# BACS_HW5

106022113, helped by 106022103

## Problem 1 : Verify Google DOI apps

### (a) : Probability of a random app that is malicious

```
prob <- pnorm(-3.7,0,1)
paste("The probability of apps being malicious is: ", prob)
```

```
## [1] "The probability of apps being malicious is:  0.000107799733477388"
```

Since the binomial distribution is considered approximately to be normal, we can just use the function *pnorm* to get the probability.

### (b) : How many apps are malicious in 2.2 million?

```
paste("Approximately ",round(prob*2.2*10^6)," apps are malicious in 2.2 million apps.")
```

```
## [1] "Approximately  237  apps are malicious in 2.2 million apps."
```

## Problem 2 : Verizon Repairing Phones

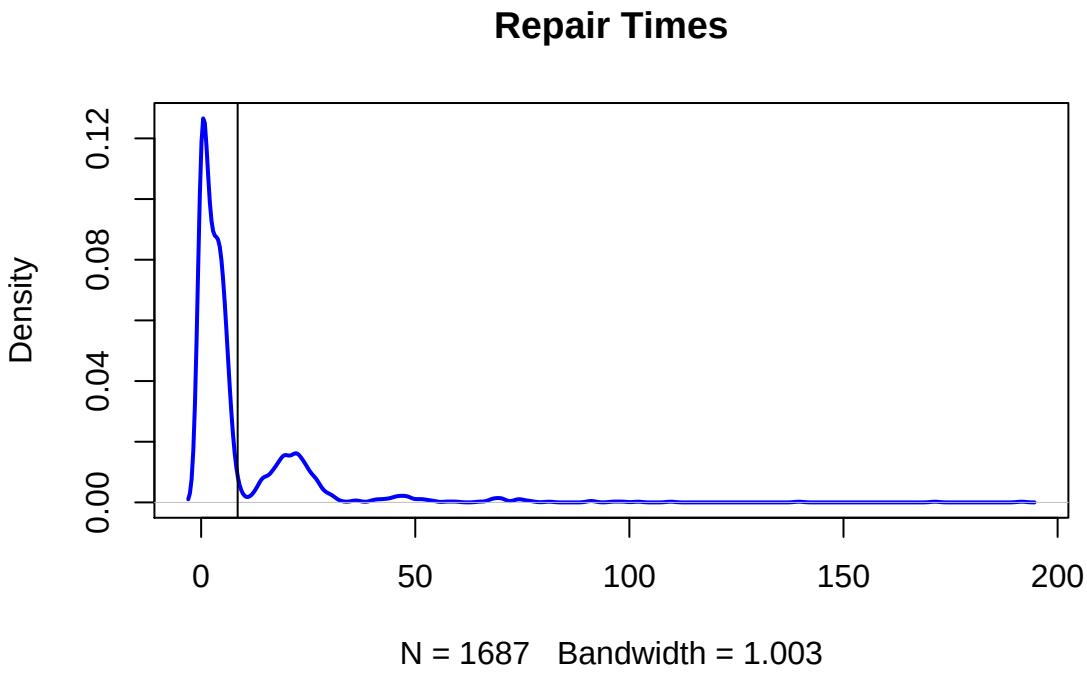
```
data <- read.csv("verizon.csv")
head(data)
```

```
##    Time Group
## 1 17.50  ILEC
## 2  2.40  ILEC
## 3  0.00  ILEC
## 4  0.65  ILEC
## 5 22.23  ILEC
## 6  1.20  ILEC
```

### (a) : NULL Distribution of t-values

```
times <- data$Time
plot(density(times), col="blue", lwd=2,main = "Repair Times")
abline(v = mean(times))
```

**i. Visualize distribution and marking the mean**




**ii. Testing Hypothesis** Since Verizon claims that the repair time is 7.6 minutes, we can set $H0 = 7.6minutes$, and then we have $H1 \neq 7.6minutes$

**iii. Estimate Population Mean and 99% confidence interval** Population mean can be estimated from the sample mean and the confidence interval can be found using the equation below.

