Problem Set6

Vivian Yeh

Question 1

```
library(stat20data)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(infer)
```

```
Attaching package: 'infer'

The following object is masked from 'package:stat20data':

    rep_sample_n
```

```
data(promote)
```

part a: Write the null and alternative hypothesis

Null hypothesis: Gender did not play a role in promotion decisions.The process of promotion is independent of gender.

Alternative hypothesis: Gender played a role in promotion decisions. The process of promotion is dependent of gender

part b: Compute the observed test statistic.

-0.292

```
test_stat <- promote %>%
  specify(response = decision,
          explanatory = gender,
          success = "promote")%>%
  calculate(stat = "diff in props")
```

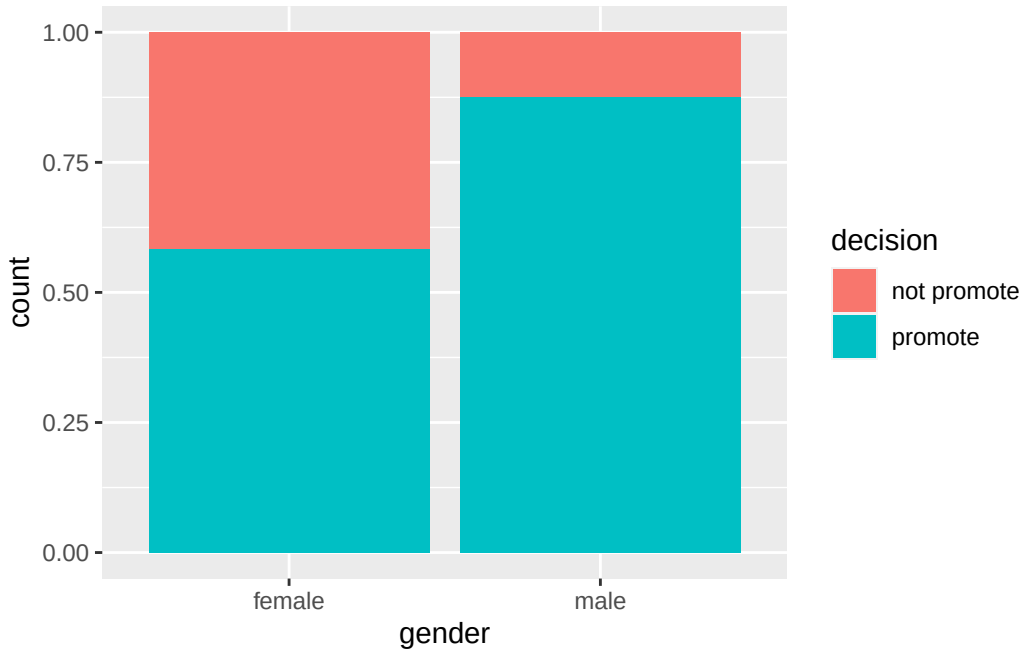
```
Warning: The statistic is based on a difference or ratio; by default, for
difference-based statistics, the explanatory variable is subtracted in the order
"female" - "male", or divided in the order "female" / "male" for ratio-based
statistics. To specify this order yourself, supply `order = c("female", "male")`
to the calculate() function.
```

```
test_stat
```

```
Response: decision (factor)
Explanatory: gender (factor)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.292
```

