Student ID: 112077423

```
library(ggplot2)
library(compstatslib)
```

```
data <- read.csv('verizon.csv', header=TRUE)
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
```

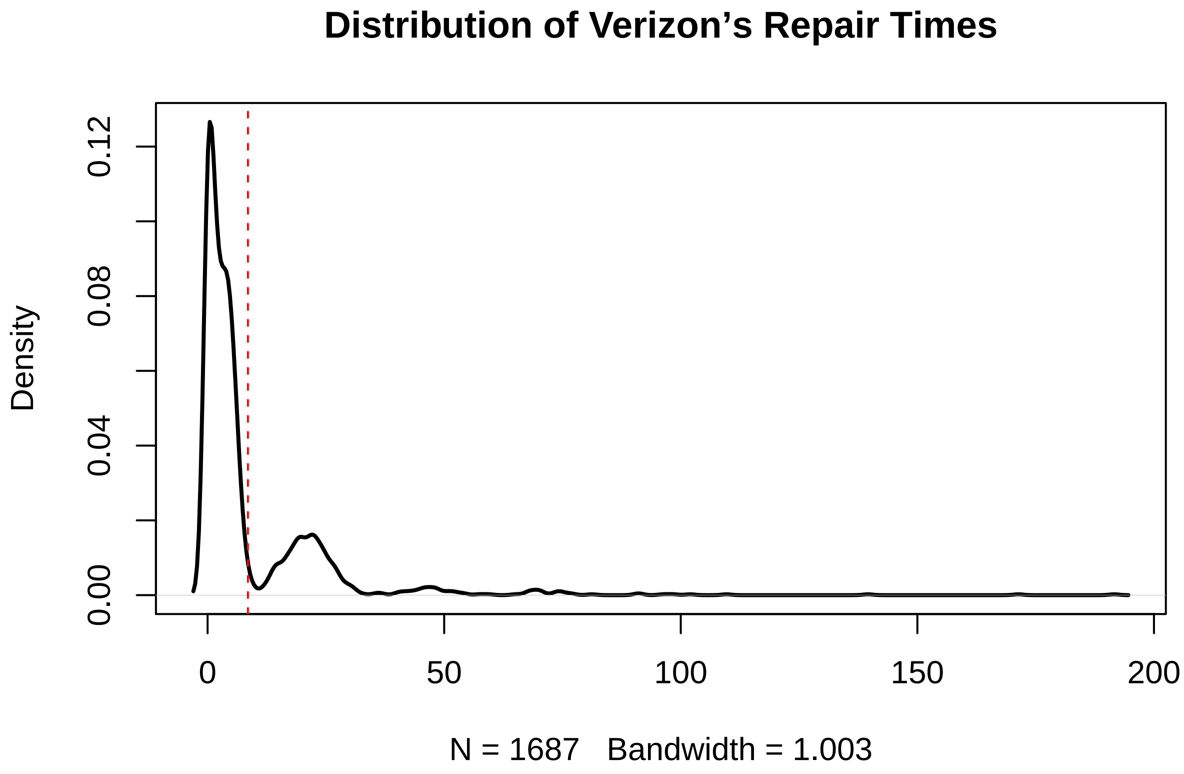
```
# check missing values
out1 <- paste('# of rows with missing values', sum(is.na(data)))
# check if negative values in time
out2 <- paste('# of rows with negative time', sum(data[data$Time<0,]))
cat(out1, out2, sep='\n')
```

```
## # of rows with missing values 0
## # of rows with negative time 0
```

Question 1(a)

(i)

```
plot(density(data$Time), lwd=2,
     main='Distribution of Verizon’s Repair Times')
abline(v=mean(data$Time), col='red', lty='dashed', lw=1)
```



(ii)

Hypothesis:

- H0: u = 7.6 *on average, it takes 7.6 minutes for Verizon to repair phone services for its customers*
- Ha: u != 7.6 *the average repair time is not equal to 7.6 minutes*

