Assignment 5

Mingrui Duan

Question 1: Enter the data from Table IV into a data frame. To do this, you should ensure that the data is in standard ggplot format, meaning one response per row, and a column for group. Note that this is not the format that the table is in, since for example there are 7 ratings of "2" in the heterosexual group, which should give 7 lines in your dataframe. DO NOT make your data frame in Excel or enter these 7 responses by hand. Instead, use the rep function in R and format your R code in a way that makes it easy to detect errors.

```
# Creating the "Group" Columns
group.homo <- rep("Homosexual", 30)
group.hete <- rep("Heterosexual", 30)
group.col <- c(group.homo, group.hete) # Bind these two groups together

# Creating the "Rating" Columns
rating.homo.2 <- rep(2, 9)
rating.homo.3 <- rep(3, 15)
rating.homo.4 <- rep(4, 6)
rating.homo <- c(rating.homo.2, rating.homo.3, rating.homo.4)

rating.hete.2 <- rep(2, 7)
rating.hete.3 <- rep(3, 19)
rating.hete.4 <- rep(4, 3)
rating.hete.5 <- rep(5, 1)
rating.hete <- c(rating.hete.2, rating.hete.3, rating.hete.4, rating.hete.5)

rating.col <- c(rating.homo, rating.hete)

# Creating a new dataframe
df <- data.frame(group.col)
df$rating <- rating.col
colnames(df)[c(1,2)] <- c("Group", "Rating")
```

Question 2: Print a summary of your data table and check against a friend's.

```
# Print the summary of new data
summary(df)
```

```
##     Group               Rating
##  Length:60          Min.   :2.000
```

```
##  Class :character   1st Qu.:2.000
##  Mode  :character   Median :3.000
##                     Mean   :2.917
##                     3rd Qu.:3.000
##                     Max.   :5.000
```

Question 3: Using your data and R's table function, replicate Hooker's Table IV. (Note R's table function may defaultly leave off the "1" ratings since there are none, and that's fine)

```
table(df)
```

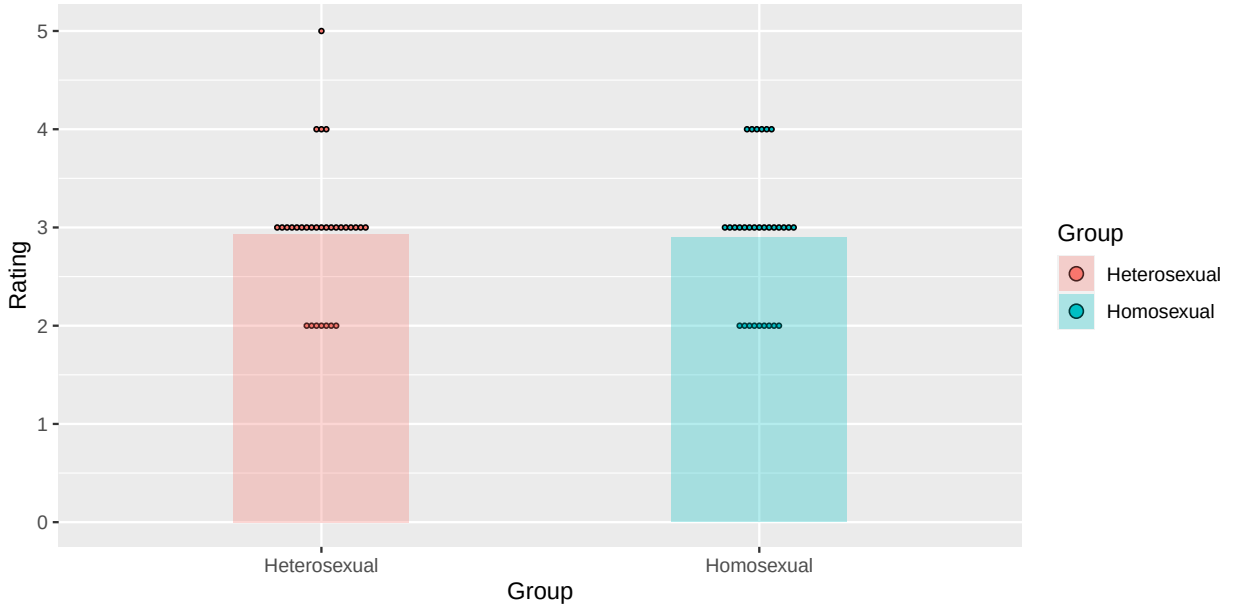
```
##               Rating
## Group           2  3  4  5
##   Heterosexual  7 19  3  1
##   Homosexual    9 15  6  0
```

