HW2 (Individual)

Due Date: Oct 4th (Monday), 23:00 SG

You can choose one of the following submission modes.

(A): submit 2 files: (1) an R Notebook, make sure the code is executable (a template is provided for you); (2) the pdf file generated from your R Notebook.

(B): use whatever programming language that you are comfortable with to solve these problems. Submit 1 pdf document with questions and answers and put the code in the appendix.

(A) is preferred but not required

If you choose (A), then you have to download R (https://www.r-project.org/) and RStudio (https://rstudio.com/products/rstudio/download/). There are plenty of resources about R online, this video teaches some basics about R and Rstudio (https://www.youtube.com/watch?v=_V8eKsto3Ug). This article is about creating Rnotebook (http://uc-r.github.io/r_notebook). If you need more help on starting with R, email TAs.

HW2 worths 9 points in the final assessment.

Problem 1. (Short Answer Questions, 11')

Supposed that you have a random sample of size N from the population of interest.Throughout the problem, assume that the Stable Unit Treatment Value Assumption (SUTVA) holds (Check the slides of Lecture 4 or Rubin's book Chapter 1 if you are not familiar with that). By default, 0 stands for the control group and 1 stands for the treatment group.

(a)-(1') Explain the notation $Y_i(1)$.

Answer: $Y_i(1)$ refers to the potential outcome for subject i with treatment.

(b)-(1') Contrast the meaning of $Y_i(0)$ with the meaning of Y_i^{obs}.

Answer: $Y_i(0)$ refers to the potential outcome for subject i without treatment. Whereas, Y_i^{obs} refers to the realized and observed outcome of subject i

(c)-(2') Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i(1)$. Is it ever possible to observe both at the same time? Why?

Answer: $Y_i(0)$ refers to the potential outcome for subject i without treatment. Whereas, $Y_i(1)$ refers to the potential outcome for subject i with treatment. It is not possible to observe both at the same time. This is because ex post, the other potential outcomes cannot be observed because the corresponding actions that would led to them being realized were not taken.

(d)-(1') Explain the notation $\mathbb{E}[Y_i(0)|D_i = 1]$, where D_i is a binary variable that gives the treatment status for subject i, 1 if treated, 0 if control.

Answer: $\mathbb{E}[Y_i(0)|D_i = 1]$ refers to the conditional expectation of the potential outcome $Y_i(0)$ given that subject i received treatment. This is a counter-factual as it is impossible to observe this.

(e)-(2') Contrast the meaning of $\mathbb{E}[Y_i(1)]$ with the meaning of $\mathbb{E}[Y_i^{obs}|D_i = 1]$.

Answer: $\mathbb{E}[Y_i(1)]$ refers to the expectation of the potential outcome $Y_i(1)$. Whereas, $\mathbb{E}[Y_i^{obs}|D_i = 1]$ refers to the actual realized outcome of subject i given that subject i received treatment.

(f)-(2') Contrast the meaning of $\mathbb{E}[Y_i(1)|D_i = 0]$ with the meaning of $\mathbb{E}[Y_i(0)|D_i = 0]$. Which one is counterfactual?

Answer: $\mathbb{E}[Y_i(1)|D_i = 0]$ refers to the conditional expectation of the potential outcome $Y_i(1)$ given that subject i did not received treatment. $\mathbb{E}[Y_i(0)|D_i = 0]$ refers to the conditional expectation of the potential outcome $Y_i(0)$ given that subject i did not received treatment. $\mathbb{E}[Y_i(1)|D_i = 0]$ is the counter-factual.

(g)-(2') Assuming that you have N (finite) subjects in your experiment. For each subject, you only observed one of the subject's potential outcomes. Which of the following null hypothesis allow(s) you to derive the exact $p-$value just like the Fisher's Sharp Null Hypothesis? (check all that apply and explain why)

(A) $H_0: \ Y_i(1)/Y_i(0) = 3.14$ for all $i = 1, 2, ..., N$;

(B) H_0 : the median of $\{Y_i(0)\}_{i=1}^N$ and the median of $\{Y_i(1)\}_{i=1}^N$ are equal;

(C) $H_0: \ Y_i(1) = [4 - Y_i(0)]^2$ for all $i = 1, 2, ..., N$;

(D) $H_0: \ Y_i(1) = 3 \ln Y_i(0)$ for all $i = 1, 2, ..., N$;

(E) $H_0: \ \frac{1}{N}\sum_{i=1}^N Y_i(1) = \frac{1}{N}\sum_{i=1}^N Y_i(0)$.