(iii)

```
# Estimate the population mean, and the 99% confidence interval (CI) of this estimate
# the sample mean is an estimate of the population mean
mean_ <- round(mean(data$Time), 2)
se <- sd(data$Time)/sqrt(length(data$Time))
CI_99 <- round(mean_ + c(-2.58*se, 2.58*se), 2)

out1 <- paste('estimate of the population mean:', mean_)
out2 <- paste('99% confidence interval:', CI_99[1], '-', CI_99[2])
cat(out1, out2, sep='\n')
```

```
## estimate of the population mean: 8.52
## 99% confidence interval: 7.59 - 9.45
```

(iv)

```r
# Find the t-statistic and p-value of the test
t <- (mean_ - 7.6) / se
df <- length(data$Time) - 1
p <- 1 - pt(t, df)
p <- p * 2
out1 <- paste('p-value:', p)
out2 <- paste('t-statistic:', t)
cat(out1, out2, sep='\n')
```

```
## p-value: 0.0107003776321966
## t-statistic: 2.55518125121811
```

(v)

The sample mean is 2.55 standard errors away from hypothesized mean. The p-value is the probability of obtaining a t-value.

(vi)

Since the p-value is larger than 1%, so we fail to reject the null hypothesis.

Question 1(b)

(i)

```r
# Estimate the bootstrapped 99% CI of the population mean
set.seed(467291)
resamples <- replicate(2000, sample(data$Time, length(data$Time), replace=TRUE))
sample_means <- c()
# calculate mean for each sample
for(i in 1:ncol(resamples)) {
  sample_means[i] <- mean(resamples[,i])
}
CI_99_ <- quantile(sample_means, probs=c(0.005, 0.995))
print(paste('99% CI of the bootstrapped means:', CI_99_[1], '-', CI_99_[2]))
```

```
## [1] "99% CI of the bootstrapped means: 7.62898055720213 - 9.46265204505038"
```

(ii)

```r
boot_mean_diffs <- function(sample0, mean_hyp) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  return(mean(resample) - mean_hyp)
}
```

```r
mean_diffs <- replicate(2000, boot_mean_diffs(data$Time, 7.6))
diff_ci_99 <- quantile(mean_diffs, probs=c(0.005, 0.995))

print(paste('99% CI of the bootstrapped difference:',
            diff_ci_99[1], '-', diff_ci_99[2]))
```

```
## [1] "99% CI of the bootstrapped difference: 0.00389576170717285 - 1.87929967397747"
```