Question 4: Make a "publication-quality" plot of the mean ratings in each group. For this plot, put bars for the means, and individual scatter points for each rating using geom_dotplot (you may want to read the help files). Fiddle with the parameters of geom_dotplot to make it look nice (you may want to set binwidth, stackdir, alpha, and colors). Save your graph as a pdf in an aspect ratio that makes it look nice.

```
library(ggplot2)
ggplot(df, aes(x = Group, y = Rating, fill = Group)) +
        geom_dotplot(binaxis =  "y",
                     binwidth = 0.05,
                     stackdir = "center") +
        geom_bar(stat = "summary",
                 fun.y = "mean",
                 alpha = 0.3,
                 width = 0.4)
```

```
## Warning in geom_bar(stat = "summary", fun.y = "mean", alpha = 0.3, width =
## 0.4): Ignoring unknown parameters: `fun.y`
```

```
## No summary function supplied, defaulting to `mean_se()`
```



Question 5: One reason to like a permutation test is that it leave the overall statistics of responses (mean, distribution, etc.) the same. Explain why. Does it leave the distribution of responses within each group the same or not? Explain why. Does it matter if we form the null distribution by shuffling the group labels or shuffling the responses? Explain why or why not.

1. The reason that "permutation test" keep the overall statistics is that what "permutation test" do is just change how our dataset are "grouped". It doesn't change each data or the size of the overall dataset. So for statistics like "mean" or "distribution", they won't change.
2. It doesn't leave the distribution of responses within each group the same. Because when conducting "permutation test" you actually change the original distribution of group by changing the data's label.
3. It matters. Shuffling labels tests the null hypothesis of no difference between groups, while shuffling values estimates the sampling distribution of a statistic. Shuffling labels is often used in hypothesis testing to evaluate whether two groups differ from each other, while shuffling values is often used in resampling methods to estimate the variability of a statistic and to construct confidence intervals.