Answer: To derive the exact $p-$value, we would need a hypothesis which would allow us to fill in all the missing data of the tables without ambiguity. This means we have to be able to derive $Y_i(1)$ from $Y_i(0)$ and vice versa.

(A) ✓. $H_0: \ Y_i(1)/Y_i(0) = 3.14$ means that under H_0, $Y_i(1) = 3.14 \times Y_i(0)$ and $Y_i(0) = Y_i(1)/3.14$
(B) X
(C) ✓. $H_0: \ Y_i(1) = [4 - Y_i(0)]^2$ means that under H_0, $Y_i(1) = [4 - Y_i(0)]^2$ and $Y_i(0) = 4 - \sqrt{Y_i(1)}$
(D) ✓. $H_0: \ Y_i(1) = 3 \ln Y_i(0)$ means that under H_0, $Y_i(1) = 3 \ln Y_i(0)$ and $Y_i(0) = e^{Y_i(1)/3}$
(E) X

Problem 2 (13').

You have developed 2 versions of a website, denoted by 0 and 1. You would like to know which version of the website can keep visitors stay longer. You did a pilot study with $N = 24$ subjects. You implemented a completely randomized experiment. Hence, half of the subjects (12) were assigned to the control group (0) and the rest were assigned to the treatment group (1).

(a)-(0.5') Load and print the data (there are 2 columns in the data)

```
data <- read.csv('./Data_2021.csv')
head(data)
```

```
##    observed Treatment
## 1 10.298260         1
## 2  9.183643         0
## 3  8.164371         0
## 4 10.595281         0
## 5  9.376818         1
## 6  4.116990         1
```

Fisher Sharp Null Hypothesis (or Exact Null Hypothesis)

Read Rubin's book Chapter 5, section 1 to 4.

Consider the following Null Hypothesis: the version of the website in the treatment group leads to 3 more seconds in sojourn time for all the subjects than the version of the website in the control group, i.e.,

$H_0 : Y_i(1) = Y_i(0) + 3, \ \forall \ i$

(b)-(0.5') Based on this null hypothesis H_0, what should be the table that describes the potential outcomes for all subjects in this experiment? Print the table (there should be 4 columns, one displays the observed outcome, one displays the actual treatment. one displays $Y_i(1)$ under H_0 and one displays $Y_i(0)$ under H_0)

```
index <- data$Treatment == 0
data[index, 'Yi(0)'] <- data[index, 'observed']
data[index, 'Yi(1)'] <- data[index, 'observed'] + 3

data[!index, 'Yi(0)'] <- data[!index, 'observed'] - 3
data[!index, 'Yi(1)'] <- data[!index, 'observed']

head(data)
```

```
##    observed Treatment     Yi(0)     Yi(1)
## 1 10.298260         1  7.298260 10.298260
## 2  9.183643         0  9.183643 12.183643
## 3  8.164371         0  8.164371 11.164371
## 4 10.595281         0 10.595281 13.595281
## 5  9.376818         1  6.376818  9.376818
## 6  4.116990         1  1.116990  4.116990
```

We choose the following test statistic

$T = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$.

(c)-(2') Generate the EXACT Probability Distribution of this test statistic T, under H_0 in this completely randomized experiment. Report the total number of possible realizations of the randomized assignment, and 25%, 50%, 75% quantile, mean, and variance of this distribution (Hint: you may want to consider the library `CombMSC`. Note that this code may take some time to run, as the total number of possible realizations is relatively large. You can see that with even just 24 subjects, a completely randomized experiment is able to generate a large number of possible assignment outcomes.)

```
n_combinations <- choose(nrow(data), nrow(data)/2)
print(paste0('There are ', n_combinations, ' combinations.'))
```

```
## [1] "There are 2704156 combinations."
```

```
library(CombMSC)
library(dplyr)
df <- data.frame()

Yi0 <- subsets(24, 12, data$`Yi(0)`) %>%
  rowMeans()
Yi1 <- subsets(24, 12, data$`Yi(1)`) %>%
  rowMeans() %>%
  rev()
t_stat <- Yi1 - Yi0

df[(1:n_combinations), c('Yi(0)_mean', 'Yi(1)_mean', 't_stat')] <- c(Yi0, Yi1, t_stat)

print('Test statistics quantile:')
```

```
## [1] "Test statistics quantile:"
```

```
quantile(df$t_stat, c(0.25, 0.5, 0.75))
```

```
##      25%      50%      75%
## 2.262575 3.000000 3.737425
```

```
print(paste0('mean: ', mean(df$t_stat)))
```

