Math 536 Homework 1

Shamir Naran

Due February 6, 2020

Problem 1:

Download the HW1P1.csv filed. In this file you will find two variables, females and males. Females contains the salary on a random sample of female employees from a tech company in Northern California. Males contains a random sample of males from the same tech company. Your goal is to investigate whether or not there is gender discrimination within that company with regards to pay (i.e. males are making more than females).

```
HW1P1 <- read.csv("HW1P1.csv", header=T, na.strings="?")
# We read in the data using the read.csv function.
# The header = T tells R that the first line of the file contains the variable names.
# na. strings tells R that when it sees a particular set of characters it should be treated as a missin
```

```
HW1P1.Females <- na.omit(HW1P1$Females) # remove NA's
HW1P1.Males <- na.omit(HW1P1$Males)
```

```
x.bar.f <- mean(HW1P1.Females, na.rm = TRUE)
x.bar.m <- mean(HW1P1.Males, na.rm = TRUE)
s.f <- sd(HW1P1.Females, na.rm = TRUE)
s.m <- sd(HW1P1.Males, na.rm = TRUE)
n.f <- length(HW1P1.Females) # count the number of non NA's in your column
n.m <- length(HW1P1.Males)
df <- ((s.f^2/n.f)+(s.m^2/n.m))^2 / ((s.f^2/n.f)^2/(n.f-1) + (s.m^2/n.m)^2/(n.m-1))
# df freedom using Satterthwaite's approximation
```

Part a:

First run a valid statistical test using a central limit theorem (i.e. the classical theoretical way). Please report and interpret your p-value.

H(null): mu(m)-mu(f) $= 0$ H(alt): mu(m)-mu(f) > 0

```
t <- (x.bar.m - x.bar.f) / sqrt((s.m^2/n.m) + (s.f^2/n.f))
# compute our test statistic for a two sample T test
p.val <- pt(-abs(t), df = df) # compute p-value
p.val
```

```
## [1] 0.01426607
```

Our p-value is approximately 0.0143. This is the probability, assuming the null hypothesis is true, that the test statistic will take a value at least as extreme as that actually observed. Judging from the p-value, it is unlikely that the true mean of female salaries is equal to the true mean of male salaries.

```r
t.star <- qt(.975, df = df) # take the lowest n
lb <- (x.bar.m - x.bar.f) - t.star*sqrt((s.f^2/n.m) + (s.m^2/n.f))
ub <- (x.bar.m - x.bar.f) + t.star*sqrt((s.f^2/n.m) + (s.m^2/n.f))
c(lb,ub)
```

```
## [1]  1301.339 27259.202
```

We compute a 95% confidence interval. Notice the value of 0 is not contained in our confidence interval. We are 95% confident that the true difference between male and female salaries lies between (1301.34, 27259.20).

Part b:

Now repeat the process of statistical inference but this time you may not assume anything about your sample summary. You must instead bootstrap a p-value.

```r
set.seed(1) # for reproducibility
bs.xbar.difference <- rep(0,10000)
for (i in 1:10000) {
  f <- sample(HW1P1.Females, size = n.f, replace = TRUE)
  m <- sample(HW1P1.Males, size = n.m, replace = TRUE)
  bs.xbar.difference[i] <- mean(m)-mean(f)
}
```

```r
quantile(bs.xbar.difference,c(.025,.975)) # 95% confidence interval
```

```
##      2.5%     97.5%
##  1648.739 26858.077
```

We compute a 95% confidence interval. Again, notice the value of 0 is not contained in our confidence interval. We are 95% confident that the true difference between male and female salaries lies between (1648.74,26858.08).

```r
bs.xbar.star <- bs.xbar.difference - (x.bar.m - x.bar.f)
# recenter our 10,000 values at 0
p.val.bs <- length(bs.xbar.star[bs.xbar.star > (x.bar.m - x.bar.f)]) / length(bs.xbar.star)
p.val.bs
```

```
## [1] 0.0143
```

Our bootstrap p-value is 0.0143. Very similar to the p-value in part a. Judging from the p-value, it is unlikely that the true mean of female salaries is equal to the true mean of male salaries. There is evidence to suggest that male salaries are higher than female salaries at this tech company. Of course, there are other variables that should be investigated such as number of hours worked, position title, years of experience, ... etc. In my opinion, gender discrimination with regards to pay can't be concluded from this study.