Stats 102A - Homework 6 - Output File

Charles Liu (304804942)

Academic Integrity Statement

By including this statement, I, **Charles Liu**, declare that all of the work in this assignment is my own original work. At no time did I look at the code of other students nor did I search for code solutions online. I understand that plagiarism on any single part of this assignment will result in a 0 for the entire assignment and that I will be referred to the dean of students.

I did discuss ideas related to the homework with **Amanda Xu, Diana Pham, and Carolyn Moor** for parts 2 through 6. We merely compared end results without showing any code. At no point did I show another student my code, nor did I look at another student's code.

Loading the Necessary Packages:

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(boot)
```

Part 1

Explain what a p-value is and how a researcher should use it. (150 words or less)

The p-value is the probability for any statistical model that its data summary would be equal to or extremely greater than its observed value. A common misconception is that p-values measure the probability that the studied hypothesis is true or accept the alternative. It is, in fact, a statement about the data of the hypothetical explanation, This does not mean the p-value explains the statements at all. Researchers should use everything in the statistical summary, including the hypotheses of the study, the data collected and its decisions made, the statistical analysis conducted, the p-values computed, and the methods used. Researchers should know that a p-value without context, hypothesis, or any other evidence provides little to no information. A researcher should use p-values to help explain your statistical model, but it should not be the answer to your hypothesis.

Part 2

Randomization test for numeric data

Randomization test

Conduct a randomization test

```
# Data credit: David C. Howell

# `no_waiting` is a vector that records the time it took a driver to leave the
# parking spot if no one was waiting for the driver
no_waiting <- c(36.30, 42.07, 39.97, 39.33, 33.76, 33.91, 39.65, 84.92, 40.70, 39.65,
39.48, 35.38, 75.07, 36.46, 38.73, 33.88, 34.39, 60.52, 53.63, 50.62)

# `waiting` is a vector that records the time it takes a driver to leave if
# someone was waiting on the driver
waiting <- c(49.48, 43.30, 85.97, 46.92, 49.18, 79.30, 47.35, 46.52, 59.68, 42.89,
49.29, 68.69, 41.61, 46.81, 43.75, 46.55, 42.33, 71.48, 78.95, 42.06)

# Checking the averages of these two data
mean(waiting)
```

```
## [1] 54.1055
```

```
mean(no_waiting)
```

```
## [1] 44.421
```

```
# Creating our obs_dif2
obs_dif2 <- mean(waiting) - mean(no_waiting)

# Hypothesis Test:
# H0: mean(waiting) = mean(no_waiting)
# Ha: mean(waiting) > mean(no_waiting)
parking <- c(no_waiting, waiting)
set.seed(1)
differences2 <- rep(NA, 10000)
for(i in seq_along(differences2)) {
  randomized <- sample(parking)
  groupA <- randomized[1:20]
  groupB <- randomized[21:40]
  differences2[i] <- mean(groupA) - mean(groupB)
}

# The empirical p-value result using one-sided alternative
mean((differences2) >= obs_dif2)
```

```
## [1] 0.0193
```

We would REJECT the NULL Hypothesis, and this means that there is a significant difference between the means of waiting and no waiting.

Comparison to traditional t-test

Conduct a traditional two-sample independent t-test.

```r
# Ha: mean(no_waiting) < mean(waiting)
# t-test (no_waiting 1st on t.test)
t_test_2.1 <- t.test(no_waiting, waiting, alternative = "less")

## Same difference

# Ha: mean(waiting) > mean(no_waiting)
# t-test (waiting 1st on t.test)
t_test_2.2 <- t.test(waiting, no_waiting, alternative = "greater")