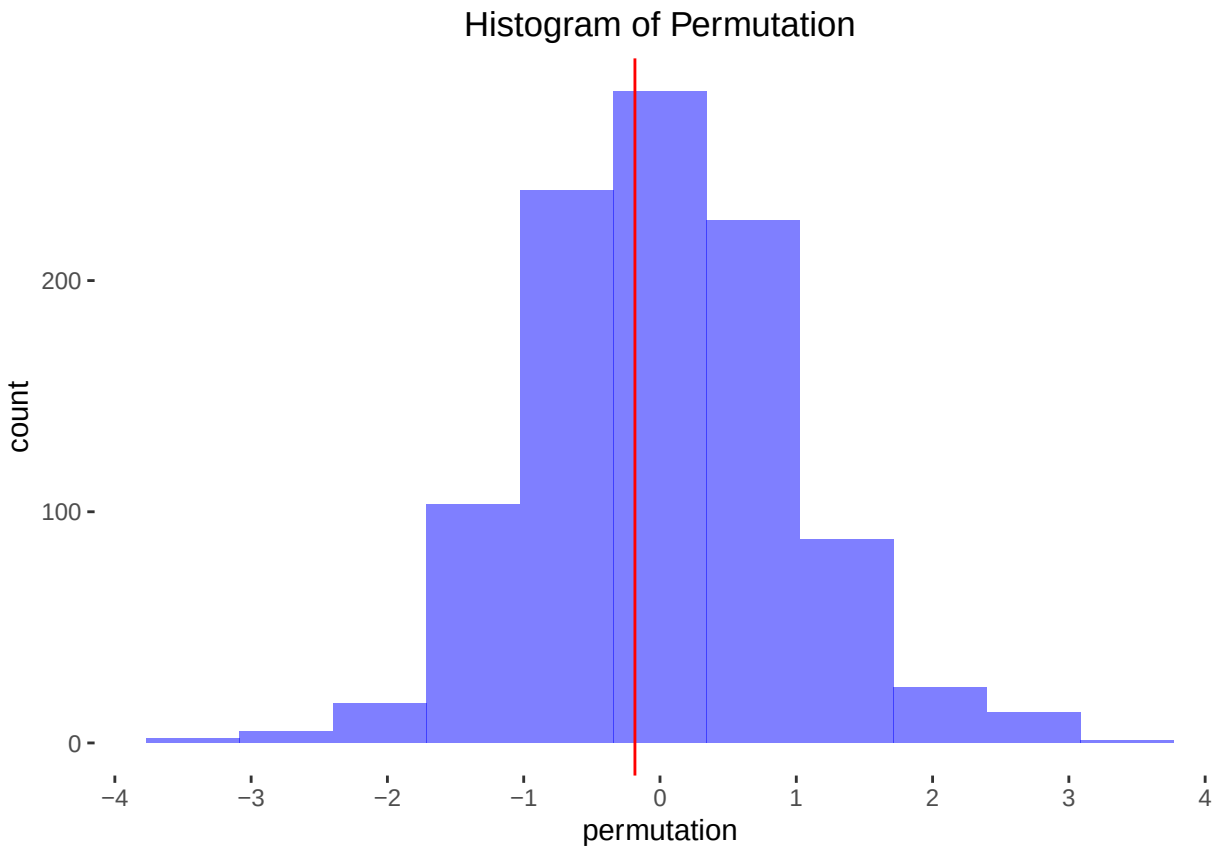
Question 6: Write 3-4 sentences explaining at an undergraduate level why shuffling labels corresponds to a null hypothesis of "no difference" between groups. Think about the groups would look like if there was no difference, and what shuffling does.

If the dataset follows our null hypothesis that there is no difference between groups, then it means that no matter how we divide the data, we will get the same conclusion after compare these groups. So what shuffling labels do is exactly re-group our data, and help us to know whether how the data are "grouped" matters.

Question 7: Write a loop that permutes the group labels, computes the test statistics on the permuted samples, and stores them in a vector called permuted.stats. Store 1000 of these permutations. Plot a histogram of these permuted differences with ggplot; if the default binning of the histogram doesn't look great, read about how to change it and change it to something reasonable. Place a vertical red line at the true (un-permuted) test statistic.

```
# Copy the functions
sem <- function (x) {
  sd(x)/(sqrt(length(x)))
}

test.statistic <- function (x, y) {
  (mean(x)-mean(y))/sqrt(sem(x)^2 + sem(y)^2)
}

# Get each group's sub dataframe
homo <- subset(df, Group == "Homosexual", select = Rating)
hete <- subset(df, Group == "Heterosexual", select = Rating)
# Create the vector store all the permuted statistics
permuted.stats <- test.statistic(homo$Rating,
                                 hete$Rating)
true.test.statistic <- test.statistic(homo$Rating,
                                 hete$Rating)
# Set the Loop and store these permutations
for (i in 1:999) {
  # shuffle the labels
  df <- data.frame("Group" = sample(df$Group), "Rating" = df$Rating)

  homo <- subset(df, Group == "Homosexual", select = Rating)
  hete <- subset(df, Group == "Heterosexual", select = Rating)

  # Add the new permuted stat to our vector
  permuted.stats <- append(permuted.stats,
                           test.statistic(homo$Rating, hete$Rating))
}

# Plot the Histogram of the test statistics
library(ggplot2)
ggplot(NULL, aes(x = permuted.stats)) +
        geom_histogram(bins = 11,
                       fill = "blue",
                       alpha = 0.5) +
        geom_vline(xintercept = true.test.statistic,
                   linetype = "solid",
                   color = "red") +
        xlab("permutation") +
        scale_x_continuous(n.breaks = 10) +
        ggtitle("Histogram of Permutation", ) +
        theme(panel.background = element_rect(fill = "transparent"),
              plot.background = element_rect(fill = "transparent", color = NA),
              plot.title = element_text(hjust = 0.5),
```

```
        panel.grid.major = element_blank(), #remove major gridlines
        panel.grid.minor = element_blank()) #remove minor gridlines
```



