HW4

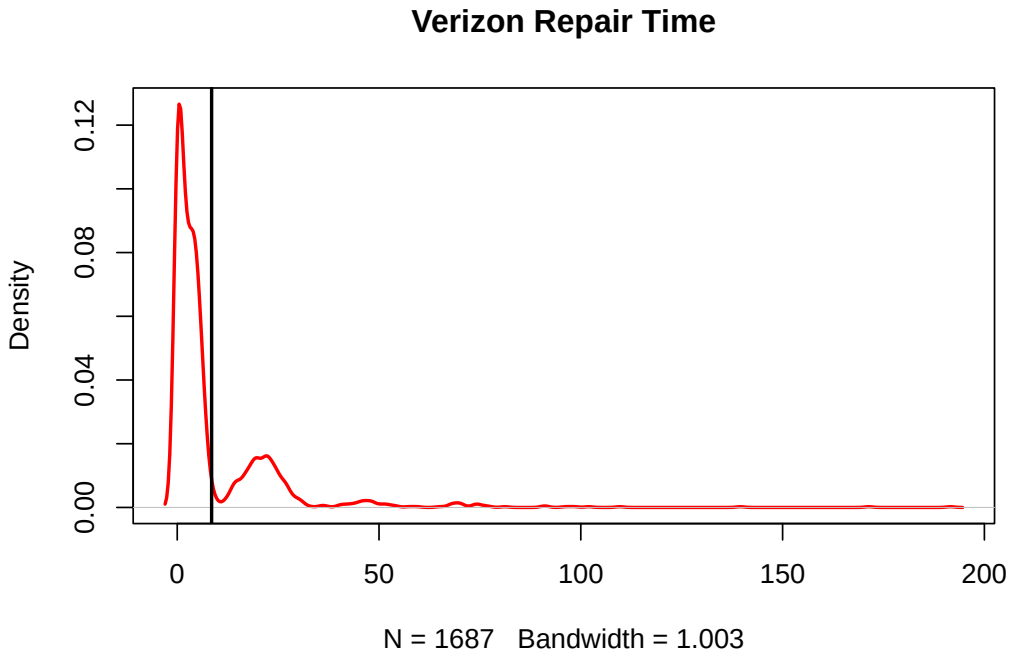
109006206

Question 1

a) Imagine that Verizon claims that they take 7.6 minutes to repair phone services for its customers on average. The PUC seeks to verify this claim at 99% confidence (i.e., significance $\alpha = 1\%$) using traditional statistical methods.

i) Visualize the distribution of Verizon's repair times, marking the mean with a vertical line

```
setwd("/Users/olivia/Documents/Documents/Study/Semester 6/BACS/HW4")
verizon<- read.csv("verizon.csv")
plot(density(verizon$Time),lwd=2,col="red",main="Verizon Repair Time")
abline(v=mean(verizon$Time),lwd = 2)
```



ii) Given what the PUC wishes to test, how would you write the hypothesis? (not graded)

h0 : $\mu \leq 7.6$ minutes h1 : $\mu > 7.6$ minutes

iii) Estimate the population mean, and the 99% confidence interval (CI) of this estimate.

```
population_mean <- mean(verizon$Time)
cat("The Population Mean is :",population_mean)
```

```
## The Population Mean is : 8.522009
```

```
sd_error <- sd(verizon$Time)/(nrow(verizon)^0.5)

ci99_low <- population_mean - (sd_error*2.58)
ci99_high <- population_mean + (sd_error*2.58)

cat("The 99% CI of the population is :", ci99_low, " to ", ci99_high)
```

```
## The 99% CI of the population is : 7.593073  to  9.450946
```

iv) Find the t-statistic and p-value of the test

```
hypothesized_mean <- 7.6
t <- (population_mean-hypothesized_mean)/sd_error

df<- nrow(verizon) - 1
p <- 1 - pt(t,df)
```

v) Briefly describe how these values relate to the Null distribution of t (not graded) Find the t-statistic and p-value of the test

By using t-test we can obtain testing statistic value which are used to calculate p-value. P-value is a value to indicate whether we should reject or not the null hypothesis.

vi) What is your conclusion about the company's claim from t his t-statistic, and why?