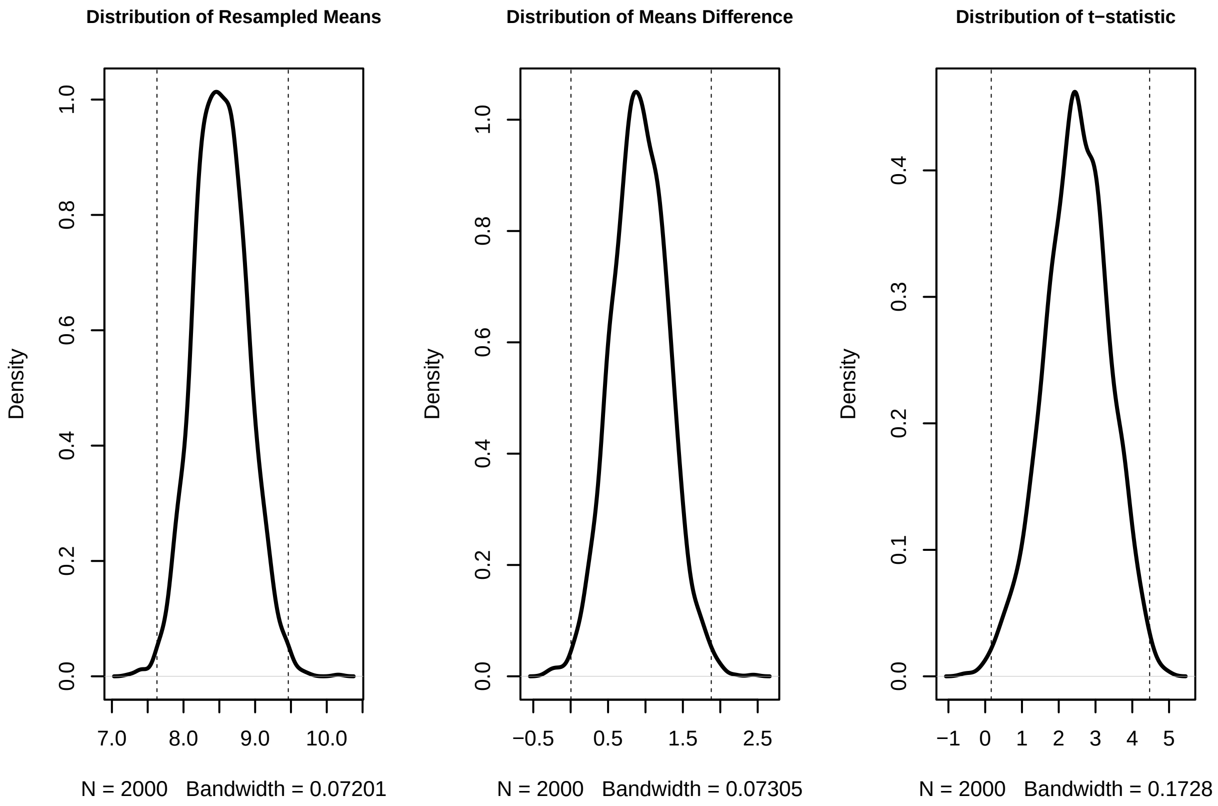
(iii)

```r
boot_t_stat <- function(sample0, mean_hyp) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  diff <- mean(resample) - mean_hyp
  se <- sd(resample)/sqrt(length(resample))
  return(diff/se )
}

t_boots <- replicate(2000, boot_t_stat(data$Time, 7.6))
t_ci_99 <- quantile(t_boots, probs=c(0.005, 0.995))

print(paste('99% CI of the bootstrapped t-statistic:',
            t_ci_99[1], '-', t_ci_99[2]))
```

```
## [1] "99% CI of the bootstrapped t-statistic: 0.164842533527012 - 4.47219520140975"
```

(iv)

```r
par(mfrow=c(1,3))
plot(density(sample_means), lwd=2, main='Distribution of Resampled Means', cex.main=0.9)
abline(v=CI_99_, lty="dashed",lw=0.5)
plot(density(mean_diffs), lwd=2, main='Distribution of Means Difference', cex.main=0.9)
abline(v=diff_ci_99, lty="dashed",lw=0.5)
plot(density(t_boots), lwd=2, main='Distribution of t-statistic', cex.main=0.9)
abline(v=t_ci_99, lty="dashed",lw=0.5)
```



(v)

Yes, the bootstrapped approach agree with the traditional t-test in part (a).

Question 1(c)

(i)

```r
resamples <- replicate(2000, sample(data$Time, length(data$Time), replace=TRUE))
sample_medians <- c()
# calculate mean for each sample
for(i in 1:ncol(resamples)) {
  sample_medians[i] <- median(resamples[,i])
}
CI_99_ <- quantile(sample_medians, probs=c(0.005, 0.995))
print(paste('99% CI of the bootstrapped medians:', CI_99_[1], '-', CI_99_[2]))
```

```
## [1] "99% CI of the bootstrapped medians: 3.17995 - 3.92005"
```

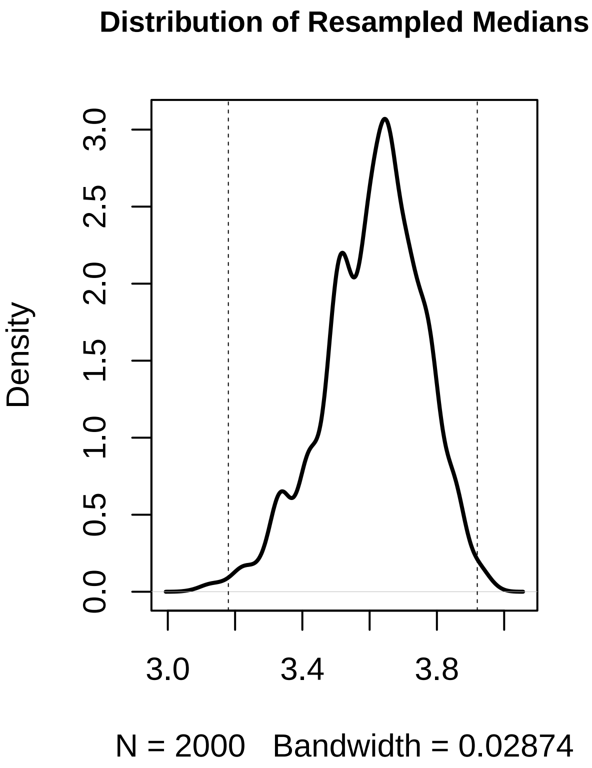
(ii)

