HWK 1 - Math 536

Elijah Amirianfar

2024-02-03

Problem 1) 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).

```
hw1_data <-
  read.csv("~/My Drive/04. CSU FULLERTON 2023-2025/2. Spring 2024/MATH 536/Data Sets/HW1P1.csv")
```

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.

```
female <- na.omit(hw1_data[,1])
male <- na.omit(hw1_data[,2])

#this computes the hypothesis test for our above experiment
t.test(male,female, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = 2.2054, df = 205.8, p-value = 0.01427
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3581.523      Inf
## sample estimates:
## mean of x mean of y
##  78034.08  63753.81
```

Here, our p-value is 0.01427 which means we reject the null and claim that our alternate is true, where the mean of the male salaries is higher than the mean of the female salaries.

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.

```
population_mean = mean(c(male,female))
# here we combine all the male and female salaries into 1 single set to get the population mean
BS.pop_male = male - mean(male) + population_mean
BS.pop_female = female -  mean(female) + population_mean
# here we take each male and female datasets, then subtract each one by the means of each gender
# then add the population mean to normalize each data set to get a better setup for the testing

single.samp.male = sample(BS.pop_male,length(BS.pop_male),replace="T")
single.xbar.male = mean(single.samp.male)
# here, we take a single sample from the male bootstrapped data, and get the mean of the sample
single.samp.female = sample(BS.pop_female,length(BS.pop_female),replace="T")
single.xbar.female = mean(single.samp.female)
# here, we take a single sample from the female bootstrapped data, and get the mean of the sample


BS.xbars_diff = rep(0,10000)

for(i in 1:10000){
    single.samp.male = mean(sample(BS.pop_male,length(BS.pop_male),replace="T"))
    single.samp.female = mean(sample(BS.pop_female,length(BS.pop_female),replace="T"))
    BS.xbars_diff[i] = single.samp.male-single.samp.female
    # here, we take a single sample from the male and female bootstrapped data,
    # get the mean of each sample, compute the difference between the male and female
    # single sample means and put it into our vector. this is done 10000 times

}

length(BS.xbars_diff[BS.xbars_diff>(mean(male)-mean(female))])/10000
```

```
## [1] 0.0155
```

```
# here, we want to count the number of times where the difference salaries is greater than the
# difference of the means for male and female salaries
```

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.

In our investigation, we were given data from a company related to the salaries of its employees. We want to test if the mean of the salaries from the male employees is larger than the mean of the salaries from the female employees. Thus, we have the following hypothesis test:

$$H_0 : \mu_{male} = \mu_{female}, \quad H_a : \mu_{male} > \mu_{female}$$

After computing a T-test, we discovered that our p-value was 0.01427, which means we reject the null and claim that the average of male salaries is greater than the average of female salaries. However, in order to double check our work, we did a bootstrap of our data to compute our p-value in a more reliable way. And by doing so, our p-value ends up being around our original value of 0.01427. Thus, we can reject the null hypothesis, and conclude that the average of male's salaries is higher than the average of female's salaries. Essentially, men get paid more than women in the company.

Problem 2) 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). Also, please submit your plots with brief discussion for each part.