Based on the rules , if p value < 0.01 we should reject the null hypothesis.
So in this case, we should reject the null hypothesis.

b) Let's re-examine Verizon's claim that they take no more than 7.6 minutes on average, but this time using bootstrapped testing:

i) Bootstrapped Percentile: Estimate the bootstrapped 99% CI of the population mean

```
set.seed(86472122)
num_boots<-7000
sample_statistic<-function(sample0) {
  resample <-sample(sample0, length(sample0), replace=TRUE)
  mean(resample)
}
sample_means<-replicate(num_boots, sample_statistic(verizon$Time))
quantile(sample_means, probs= c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 7.633418 9.460080
```

ii) Bootstrapped Difference of Means:

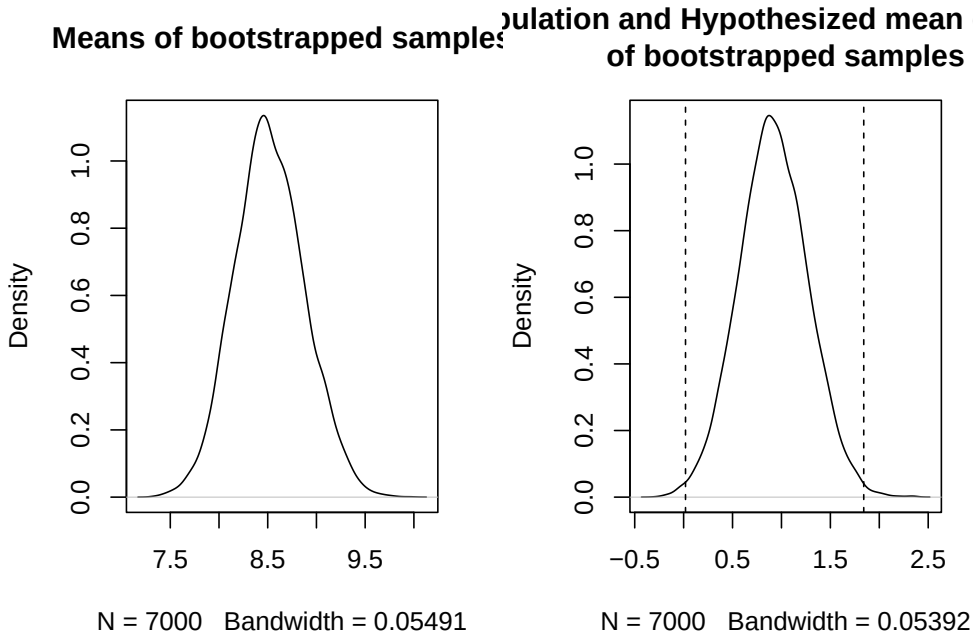
What is the 99% CI of the bootstrapped difference between the sample mean and the hypothesized mean?

```
boot_mean_diffs<-function(sample0, mean_hyp) {
  resample <-sample(sample0, length(sample0), replace=TRUE)
  return( mean(resample)-mean_hyp)
}
mean_diffs<-replicate(num_boots, boot_mean_diffs(verizon$Time, 7.6))
quantile(mean_diffs, probs=c(0.005, 0.995))
```

```
##      0.5%     99.5%
## 0.0193008 1.8410805
```

iii) Plot distribution the two bootstraps above

```
par(mfrow=c(1, 2))
plot(density(sample_means), main="Means of bootstrapped samples")
plot(density(mean_diffs), main="Population and Hypothesized mean difference\nof bootstrapped samples")
abline(v=quantile(mean_diffs, probs=c(0.005, 0.995)), lty="dashed")
```



iv) Does the bootstrapped approach agree with the traditional t-test in part [a]?

Yes, the bootstrapped difference means doesn't shows that there are any 0 difference between bootstrap mean and verizon's claim so we should reject the null hypothesis. The bootstrapped means also shows that the actual t value (2.5607623) is outside the range of the CI (7.6334178, 9.4600801) so we should reject the null hypothesis.

c) Finally, imagine that Verizon notes that the distribution of repair times is highly skewed by outliers, and feel that testing the mean in not fair because the mean is sensitive to outliers. They claim that the median is a more fair test, and claim that the median repair time is no more than 3.5 minutes at 99% confidence (i.e., significance $\alpha = 1\%$).

i) Bootstrapped Percentile: Estimate the bootstrapped 99% CI of the population median

```
sample_medians<-replicate(num_boots, sample_statistic(verizon$Time))
quantile(sample_medians, probs= c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 7.625340 9.484112
```

