Homework 1

Michael Pena

2024-02-07

Question 1

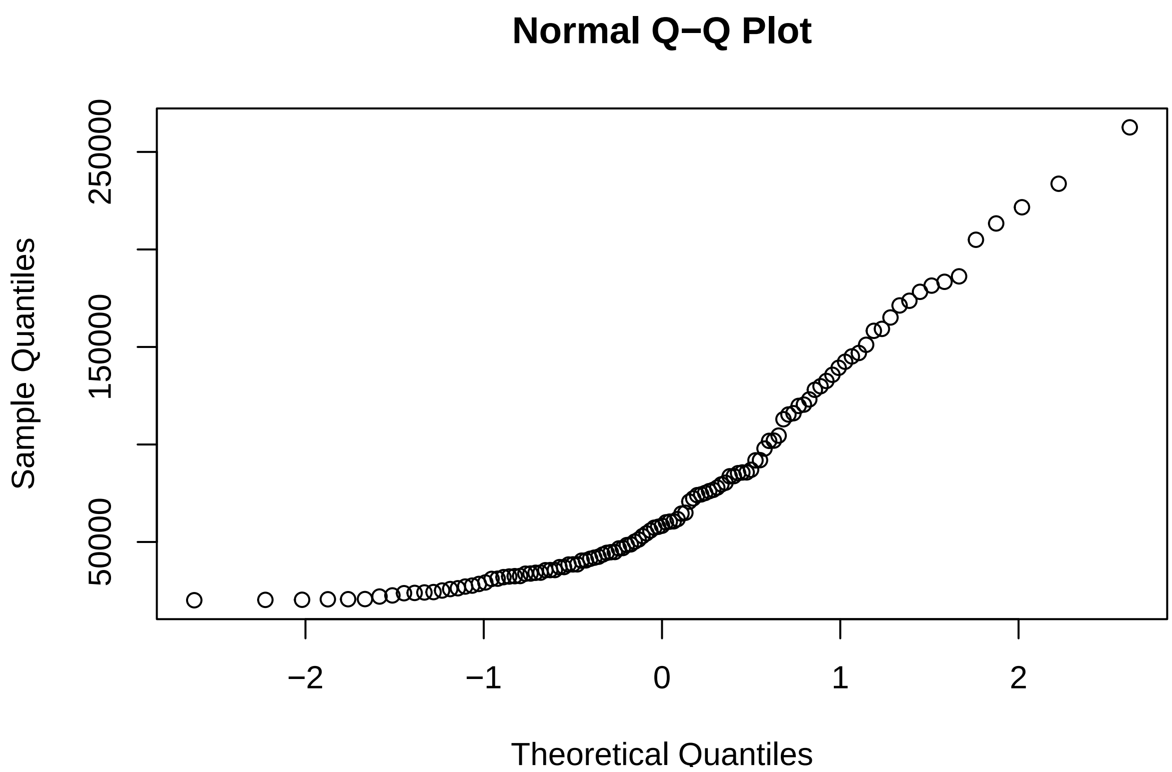
++
DIRECTIONS: Download the HW1P1.csv file. In this file you will find two variables, females and males. Females contains the salary on a random sample of female employees from a tech company in Northern California. Males contains a random sample of males from the same tech company. Your goal is to investigate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

```r
# load data
data <- read.csv("HW1P1.csv")
```