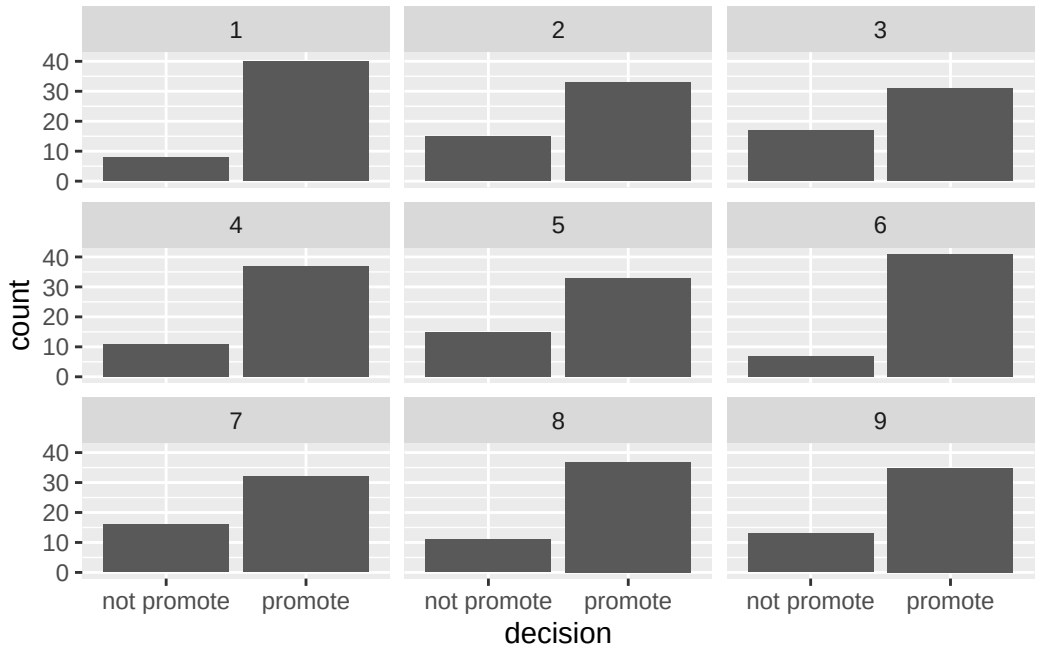
part c: Visualize the observed data using an appropriate plot.

```
library(ggplot2)
promote %>%
  ggplot(aes(x = gender,
             fill = decision)) +
  geom_bar(position = "fill")
```



part d: Construct a plot featuring 9 subplots, each one featuring a visualization of a data set generated under the null hypothesis. Does your visualization of the observed data from the previous part look like it could be one of these plots?

```
promote %>%
  specify(response = decision,
          success = "promote") %>%
  generate(reps = 9,
           type = "bootstrap") %>%
  ggplot(aes(x = decision)) +
  geom_bar() +
  facet_wrap(vars(replicate),
             nrow = 3)
```



part e: Construct and save the null distribution of statistics.

```
promote %>%
  specify(response = decision,
          explanatory = gender,
          success = "promote") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500,
           type = "permute")%>%
  calculate(stat = "diff in props")
```