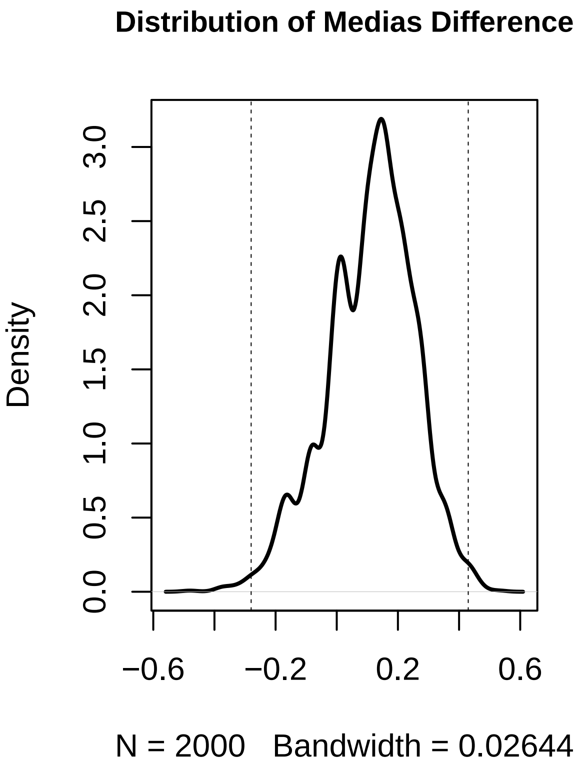
```
boot_median_diffs <- function(sample0, median_hyp) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  return(median(resample) - median_hyp)
}

median_diffs <- replicate(2000, boot_median_diffs(data$Time, 3.5))
diff_ci_99 <- quantile(median_diffs, probs=c(0.005, 0.995))

print(paste('99% CI of the bootstrapped difference:',
            diff_ci_99[1], '-', diff_ci_99[2]))
```

```
## [1] "99% CI of the bootstrapped difference: -0.28 - 0.43"
```

(iii)

```
par(mfrow=c(1,2))
plot(density(sample_medians), lwd=2,
     main='Distribution of Resampled Medians',
     cex.main=0.9)
abline(v=CI_99_, lty="dashed",lw=0.5)
plot(density(median_diffs), lwd=2,
     main='Distribution of Medias Difference',
     cex.main=0.9)
abline(v=diff_ci_99, lty="dashed",lw=0.5)
```





(iv)

Verizon’s claim about the median sounds reasonable because its in 99% CI and median difference range is not too large.

Question 2

(i) Would this scenario create systematic or random error (or both or neither)?

(ii) Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?

(iii) Will it increase or decrease our power to reject the null hypothesis?

(iv) Which kind of error (Type I or Type II) becomes more likely because of this scenario?

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) systematic error;

(ii) diff and sd;

(iii) increase;

(iv) Type II Error (we are more likely to reject H0).

Question 2(b)

You find that 20 of the respondents are reporting data from the wrong wearable device, and should not have been in the sample. These 20 people are just like the others in every other respect.

(i) Both;

(ii) n (since we can't consider 20 users as part of sample);

(iii) decrease (smaller dataset -> decreased sensitivity);

(iv) Type II Error.

Question 2(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) neither;

(ii) alpha;

(iii) increase;

(iv) Type I Error.

Question 2(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) systematic;

(ii) diff, sd;

(iii) decrease;

(iv) Type II Error.