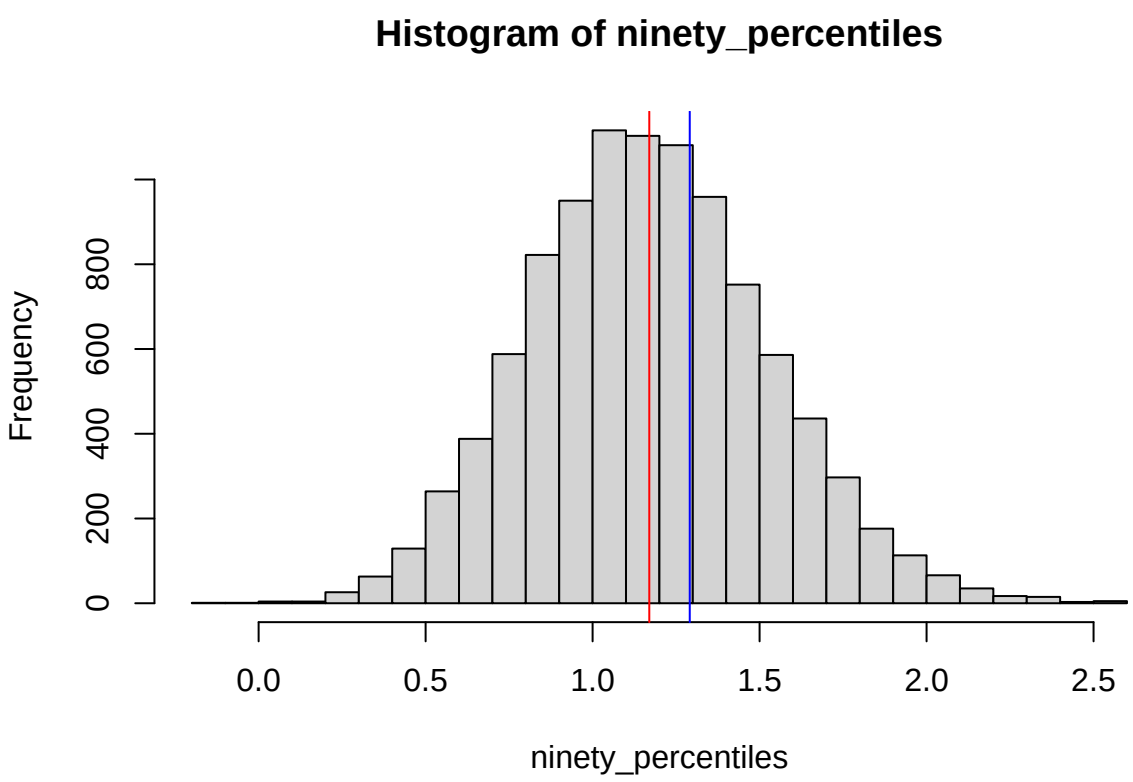
Part a) For a given sample size of 20, draw 10,000 samples, all of size 20. Compute the 90%tile of each sample. Plot the 90%tiles of each sample along with the true 90%tile of your population. Do you believe that the 90%tile of a sample of size 20 is an unbiased estimator of the population parameter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

```r
set.seed(536)
pop_data <- rnorm(50000, mean = 0, sd = 1.0)
# above is our random data from a Standard normal distribution

ninety_percentiles <- seq(0,0,length=10000)
# an empty vector where we add the 90th percentile of each sample

for (i in 1:10000){
  ninety_percentiles[i] <- quantile(sample(pop_data, size = 20,replace="T"), probs = .9)
  # computes the 90th percentile of each sample from the population
}

hist(ninety_percentiles, breaks = 20)
abline(v = mean(ninety_percentiles), col = "red")
abline(v = quantile(pop_data, prob = 0.9), col = "blue")
```




```
first_bias = abs(mean(ninety_percentiles)-quantile(pop_data, prob = 0.9))
first_bias
```

```
##       90%
## 0.1211522
```