$$\mu \pm z_{0.995}\frac{\sigma}{\sqrt{n}}$$

However, this equation assumes the distribution is normal since it uses a factor calculated from normal distribution percentile. Hence, the best way to estimate the confidence intervals is by using quantile.

```
#ci2_99 <- quantile(times,probs = c(0.005,0.995))
conf1 <- mean(times)-(qnorm(0.995,0,1)*sd(times)/sqrt(length(times)))
conf2 <- mean(times)+(qnorm(0.995,0,1)*sd(times)/sqrt(length(times)))
paste("Population Mean: ",mean(times),", 99% Confidence interval:(",conf1,",",conf2,")")
```

```
## [1] "Population Mean:  8.52200948429164 , 99% Confidence interval:( 7.59457509656072 , 9.44944387202
```

**iv. Find t-value, p-value using statistics measures** We can have the t-value with the equation below and then derive p-value from it.

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

```r
t <- (mean(times)-7.6)/(sd(times)/sqrt(length(times)))
p <- 2*(1-pt(t,length(times)-1))# two-tailed
paste("t-statistic: ",t, ", p-value: ",p)
```

```
## [1] "t-statistic:  2.56076233446444 , p-value:  0.010530684588578"
```

**v. Describe these values to NULL Distribution** The t-statistic is the ratio of the difference of the estimated value from its hypothesized value to its standard error. It will give us the answer if the hypothesis is true or it lacks further evidence. The p-value is the probability that the results from your sample data occurred by chance.

**vi. Conclusions of the test** Since the p-value is $0.0106 > 0.01$, we can arrive at the conclusion that we don't have enough evidence to reject the null hypothesis. Because our stats indicate that we have covered $1 - 0.0053$ percentage of the distribution, which is more than 99%.

**(b) Using bootstrapping on the sample data:**

```r
sample_statistic <- function(stat_func, sample0){
  resample <- sample(sample0, length(sample0),replace = TRUE)
  stat_func(resample)
}
sample_means <- replicate(length(times),sample_statistic(mean,times))
#Confidence Interval:
ci_99 <- quantile(sample_means, probs = c(0.005,0.995))
paste("The confidence interval is: ",ci_99[1],",",ci_99[2])
```

**i. 99% CI of Bootstrapped means**

```
## [1] "The confidence interval is:  7.68726490812092 , 9.50289952578542"
```

```r
boot_mean_diffs <- function(sample0, hyp){
  resample <- sample(sample0, length(sample0),replace = TRUE)
  return(mean(resample)-hyp)
}
mean_diffs <- replicate(length(times),boot_mean_diffs(times,7.6))
#Confidence Interval
ci1_99 <- quantile(mean_diffs,c(0.005,0.995))
paste("The confidence interval is: ",ci1_99[1],",",ci1_99[2])
```

**ii. 99% CI of Bootstrapped difference between means**

```
## [1] "The confidence interval is:  0.100814463544755 , 1.84345002963841"
```