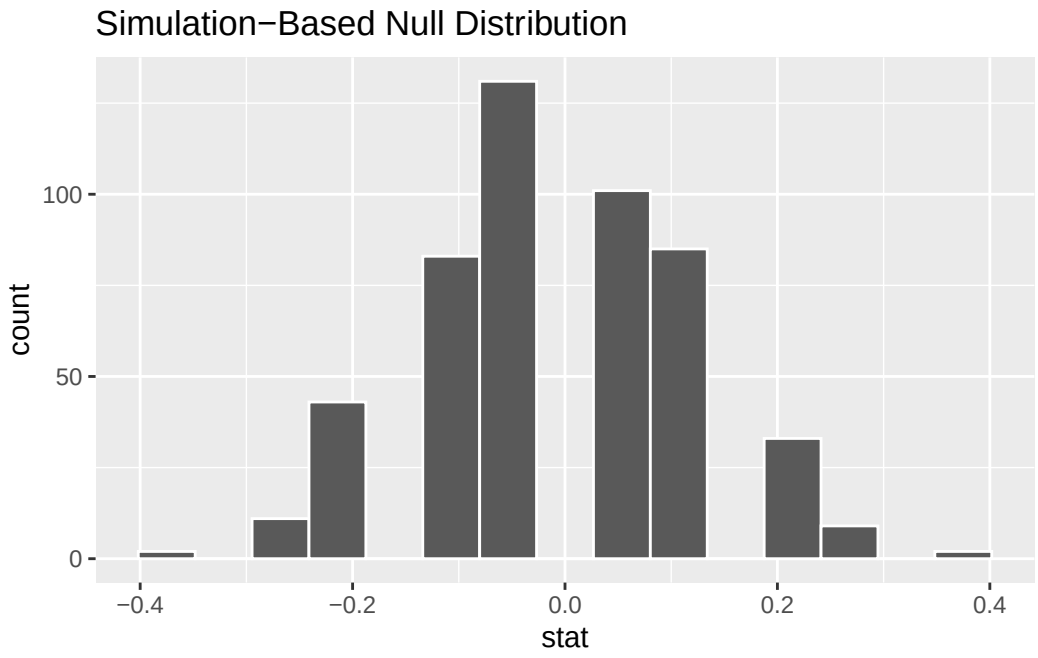
```
Warning: The statistic is based on a difference or ratio; by default, for
difference-based statistics, the explanatory variable is subtracted in the order
"female" - "male", or divided in the order "female" / "male" for ratio-based
statistics. To specify this order yourself, supply `order = c("female", "male")`
to the calculate() function.

Response: decision (factor)
Explanatory: gender (factor)
Null Hypothesis: independence
# A tibble: 500 x 2
   replicate    stat
```

```
      <int>   <dbl>
 1        1  0.125
 2        2 -0.208
 3        3 -0.125
 4        4  0.0417
 5        5 -0.125
 6        6 -0.0417
 7        7 -0.208
 8        8 -0.125
 9        9  0.0417
10       10  0.125
# ... with 490 more rows
```

part f: Visualize the null distribution.

```
promote %>%
  specify(response = decision,
          explanatory = gender,
          success = "promote") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500,
           type = "permute") %>%
  calculate(stat = "diff in props") %>%
  visualize()
```

```
Warning: The statistic is based on a difference or ratio; by default, for
difference-based statistics, the explanatory variable is subtracted in the order
"female" - "male", or divided in the order "female" / "male" for ratio-based
statistics. To specify this order yourself, supply `order = c("female", "male")`
to the calculate() function.
```



part g: Compute the p-value.

```
promote %>%
  specify(response = decision,
          explanatory = gender,
          success = "promote") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 500,
           type = "permute") %>%
  calculate(stat = "diff in props") %>%
  get_p_value(obs_stat = test_stat,
              direction = "both")
```

```
Warning: The statistic is based on a difference or ratio; by default, for
difference-based statistics, the explanatory variable is subtracted in the order
"female" - "male", or divided in the order "female" / "male" for ratio-based
statistics. To specify this order yourself, supply `order = c("female", "male")`
to the calculate() function.

# A tibble: 1 x 1
  p_value
    <dbl>
1    0.02
```

part h: Interpret your p-value based on the significance level we set at the beginning of the problem. What does it say about the consistency between the null hypothesis and the observed data?

The P-value, which is 0.068, shows that there is no statistical significance difference between two genders based on the alpha = 0.05. Therefore, we cannot reject the null hypothesis.

part i: What is the probability that we conclude that there is a difference in promotion decisions by gender when in truth there is no difference?

The probability is 6.8%, which is also the probability of the Type I Error.

part j: What would happen to the power of your hypothesis test?

The power of the hypothesis test would rise up because the sample size increases. In this case, we will have a more accurate test results.

Question 2

The 95% confidence level for this statistic is from 0.604 to 0.844, and it doesn't contain 0. This indicates that the promotion rate is different from each gender.

```
set.seed(805)

promote %>%
  specify(response = decision,
          success = "promote") %>%
  generate(reps = 500, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(.95)
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1    0.604    0.844
```

Question 3

For each of the following sets of hypotheses, determine whether committing a Type I error or a Type II error has greater real-world consequences. Explain your decision.

part a: Imagine that the San Francisco Giants hold a book donation promotion for one of their games where every individual who brings a used book to the stadium is allowed free entry.