# Randomization Test
random_test_2 <- mean((differences2) >= obs_dif2)

t_test_2.1$p.value
```

```
## [1] 0.01900965
```

```r
t_test_2.2$p.value
```

```
## [1] 0.01900965
```

```r
random_test_2
```

```
## [1] 0.0193
```

```r
random_test_2 - t_test_2.1$p.value
```

```
## [1] 0.0002903549
```

We see there is a 0.0002903549 value difference between our Randomized Test's p-value and Traditional T-test's p-value. They both yield the same conclusion, which is REJECT the NULL Hypothesis. This difference is far too small to make any difference, and we can say either methods work.

Part 3

Another Randomization test for numeric data

Exploratory Analysis

Carry out an exploratory analysis.

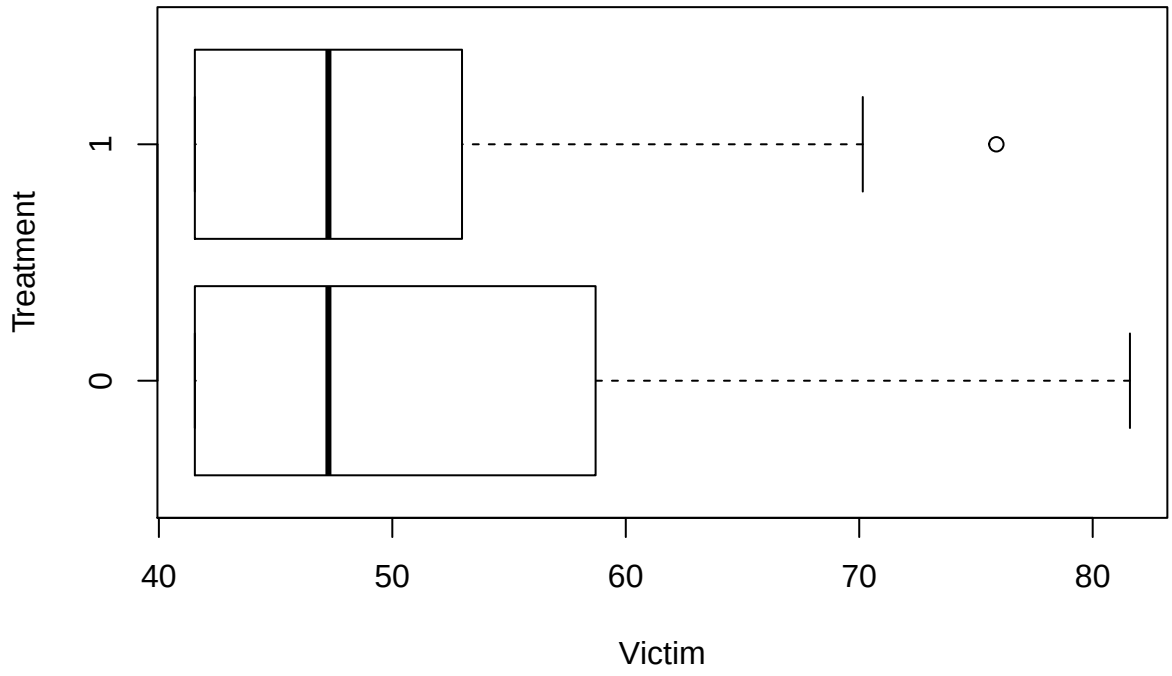
```r
rm(.Random.seed, envir=globalenv())

# use read.csv() and the raw data link
afterschool <- read.csv('https://raw.githubusercontent.com/zief0002/comparing-groups/master/data/AfterS
attach(afterschool)

# Exploratory analysis
explo_data3 <- afterschool %>% group_by(Treatment) %>% summarise(mean = mean(Victim), sd = sd(Victim), n
explo_data3
```

```
## # A tibble: 2 x 4
##   Treatment  mean    sd     n
##       <int> <dbl> <dbl> <int>
## 1         0  50.6 10.9    187
## 2         1  49.3  8.84   169
```

```r
# Boxplot
boxplot(Victim ~ Treatment, horizontal = TRUE, data = afterschool)
```



Randomization Test

Use the randomization test.

```r
# Creating our obs_dif3
controlled <- Victim[1:187] # n = 187
treated <- Victim[188:356] # n = 169
obs_dif3 <- mean(controlled) - mean(treated)

# Hypothesis Test:
# H0: mean(control) = mean(treatment)
# Ha: mean(control) > mean(treatment)
set.seed(1)
differences3 <- rep(NA, 10000)

for(i in seq_along(differences3)){
randomized <- sample(Victim)
control <- randomized[1:187]
treatment <- randomized[188:356]
differences3[i] <- mean(control) - mean(treatment)
}

# The empirical p-value result using two-sided alternative
```

```r
mean(abs(differences3) >= obs_dif3)
```

```
## [1] 0.2227
```

We would NOT REJECT the NULL Hypothesis, and this means that there may not be a significant difference between the means of controlled and treated.

Part 4

Randomization test

Perform a randomization test. I created the data having Outcome: Live = TRUE.