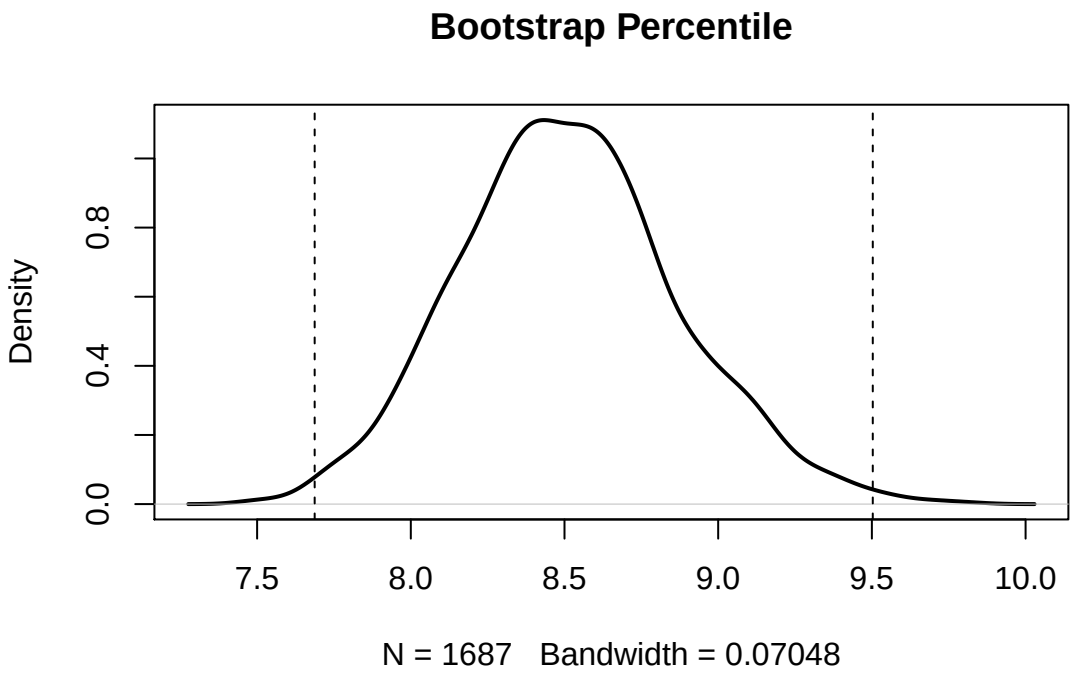
```r
boot_t_stat <- function(sample0, hyp){
  resample <- sample(sample0, length(sample0), replace = TRUE)
  diff <- mean(resample)- hyp
  se <- sd(resample)/sqrt(length(resample))
  return (diff/se)
}
t_boots <- replicate(length(times), boot_t_stat(times,7.6))
t_ci_99 <- quantile(t_boots, probs = c(0.005,0.995))
paste("The confidence interval is: ",t_ci_99[1],",",t_ci_99[2])
```

**iii. 99% CI of Bootstrapped t-intervals**

```
## [1] "The confidence interval is:  0.134146404771181 , 4.52055464838781"
```

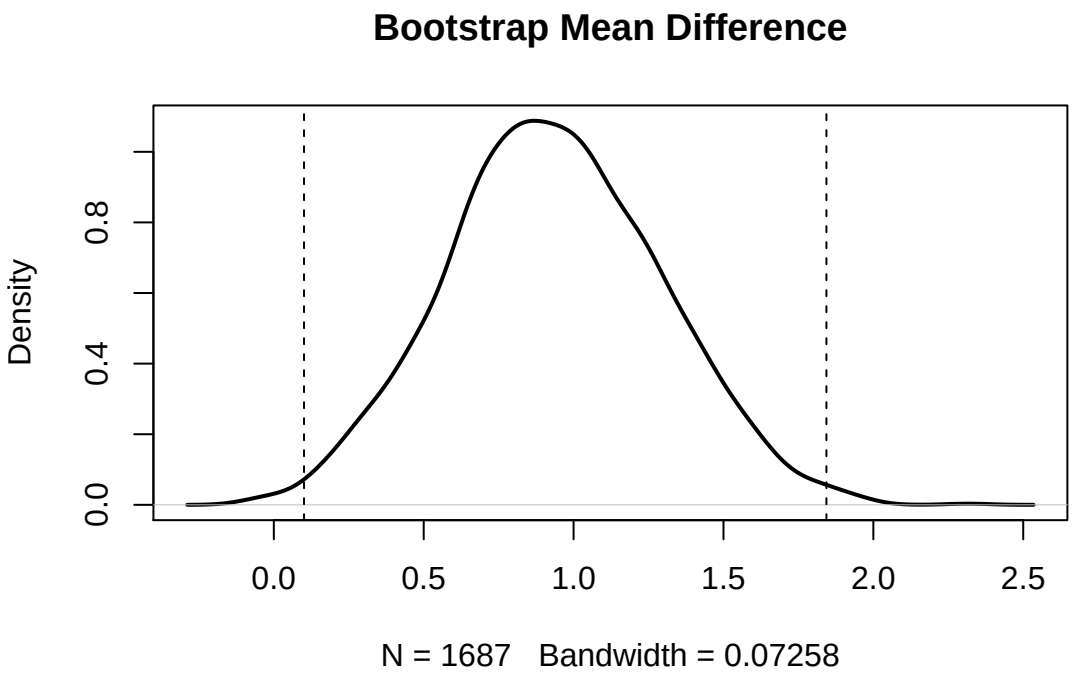
**iv. Plots of above** i.Percentile

```r
plot(density(sample_means),lwd =2, main = "Bootstrap Percentile")
abline(v = ci_99,lty = "dashed")
```