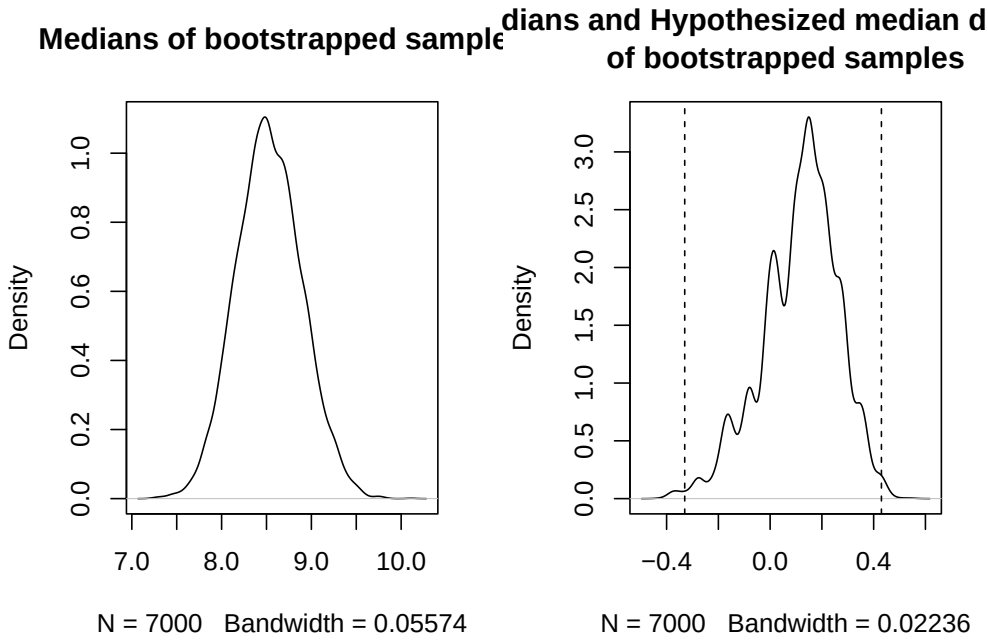
ii) Bootstrapped Difference of Medians:
What is the 99% CI of the bootstrapped difference between the sample median and the hypothesized median?

```
boot_median_diffs<-function(sample0, median_hyp) {
  resample <-sample(sample0, length(sample0), replace=TRUE)
  return(median(resample)-median_hyp)
}
median_diffs<-replicate(num_boots, boot_median_diffs(verizon$Time, 3.5))
quantile(median_diffs, probs=c(0.005, 0.995))
```

```
##  0.5% 99.5%
## -0.33  0.43
```

iii) Plot distribution the two bootstraps above on two separate plots.

```
par(mfrow=c(1, 2))
plot(density(sample_medians), main="Medians of bootstrapped samples")
plot(density(median_diffs), main="Medians and Hypothesized median difference\nof bootstrapped samples")
abline(v=quantile(median_diffs, probs=c(0.005, 0.995)), lty="dashed")
```



iv What is your conclusion about Verizon's claim about the median, and why?

The bootstrapped median also shows that the actual t value (2.5607623)
is outside the range of the CI (7.6253401, 9.4841117)
so we should reject the null hypothesis. The average time should be longer than 7.6 minutes.

Question 2

a) You discover that your colleague wanted to target the general population of Taiwanese users of the product. However, he only collected data from a pool of young consumers, and missed many older customers who you suspect might use the product much less every day.

i) This scenario would create systematic error (bias). This situation means that the sample data is biased towards young consumers and may not accurately reflect the opinions and behaviors of older consumers.

ii) Since the sample data is biased it would effect the variance. Thus, standard deviation(sd) will also affect since the formula of sd itself is the square root of variance. diff would also be affected since diff itself calculated the difference between means of two groups.

iii) This situation is likely to decrease the power.

iv) This situation would most likely to result in a "Type II Error". It would not affect "Type I Error".

b) You find that 20 of the respondents are reporting data from the wrong wearable device, so they should be removed from the data. These 20 people are just like the others in every other respect.

i) This scenario would create random error. This situation means that there are some sample data that affects the measurements in unpredictable ways.

ii) Since this situation will remove some data it would affect the n. The variance would also be affected, thus sd will also be affected in this case. diff would also be affected since diff itself calculated the difference between means of two groups.

iii) This situation is likely to decrease the power since the sample size is reduced.

iv) In this situation the probability of a Type II error will decrease since the sample size is reduced. The probability of Type I Error will remain the same since there is no difference in the alpha.

c) A very annoying professor visiting your company has criticized your colleague's "95% confidence" criteria, and has suggested relaxing it to just 90%.

i) This scenario would create neither error.

ii) This situation will only affect alpha.

iii) The power will be increase. Based on the theory itself if the higher the significance level is it will cause higher power.

iv) This situation would most likely to result in a "Type I Error"

d) Your colleague has measured usage times on five weekdays and taken a daily average. But you feel this will underreport usage for younger people who are very active on weekends, whereas it over-reports usage of older users.

i) This scenario would create systematic error (bias). The error is systematic because the design of the study is biased in a way that consistently underreports the usage times for younger people who are active on weekends, and over-reports the usage times of older users.

ii) Since the sample data is biased it would effect the variance. Thus, standard deviation(sd) will also affect since the formula of sd itself is the square root of variance. diff would also be affected since diff itself calculated the difference between means of two groups.

iii) This situation is likely to decrease the power.

iv) This situation would most likely to result in a "Type II Error". It would not affect "Type I Error".