```r
rm(.Random.seed, envir=globalenv())
# Outcome: Live = TRUE

# Creating the data for ecmo_data
ecmo_data <- c(rep(TRUE, 6), rep(FALSE, 4), rep(TRUE, 28), rep(FALSE, 1))

# Creating our obs_dif2
cmt_total <- ecmo_data[1:10]
ecmo_total <- ecmo_data[11:39]
obs_dif4 <- mean(ecmo_total) - mean(cmt_total)

# Checking the averages of these two data
mean(cmt_total) # 6 out of 10
```

```
## [1] 0.6
```

```r
mean(ecmo_total) # 28 out of 29
```

```
## [1] 0.9655172
```

```r
# Hypothesis Test:
# H0: mean(Treatment: ECMO) = mean(Treatment: CMT)
# Ha: mean(Treatment: ECMO) > mean(Treatment: CMT)
set.seed(1)
differences4 <- rep(NA, 10000)
for(i in seq_along(differences4)) {
  randomized <- sample(ecmo_data)
  groupA <- randomized[11:39]
  groupB <- randomized[1:10]
  differences4[i] <- mean(groupA) - mean(groupB)
}

# The empirical p-value result using one-sided alternative
mean((differences4) >= obs_dif4)
```

```
## [1] 0.0106
```

We would REJECT the NULL Hypothesis, and this means that there is a significant difference between the means of the Treatment type for ECMO and CMT.

Comparison to Fisher's Exact Test

Use R's `fisher.test()`

```r
fisher_test <- fisher.test(rbind(c(4, 1), c(6, 28)), alternative = "greater")$p.value

random_test_4 <- mean((differences4) >= obs_dif4)

fisher_test - random_test_4
```

```
## [1] 0.0004150636
```

We see there is a 0.0004150636 value difference between our Randomized Test's p-value and Fisher Test's p-value. They both yield the same conclusion, which is REJECT the NULL Hypothesis. This difference is far too small to make any difference, and we can say either methods work.

Part 5

Comparing Groups, Chapter 7, Exercise 7.1

Non-parametric bootstrap test

Use a non-parametric bootstrap test

```r
rm(.Random.seed, envir=globalenv())
# Non-parametric bootstrap

# use read.csv() and the raw data link
hsb <- read.csv('https://raw.githubusercontent.com/zief0002/comparing-groups/master/data/HSB.csv')
attach(hsb)

# Exploratory Analysis
explo_data5 <- hsb %>% group_by(Schtyp) %>% summarize(mean = mean(Sci), sd = sd(Sci), n = n())
explo_data5
```

```
## # A tibble: 2 x 4
##   Schtyp  mean    sd     n
##    <int> <dbl> <dbl> <int>
## 1      0  53.3  8.19    32
## 2      1  51.6 10.2    168
```

```r
# Creating our obs_dif5
public <- hsb[Schtyp == 1,]
private <- hsb[Schtyp == 0,]
obs_dif5 <- var(public$Sci) - var(private$Sci)

# Hypothesis Test:
# H0: var(Public) = var(Private)
# Ha: var(Public) > var(Private)
set.seed(1)
differences5 <- rep(NA, 10000)
for(i in seq_along(differences5)){
  sample_public <- sample(Sci, nrow(public), replace = TRUE)
  sample_private <- sample(Sci, nrow(private), replace = TRUE)
  differences5[i] <- var(sample_public) - var(sample_private)
}

# The empirical p-value result using two-sided alternative
mean(abs(differences5) >= obs_dif5)
```

```
## [1] 0.1061
```

We would NOT REJECT the NULL Hypothesis, and this means that there may not be a significant difference between the variances of Public and Private school types.

Parametric bootstrap test

Use a parametric bootstrap test