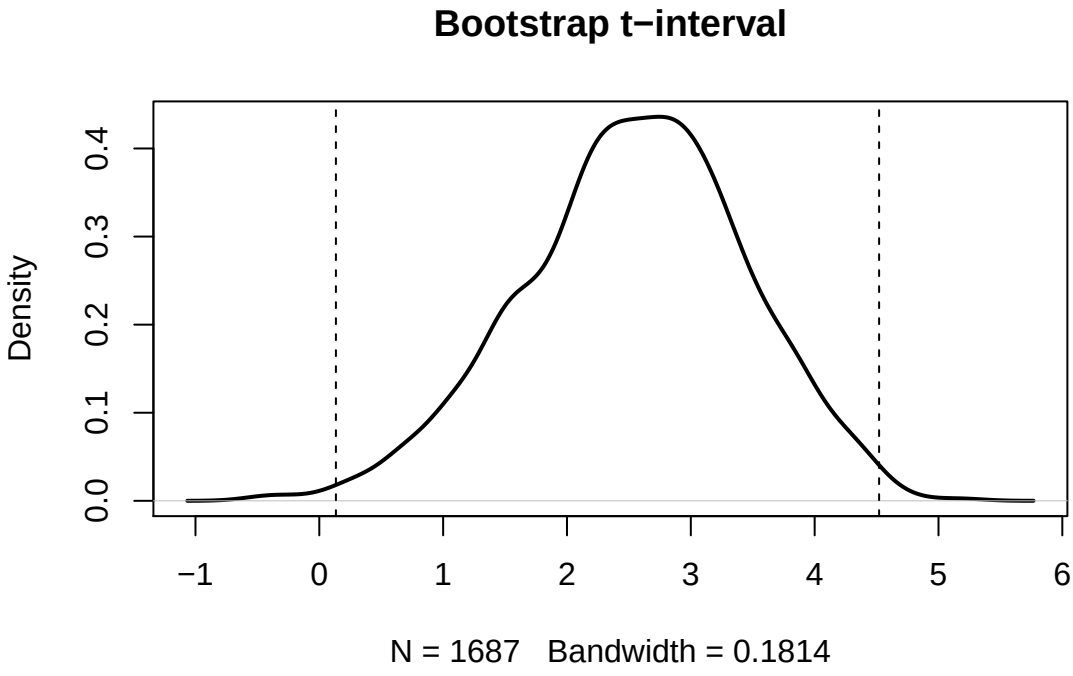

ii. Difference of Means

```r
plot(density(mean_diffs),lwd =2, main = "Bootstrap Mean Difference")
abline(v = ci1_99,lty = "dashed")
```




iii. Bootstrapped t-interval

```
plot(density(t_boots),lwd =2, main = "Bootstrap t-interval")
abline(v = t_ci_99,lty = "dashed")
```



**(c) Do they agree with each other?**

**ANSWER** :Comparing the traditional CI to the bootstrapped percentile, the intervals are extremely close. Moreover, the bootstrapped difference of means confidence interval covered 0, so that it does agree to our test results that we don't have enough evidence to reject the NULL hypothesis. However, the bootstrapped t-interval does not cover 0, but the mean rests around our test results. Basically they all agree with each other.