0 : The number of individuals who wish attend the game will be no more than the capacity of the stadium.

: The number of individuals who wish to attend the game will exceed the capacity of the stadium.

In this case, the Type II error has greater real-world consequences because the excess number of individuals will have no access to enter the stadium, and the San Francisco Giants will be lack of both the human and capital resources and preparation to handle the excess amount of people.

part b:

0: A new operating system is no more effective for Carbon Health to use on the computers at its clinics than the old operating system.

: A new operating system is more effective for Carbon Health to use on the computers at its clinics than the old operating system.

In this situation, the Type I Error has greater real-world consequences because the clinic will keep using the bad-functioned operating system and miss the chance to improve the operating system.

part c:

0: The results of the 2020 United States presidential election were legitimate (the right candidate won).

: The results of the 2020 United States presidential election were fraudulent (the wrong candidate won).

Under this condition, the Type II Error has greater real-world consequences because the wrong candidate won is the true even if the test shows the right candidate won. The wrong candidate still win the election.

Question 4: stat20 website

Critique our course website, www.stat20.org/

- **The intended audience**: Students who take stat 20 class
- **Four principles:**

Adapt to your audience: Used. The contents are designed for the students who take stat 20, and the contents is written in the way that teach students to understand the concepts and applications of the statistic knowledge.

Maximize the signal-to-noise ratio: Used. The website make all formatting uniform throughout such as the font, text size, text color, webpage style. The reading notes focus on the important concepts of the class, and this shows the technique of minimizing the noise.

Use effective redundancy: Used. The webpage states the main points and the important phrase repetitively to help students notice the important point.

Trade-offs: Yes. The class goes in a limited time, so there are some other important concepts that cannot be covered in this class.

- **Strategies**:

Revise: Used. When the professor notice there's a mistake in the materials, notes, or assignments, they will revise it and notify students.

Explicitly then the details: Used. There are bullets points showing first, then explain each point in details.

Hierarchical is better than sequential: Used. The arrange of the notes followed this strategy. It talks about the unit first, then the detail topic; for example, The Taxonomy of Data comes before the detailed introduction of categorical and numerical data.

Make the structure easy to navigate and understand: Used. There is a floating table of contents at the top right for us to keep track of where we are reading.

Communicate at multiple levels: Not used. Also, it is difficult to apply in the class notes.

- **Improve how the content was communicated:** If the website could apply the technique ofCommunicate at multiple levels, students with different learning level can have a better learning outcome. Some students have never learned about statistic may need more basic concepts to start with.

Question 5: a data visualization

Critique – – Where bars out number grocery stores.

- **The intended audience**: Regular readers
- **Four principles**:

Adapt to your audience: Used. The contents are easy to follow with and not too complex to understand.

Maximize the signal-to-noise ratio: Used but can be improved. The website make all formatting uniform throughout such as the font, text size, text color, webpage style. There can be more bullet points instead of the long paragraph.

Use effective redundancy: Used. The webpage keeps stating the bars and grocery stores to make the audience keep on the right track.

Trade-offs: Yes. It mainly talks about the condition in the US, so there are less focus on other countries part.

- **Strategies:**

Revise: Not sure whether Used. Cannot see the edit date. However, the author stated, "I came back to the map recently," and this shows that the author had review the data.

Explicitly then the details: Used. The author started from three questions then answered each of them in detailed.

Hierarchical is better than sequential: Not Used. It shows the maps of different countries sequentially.

Make the structure easy to navigate and understand: Not Used. There is not a floating table of contents for us to keep track of where we are reading.

Communicate at multiple levels: Not used. However, the contents is easy for anyone to understand without the multiple levels communication.

- **Improve how the content was communicated:** I would suggest that the author can add some bullets points and bold keywords for audience to follow up.