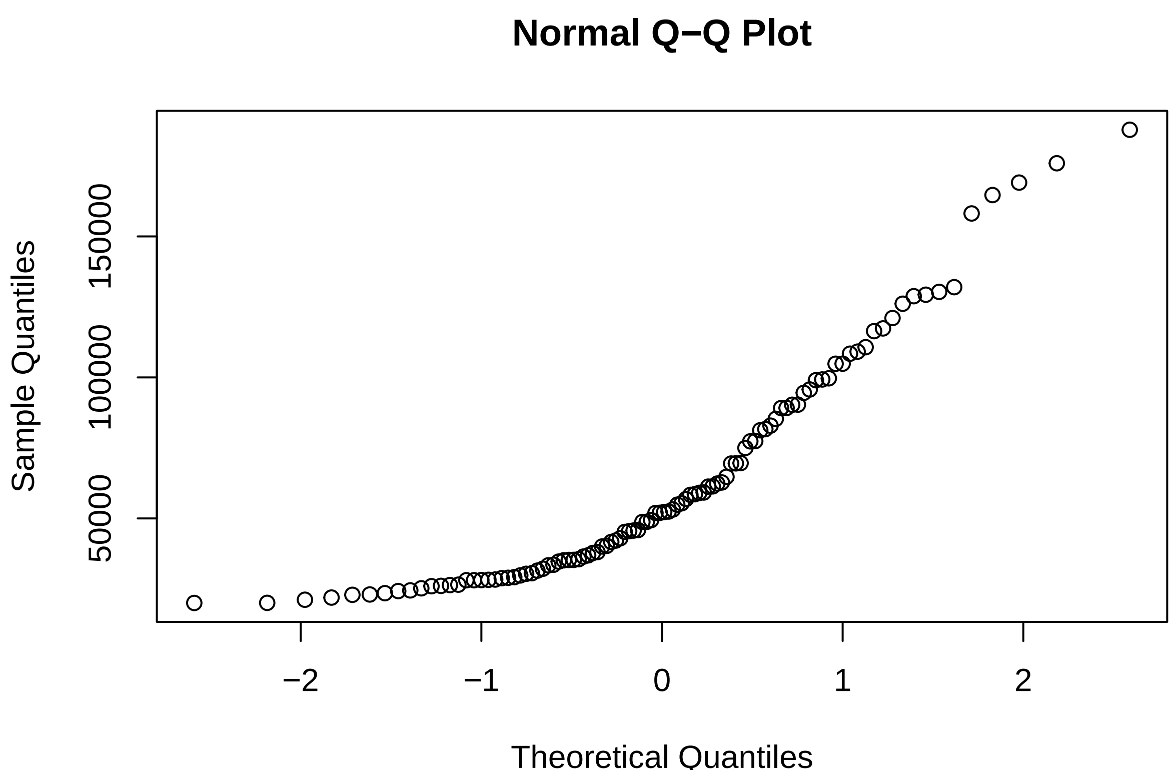
Part a:

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

```r
# make a new column that is the difference between the two
male_vec <- data$Males
fem_vec <- na.omit(data$Females)
# check normality of the difference
qqnorm(male_vec)
```




```
qqnorm(fem_vec)
```




```
# let H_alt be mu > 0 with 99% confidence
# run student t.test
t.test(male_vec,fem_vec, alternative = "greater", paired = F, conf.level = .99)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_vec and fem_vec
## t = 2.2054, df = 205.8, p-value = 0.01427
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -901.2662       Inf
## sample estimates:
## mean of x mean of y
##  78034.08  63753.81
```

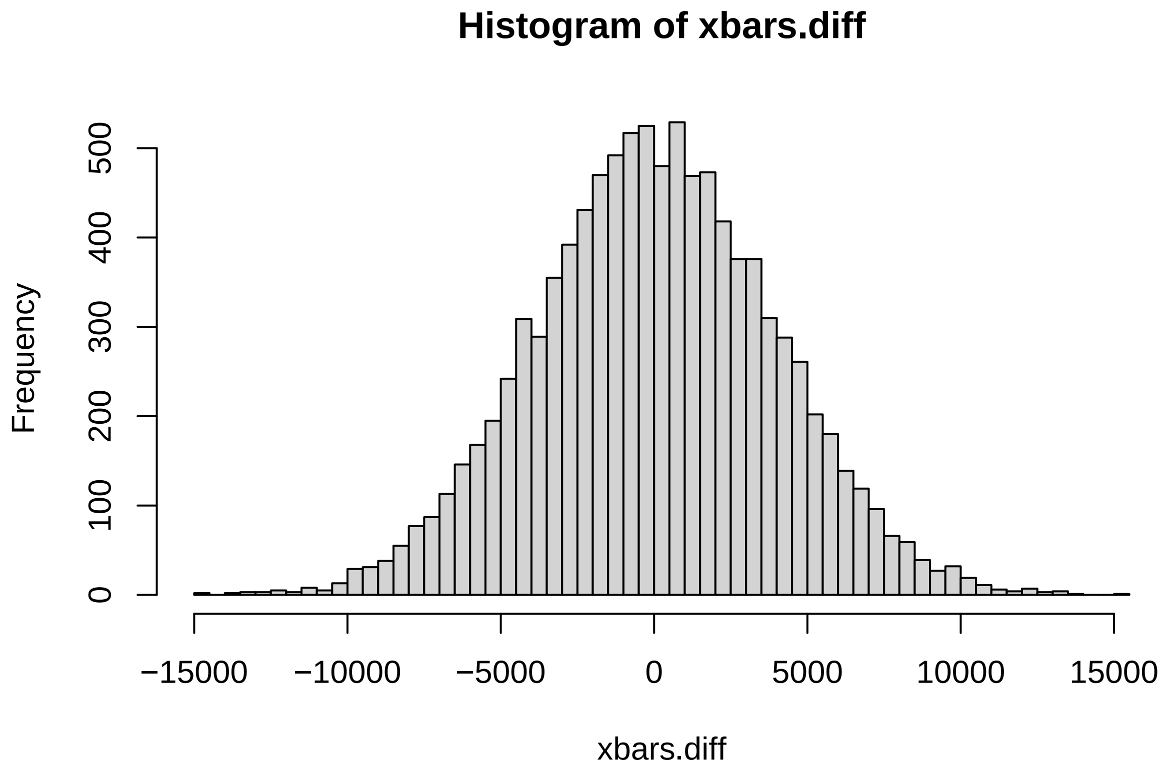
The selected $\alpha = 0.01$ and here p-value is greater that 0.01. Thus we fail to reject the null hypothesis.