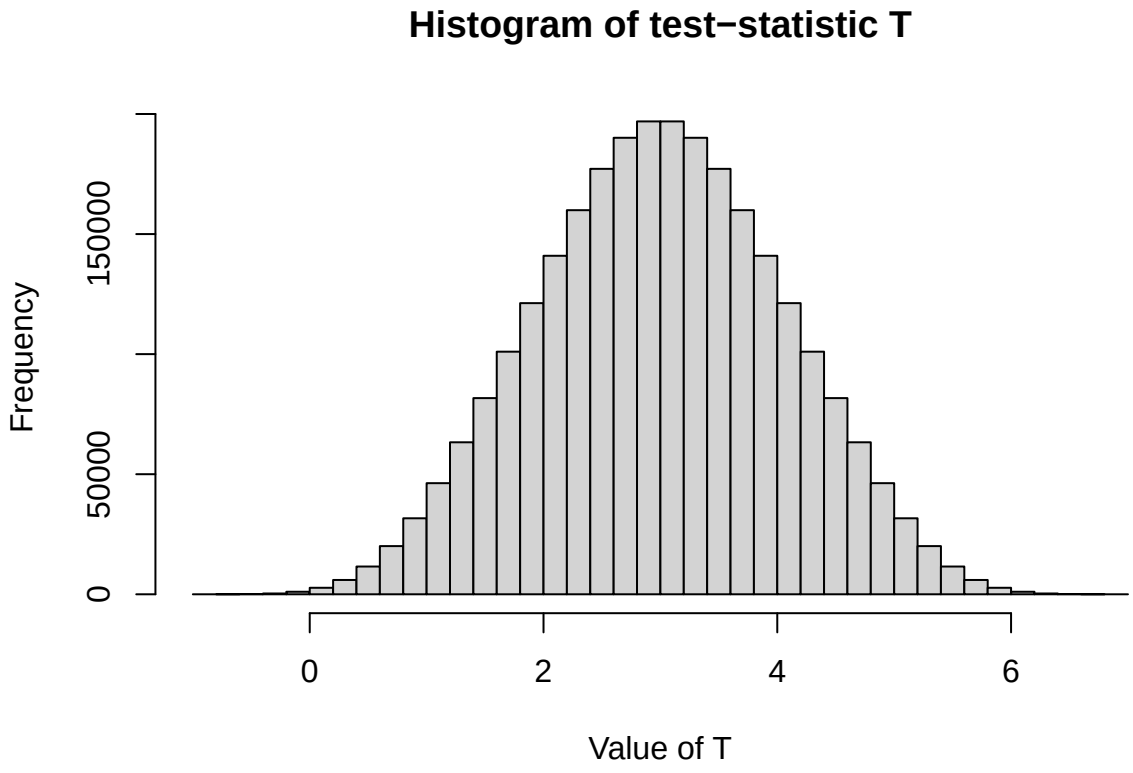
```
## [1] "mean: 3"
```

```
print(paste0('variance: ', var(df$t_stat)))
```

```
## [1] "variance: 1.11079030679931"
```

(d)-(1') Plot the histogram of this test statistic T with `breaks = 50` (a parameter in the `hist` command).

```
hist(df$t_stat,
     breaks = 50,
     main = 'Histogram of test-statistic T',
     xlab = 'Value of T')
```



(e)-(2') In the original data, what is the value of this test statistic T? If it is positive, then what is the probability of observing T being no less than that value under H_0 based on the distribution of T you have derived? If it is negative, then What is the probability of observing T being no greater than that value? What do you think about your H_0?

```
Yi0_mean <- data[data$Treatment == 0, 'observed'] %>%
  mean()
Yi1_mean <- data[data$Treatment == 1, 'observed'] %>%
  mean()
obs_t <- Yi1_mean - Yi0_mean
```

```
print(paste0('The observed test statistic T = ', round(obs_t, 4)))
```

```
## [1] "The observed test statistic T = -0.2858"
```

```
p_leq_t <- (df$t_stat <= obs_t) %>%
  sum()
p_leq_t <- p_leq_t / length(df$t_stat)
print(
  paste0('The probability of observing T being no greater than that value is: ',
         signif(p_leq_t, 4))
)
```

```
## [1] "The probability of observing T being no greater than that value is: 0.00008616"
```

As the probability of observing T being no greater is very unlikely under the assumption of the null hypothesis, the observation is statistically significant ($p \leq 0.05$) and therefore, we reject the null hypothesis H_0.

