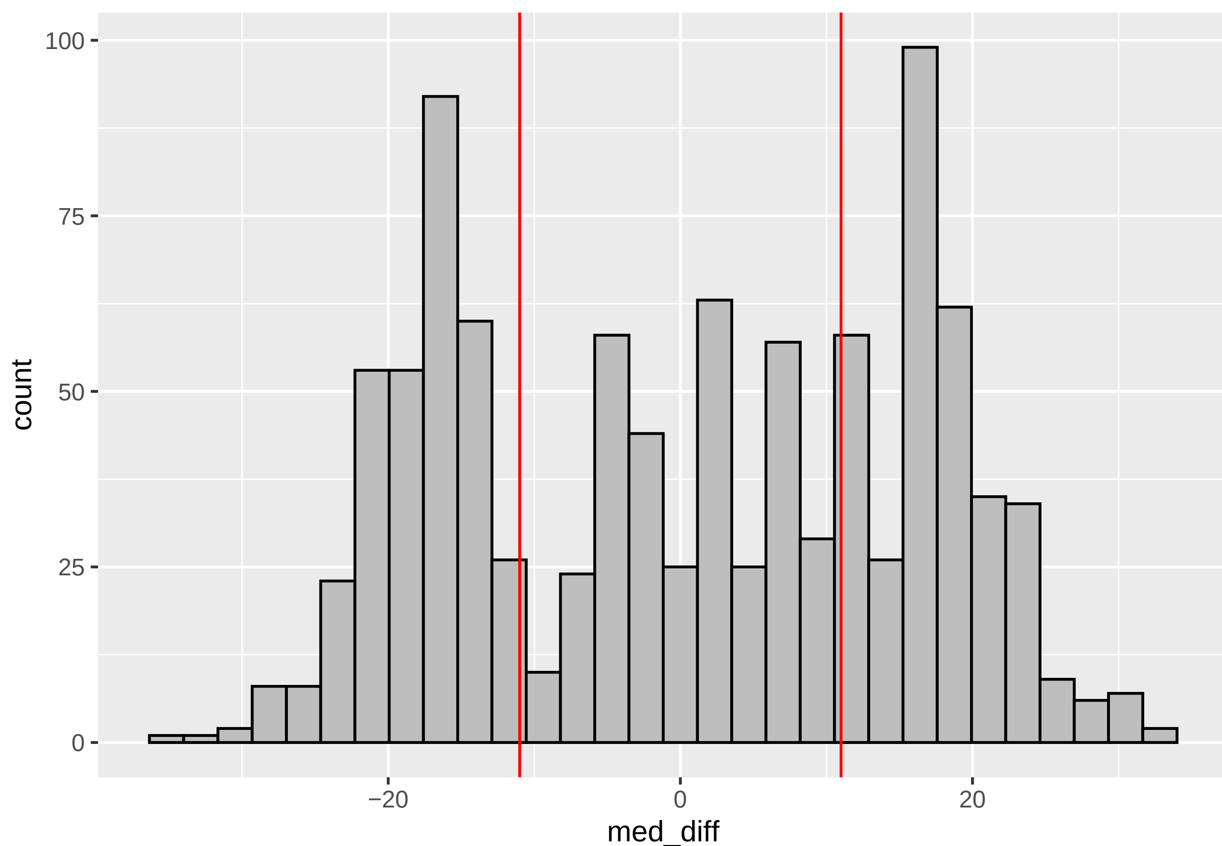
```
set.seed(123456789)
repetitions <- 1000
simulated_values <- rep(NA, repetitions)
test_stat <- test_stat_num # this line uses the work from the previous question
```

```r
for(i in 1:repetitions){
  simdata <- gratitude_v2 %>%
    mutate(treatment = sample(treatment)) %>%
    group_by(treatment) %>%
    summarise(medians = median(fav_num)) %>%
    summarise(value = diff(medians))

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

sim <- tibble(med_diff = simulated_values)

sim %>% ggplot(aes(x=med_diff)) +
  geom_histogram(color="black", fill="gray") +
  geom_vline(xintercept = abs(test_stat), colour = "red") +
  geom_vline(xintercept = -abs(test_stat), colour = "red")
```




```r
# Calculate p-value
num_more_extreme <- sim %>%
  filter(abs(med_diff) >= abs(test_stat_num)) %>%
  summarise(n())

p_value <- as.numeric(num_more_extreme / repetitions)
p_value
```

```
## [1] 0.665
```

Which of the following is the BEST interpretation of the p-value calculated above?

A. There is a 67% chance that there is no difference between the median favourite numbers of the two groups.

B. We have strength of evidence against the null hypothesis of $\alpha = 0.665$.

C. We have no evidence against the hypothesis that there is no difference in the median favourite numbers between the treatment and control groups.

D. At the 50% significance level we fail to reject the alternative hypothesis.