conclusion: There is not sufficient sample evidence to support the claim that "males are making more than females."

part b.

++ DIRECTIONS: 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. Please write a small report, no more than a paragraph or two relating your results to the company's interests in discovery. Please only provide relevant statistical output. ++

```r
# find population mean
set.seed(536)
BS.male_vec <- male_vec - mean(male_vec)
BS.fem_vec <- fem_vec - mean(fem_vec)
BS.pop.mean <- mean(male_vec) - mean(fem_vec)


# bootstrapping

xbars.diff <- rep(0,10000)
for (i in 1:10000){
  sampleM <- sample(BS.male_vec,300, replace = T)
  sampleF <- sample(BS.fem_vec,300, replace = T)
  xbars.diff[i] <- mean(sampleM - sampleF)
}

# render histogram
hist(xbars.diff, breaks = 80)
```




```
# finding a p-value
pval <- length(xbars.diff[xbars.diff > BS.pop.mean])/10000
pval
```

```
## [1] 1e-04
```

The histogram suggests we can follow a normally distributed data set with our bootstrapped data. Because the original question was asking if the company paid men more than women; we ran a right tailed test with an alternative hypothesis $\mu_d > 0$ where μ_d represents the average male income minus the average female income. We took several samples samples of 300 with replacement from male and female incomes, recorded the average difference, and repeated this process ten thousand times. The P-value we return is very small number that is degrees smaller than our chosen $\alpha = 0.01$ (I chose 99% significance). This bootstrap method concludes that there is significant evidence that males incomes is by average higher than that of their female counterparts. It may be necessary for the company to address this pay disparity.

Question 2

++ DIRECTIONS: Does bootstrapping always work? In this problem we want to begin with a population, I don't care what your population is but something robust (maybe like 50,000 data observations from a well defined numerical distribution). ++

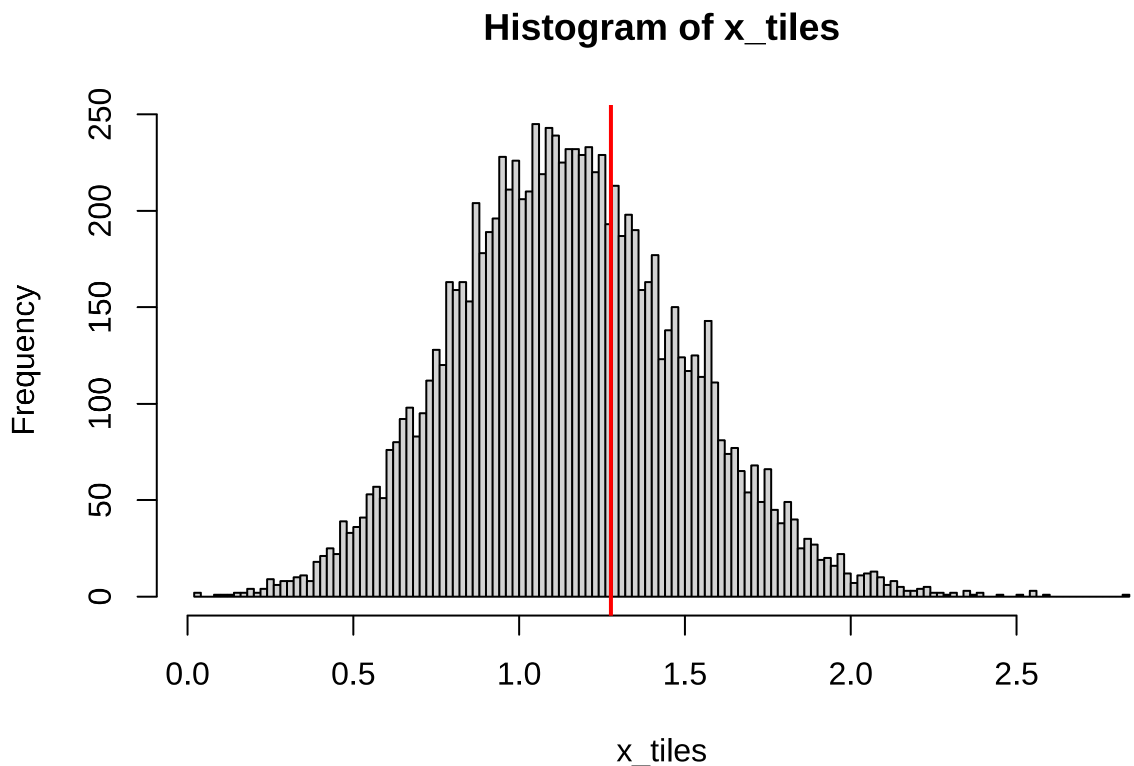
```
#generate observations
obs <- rnorm(100000)
```