Read Rubin's book Chapter, section 1 to 6. Compute the estimate of the Neyman variance estimator. Note that the test statistic is still $T = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$. The true variance of T is $S_c^2/N_c + S_t^2/N_t - S_{ct}^2/N$, where the subscript "c" is control and "t" is treatment. The Neyman variance estimator ignores S_{ct}^2/N.

(f)-(1') Print the estimate of the Neyman variance estimator, assuming heterogenous variance (i.e., the variability of the $Y(1)$ and $Y(0)$ are not the same).

```
treated <- which(data$Treatment == 1)
control <- which(data$Treatment == 0)
Nt <- length(treated)
Nc <- length(control)

SCsquare <- 1 / (Nc - 1) *
  sum((data[control, 'Yi(0)'] -
         mean(data[control, 'observed'])) ^ 2)
STsquare <- 1 / (Nt - 1) *
  sum((data[treated, 'Yi(1)'] -
         mean(data[treated, 'observed'])) ^ 2)

Neyman_hetero <- (SCsquare / Nc) + (STsquare / Nt)
print(
  paste0(
    'The estimate of the Neyman variance estimator, assuming heterogenous variance is: ',
         round(Neyman_hetero, 4)
  )
)
```

```
## [1] "The estimate of the Neyman variance estimator, assuming heterogenous variance is: 0.6705"
```

(g)-(1') Using the estimate of the Neyman variance estimator, construct a 95% confidence interval, assuming normality. Based on this confidence interval you have derived, if someone claim that the average sojourn time in the treatment group is 3 seconds longer than that in the control group, then what is your opinion?

```
alpha <- 0.05
lb <- obs_t + qnorm((alpha / 2)) * sqrt(Neyman_hetero)
ub <- obs_t + qnorm(1 - (alpha / 2)) * sqrt(Neyman_hetero)

print(paste0('95% CI: (', round(lb, 4), ' , ', round(ub, 4), ')'))
```

```
## [1] "95% CI: (-1.8908 , 1.3191)"
```

Based on this confidence interval, I would reach the conclusion that the claim is false as 3 seconds is outside the 95% confidence interval.

Now we consider a different test statistic $G = \max(Y_1^{obs}) - \min\{Y_0^{obs}\}$.

(h)-(2') Generate the EXACT Probability Distribution of this test statistic G, under H_0 in a completely randomized experiment. Report the 25%, 50%, 75% quantile of this distribution, and variance

```
df <- data.frame()

Yi0 <- subsets(24, 12, data$`Yi(0)`) %>% data.frame() %>% apply(1, FUN = min)
Yi1 <- subsets(24, 12, data$`Yi(1)`) %>% data.frame() %>% apply(1, FUN = max) %>% rev()
g_stat <- Yi1 - Yi0

df[(1:n_combinations), c('Yi(0)_min', 'Yi(1)_max', 'g_stat')] <- c(Yi0, Yi1, g_stat)

print('Test statistics quantile:')
```

```
## [1] "Test statistics quantile:"
```

```
quantile(df$g_stat, c(0.25, 0.5, 0.75))
```

```
##      25%      50%      75%
## 11.42722 12.10352 12.47829
```

```
print(paste0('mean: ', mean(df$g_stat) %>% round(4)))
```