```
rm(.Random.seed, envir=globalenv())
# Parametric bootstrap

# Exploratory Analysis
explo_data5 <- hsb %>% group_by(Schtyp) %>% summarize(mean = mean(Sci), sd = sd(Sci), n = n())
explo_data5
```

```
## # A tibble: 2 x 4
##   Schtyp  mean    sd     n
##    <int> <dbl> <dbl> <int>
## 1      0  53.3  8.19    32
## 2      1  51.6 10.2    168
```

```
# Creating our obs_dif5 & finding the mean(Sci) and var(Sci)
public <- hsb[Schtyp == 1,]
private <- hsb[Schtyp == 0,]
obs_dif5 <- var(public$Sci) - var(private$Sci)
m <- Sci %>% mean()
s <- Sci %>% sd()

# Hypothesis Test:
# H0: var(Public) = var(Private)
# Ha: var(Public) > var(Private)
set.seed(1)
differences5 <- rep(NA, 10000)
for(i in seq_along(differences5)) {
  groupA <- rnorm(168, m, s)
  groupB <- rnorm(32, m, s)
  differences5[i] <- var(groupA) - var(groupB)
}

# The empirical p-value result using two-sided alternative
mean(abs(differences5) >= obs_dif5)
```

```
## [1] 0.1619
```

We would NOT REJECT the NULL Hypothesis, and this means that there may not be a significant difference between the variances of Public and Private school types.

Part 6

Non-parametric bootstrap test

Perform a bootstrap test

```
rm(.Random.seed, envir=globalenv())

# Creating the data for light
```

```r
light <- c(28, 26, 33, 24, 34,-44, 27, 16, 40, -2, 29, 22, 24, 21, 25, 30, 23, 29, 31, 19,
24, 20, 36, 32, 36, 28, 25, 21, 28, 29, 37, 25, 28, 26, 30, 32, 36, 26, 30, 22,
36, 23, 27, 27, 28, 27, 31, 27, 26, 33, 26, 32, 32, 24, 39, 28, 24, 25, 32, 25,
29, 27, 28, 29, 16, 23)

# Checking the average center for light
mean(light)
```

```
## [1] 26.21212
```

```r
# Recenter the vector for light called newlight
newlight <- light - mean(light) + 33

# Checking the average center for newlight
mean(newlight)
```

```
## [1] 33
```

```r
# Creating our obs_dif6
obs_dif6.1 <- mean(newlight) - mean(light)


# Non-parametric bootstrap function called dif_func6 for our boot() function
dif_func6 <- function(data, indices) {
  a <- data[indices]
  mean(a) - mean(newlight)
}

# Hypothesis Test:
# H0: mean(newlight) = mean(light)
# Ha: mean(newlight) > mean(light)
set.seed(1)
bootstrap_light <- boot(newlight, dif_func6, 100000)$t

# The empirical p-value result using two-sided alternative
mean(abs(bootstrap_light) >= obs_dif6.1)
```

```
## [1] 2e-05
```

We would REJECT the NULL Hypothesis, and this means that there is a significant difference between the means of light and newlight.

Non-parametric bootstrap test with outliers removed

Perform the bootstrap test again after removing the two negative outliers (-2, and -44)

```r
rm(.Random.seed, envir=globalenv())

# Remove the two outliers (-2 and -44 are outliers as seen in the data)
improved_light <- light[which(light != "-2" & light != "-44")]

# recenter the vector again
newlight2 <- improved_light - mean(improved_light) + 33

# Checking the average center for newlight2
mean(newlight2)
```

```
## [1] 33
```

```
# Creating our obs_dif6.2
obs_dif6.2 <- mean(newlight2) - mean(light)

# Non-parametric bootstrap function called dif_func6 for our boot() function
dif_func6 <- function(data, indices) {
  a <- data[indices]
  mean(a) - mean(newlight)
}

# Hypothesis Test:
# H0: mean(newlight2) = mean(light)
# Ha: mean(newlight2) > mean(light)
set.seed(1)
bootstrap_light2 <- boot(newlight2, dif_func6, 100000)

# The empirical p-value result using two-sided alternative
mean(abs(bootstrap_light2$t) >= obs_dif6.2)
```

```
## [1] 0
```

We would REJECT the NULL Hypothesis, and this means that there is a significant difference between the means of light and newlight2.