part a.

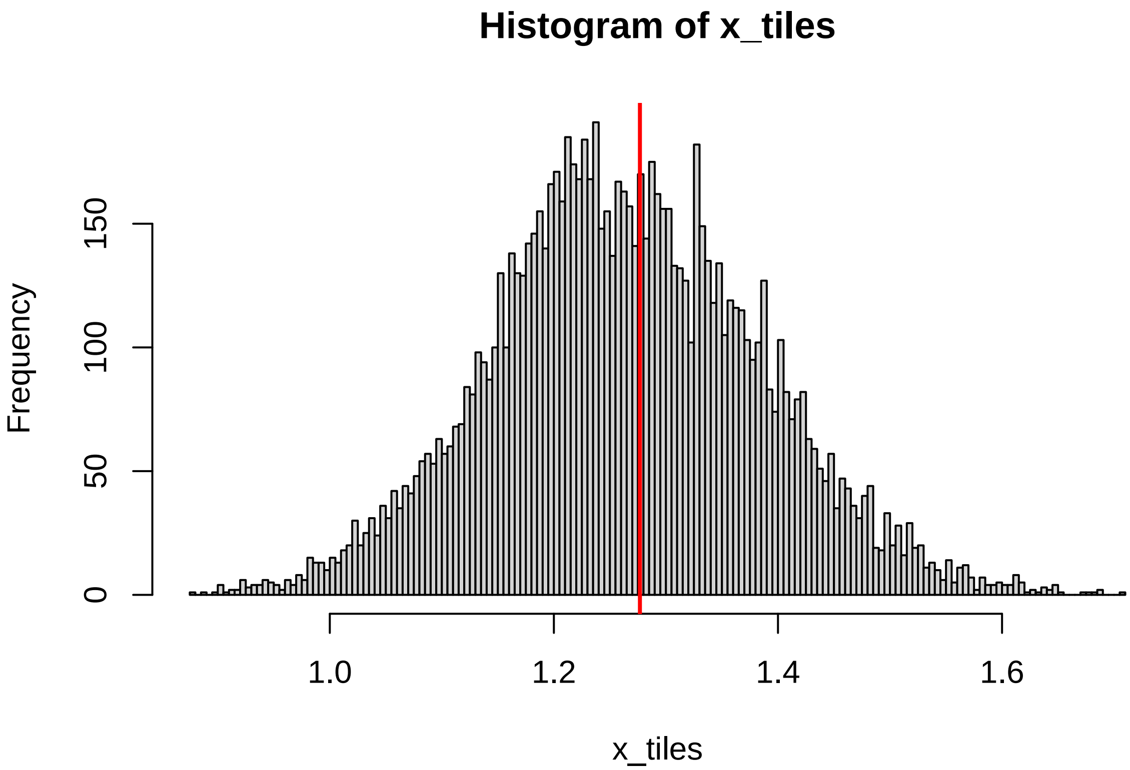
++ DIRECTIONS: For a given sample size of 20, draw 10,000 samples, all of size 20. Compute the 90%tile of each sample. Plot the 90%tiles of each sample along with the true 90%tile of your population. Do you believe that the 90%tile of a sample of size 20 is an unbiased estimator of the population parameter 90%tile?