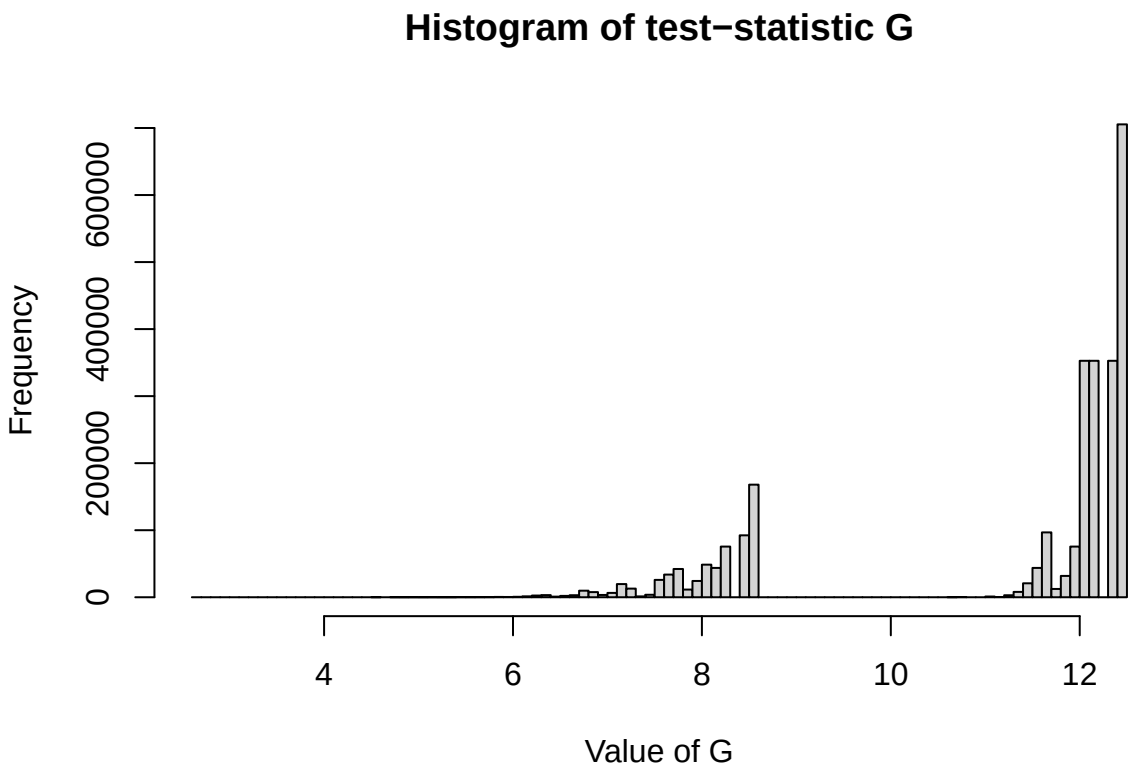
```
## [1] "mean: 11.2117"
```

```
print(paste0('variance: ', var(df$g_stat) %>% round(4)))
```

```
## [1] "variance: 3.2875"
```

(i)-(1') Plot the histogram of this test statistic G with `breaks = 100` (a parameter in the `hist` command).

```
hist(df$g_stat,
     breaks = 100,
     main = 'Histogram of test-statistic G',
     xlab = 'Value of G')
```



Note that when N (the number of subjects) is large, deriving the EXACT distribution of the statistic under Fisher Sharp NUll Hypothesis becomes very tedious. In this case, we often use "bootstrap", i.e., we will re-execute the randomized assignment procedure for M times. Then, within each iteration, we randomly assign half of the subjects to the control group and the remaining to the treatment group, and we compute the value of our test statistic.

(j)-(2') Still use T as the test statistic, set $M = 100000$. Report the 25%, 50%, 75% quantile of this distribution, and the variance of this distribution. Plot the histogram of this test statistic T with `breaks = 50` (a parameter in the `hist` command)

```
M <- 100000
df <- data.frame()
output <- c()

for (row in 1:M){
  index <- sample((1:24), 12)
  Yi0 <- data$`Yi(0)`[index] %>%
    mean()
  Yi1 <- data$`Yi(1)`[(25 - index)] %>%
    mean()
  t_stat <- Yi1 - Yi0
  output <- c(output, t_stat)
}

print('Test statistics quantile:')
```

```
## [1] "Test statistics quantile:"
```

```r
quantile(output, c(0.25, 0.5, 0.75))
```

```
##      25%      50%      75%
## 2.456497 3.001488 3.545843
```

```r
print(paste0('mean: ', mean(output) %>% round(4)))
```

```
## [1] "mean: 3.0013"
```

```r
print(paste0('variance: ', var(output) %>% round(4)))
```

```
## [1] "variance: 0.6128"
```

```r
hist(output,
     breaks = 50,
     main = 'Histogram of test-statistic T',
     xlab = 'Value of T')
```