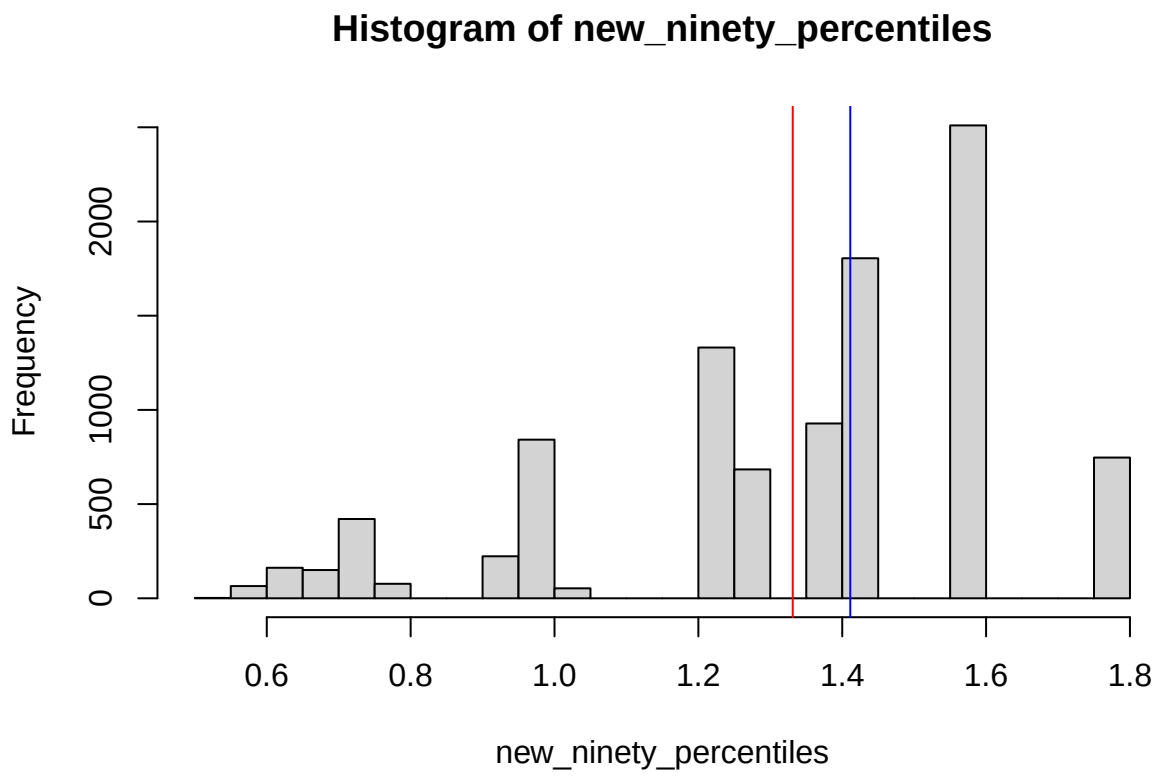
I believe that the sample of size 20 is a biased estimator of the population parameter 90th percentile since our bias is 0.121.

Part b) 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?

```
set.seed(536)
new_sample = sample(pop_data, size = 20,replace="T")
# here we take a single sample of size 20
new_90_percentile = quantile(new_sample,0.9)
# here we record the 90th percentile of that sample we found

new_ninety_percentiles = seq(0,0,length=10000)
# an empty vector where we add the 90th percentile of each new sample
```

```r
for (i in 1:10000){
  new_ninety_percentiles[i] <- quantile(sample(new_sample, size = 20,replace="T"), probs = .9)
  # computes the 90th percentile of a new sample from NEW_SAMPLE
}
hist(new_ninety_percentiles, breaks = 20)
abline(v = mean(new_ninety_percentiles), col = "red") # represents mean of new 90th percentiles
abline(v = quantile(new_sample, prob = 0.9), col = "blue") # represents 90th percentile of new sample
```




```r
second_bias = abs(mean(new_ninety_percentiles)-quantile(new_sample, prob = 0.9))
second_bias
```

```
##        90%
## 0.08001447
```

Here, my mean of the new_ninety_percentiles is 1.331203 (red line) and the quantile of the new sample (blue line) is above 1.4. Thus, we can quantify the amount of bias present in our situation. However, every time we calculate bias, we get vastly different results. And so it is not accurate.

Part c) 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!

```r
set.seed(536)
final_sample = sample(pop_data, size = 20,replace="T")
# new single sample from population

final_ninety_percentiles = seq(0,0,length=10000)
# an empty vector where we add the 90th percentile of each new sample

for (i in 1:10000){
  final_ninety_percentiles[i] <- quantile(sample(final_sample, size = 20,replace="T"), probs = .9) + fi
  # computes the 90th percentile of a new sample from FINAL_SAMPLE
}

a = quantile(final_ninety_percentiles,.025)
b = quantile(final_ninety_percentiles,.975)
c(a,b)
```

```
##      2.5%     97.5%
## 0.7924439 1.8718824
```