++

```
x_tiles <- rep(0,10000)
true_90th <- quantile(obs, .9)

for (i in 1:10000){
  samp <- sample(obs,20,replace = T)
  x_tiles[i] <- quantile(samp, .9)
}


hist(x_tiles, breaks = 160)
abline(v = true_90th,col = "red", lwd = 2)
```




```
# bootstrap now but with a larger sample size in each iteration

for (i in 1:10000){
  samp <- sample(obs,200,replace = T)
  x_tiles[i] <- quantile(samp, .9)
}


hist(x_tiles, breaks = 160)
abline(v = true_90th,col = "red", lwd = 2)
```

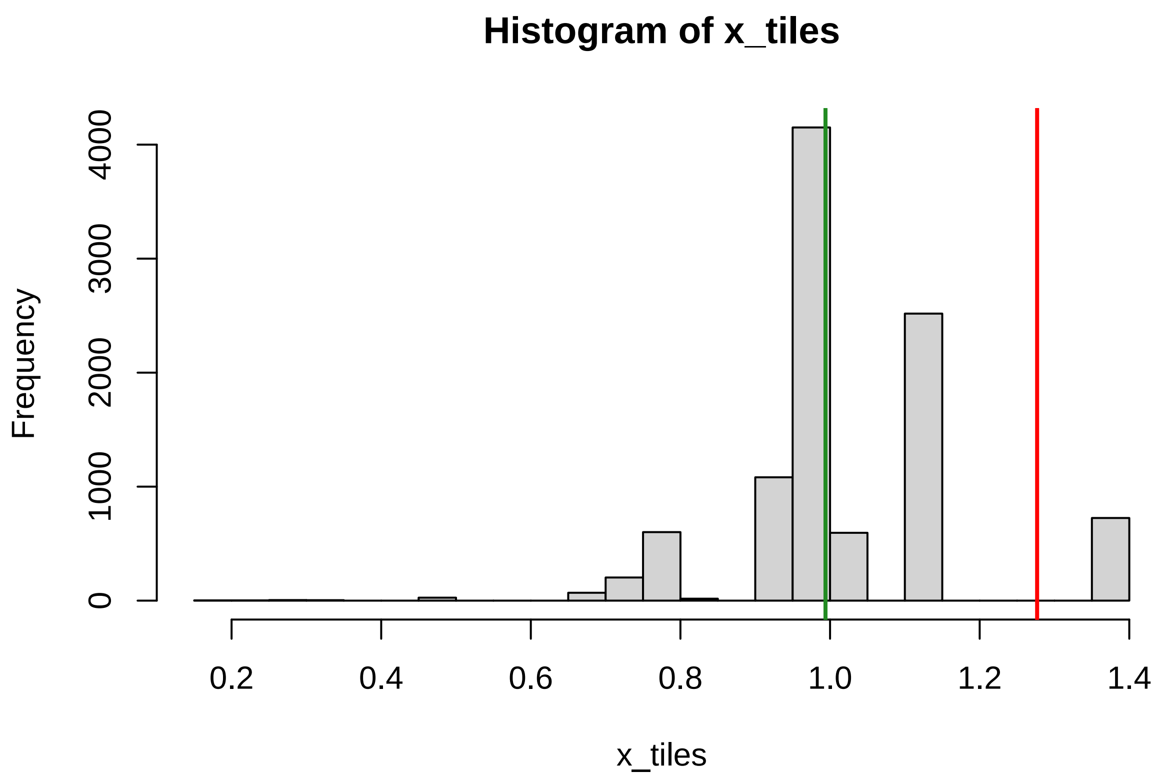


Using only a sample size of 20 is biased as when I highered the sample size in the second graphic, the true 90%tile got closer to median.

part b.

++
DIRECTIONS: Take a single sample of size 20. Record the 90%tile. Now draw 10,000 bootstrap samples of size 20 from your original sample. Plot the 10,000 BS estimates of the 90%tile along with the true 90%tile of your original sample. Are your bootstrap estimates of the 90%tile biased? Can you quantify the amount of bias?
++

```
single.samp <- sample(obs,20,replace = T)
single.samp.90 <- quantile(single.samp,.9)

for (i in 1:10000){
  samp <- sample(single.samp,20,replace = T)
  x_tiles[i] <- quantile(samp, .9)
}

hist(x_tiles, breaks = 20)
abline(v = true_90th,col = "red", lwd = 2)
abline(v = single.samp.90,col = "forestgreen", lwd = 2)
```



There seems to be heavy bias in our current method but we also notice how unnormal the data is distributed. This would be difficult to argue as the graph looks really bad (lacking data to make an solid conclusion).

```r
# quantifying bias
bias <- mean(x_tiles) - single.samp.90
```

part c.

++ DIRECTIONS: Combining parts a. and b. Take a single sample of size 20 from your population and come up with a Bootstrapped Confidence interval for the 90%tile of the population… Don't forget to correct for bias! Please submit your plots with brief discussion for each part. (I really want you to argue how you're going to account for bias) ++

we will account for bias by taking the mean of our boostrapped 90th percentiles vector and and finding the difference between that and our randomly sampled 90th percentile from the sample of size 20.