To determine our best bin for the histogram, I use the "Sturges' rule". Use $log_2(n) + 1$ as the number of bins. Here n is our data points size: 1000. So the best bin after computation is 11.

Question 8: Why does the count on the histogram go so high, when there are so few subjects? What determines the count and is the count relevant to the inferences we'll make, why or why not?

The count on the histogram goes so high because histograms are used to represent the distribution of a dataset by visualization. If there are only a few data points in the dataset, each data point will be represented by a single bar in the histogram. As a result, the height of each bar will be relatively high compared to histograms of larger datasets, where the height of each bar is typically much lower. Also, the dataset can be effected by the outliers. The count is relevant to the inferences we make. Because what we do our null hypothesis is that there is no difference between these two groups. So if the count of the permutation value near 0 is high, it can show that our null hypothesis is correct.

Question 9: What should the mean of the permuted test statistics be? Explain why. Does yours match?

```r
print(mean(permuted.stats))
```

```
## [1] -0.001823388
```

The mean of the permuted test statistics should be close to zero, for the reason that our null hypothesis is that there is no difference between two groups. Mine is match, the mean of the permuted test is 0.073

Question 10: From permuted.stats, print a one-tailed p-value, testing the hypothesis that the homosexual group has higher ratings than the heterosexual.

```r
p.set <- subset(permuted.stats, permuted.stats < true.test.statistic)
p.value <- (length(p.set)/1000)/2
print(p.value)
```

```
## [1] 0.198
```

Since my p-value is greater than 0.05, I don't have enough evidence to conclude that my hypothesis.

Question 11: From permuted.stats, print a two-tailed p-value, testing the hypothesis that the groups differ.

```r
p.set <- subset(permuted.stats,
                permuted.stats < true.test.statistic | permuted.stats > abs(true.test.statistic))
p.value <- (length(p.set)/1000)/2
print(p.value)
```

```
## [1] 0.3905
```

Question 12: Explain what the p-value for Q11 means at the level a college undergraduate could understand.

Since my p-value is 0.3835, it illustrates that there is 38.35% chance of observing a difference in rating as extreme or more extreme than -0.184 points assuming the null hypothesis that the groups have the same ratings is true, and that the homosexual group has higher ratings than the heterosexual group.

Question 13: You've probably heard that $p < 0.05$ is a threshold for getting published. Your pvalues should not have been < 0.05 in the previous questions. Should this work have been published? Write a few sentences as you might in a review or discussion, explaining why or why not.

This work should have been published. Because the p-value I got is greater than the 0.05, so that the null hypothesis we make in the previous question is not sufficient to prove. But it doesn't mean my study cannot be publishable. Because in Evelyn Hooker's paper, it was 1950s. The study contribute to these particular topic in research on the adjustment of homosexual and heterosexual. Even if it doesn't get enough evidence, it is still innovative and important to the area at that time. So it's worth to be published.