Question 1)

(a) what is the probability that a randomly chosen app from Google's app store will turn off the Verify security feature?

We determine the app is harmful by Z-score < 3.7, so we can use "pnorm" function to get the probability of Z-score < 3.7.

```
> pnorm(-3.7)
[1] 0.0001077997
```

(b) what number of apps on the Play Store did Google expect would maliciously turn off the Verify feature once installed?

The number of malicious apps is number of all apps multiply by the probability we get last question:

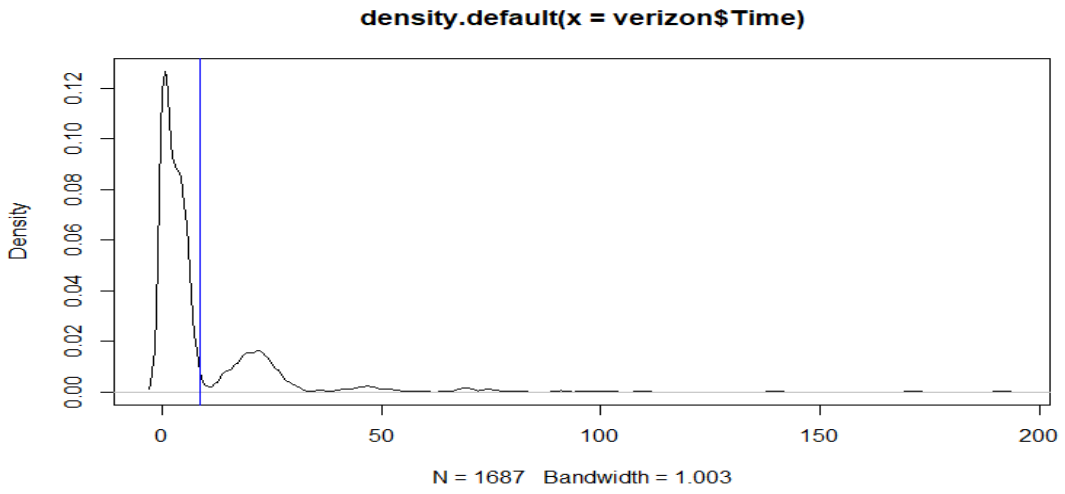
```
> 220000000*pnorm(-3.7)
[1] 23715.94
```

Question 2)

(a) The Null distribution of t-values:

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

```
#i visualize
plot(density(verizon$Time))
abline(v = mean(verizon$Time), col = 'blue')
```



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

The null hypothesis(H0) : the mean of repair time is less than 7.6

The Alternative Hypothesis(H1) : the mean of repair time is larger than 7.6

The level of significance: 0.01

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

Use the sample mean to estimate the population mean, by adding and subtracting 2.58*standard error of sample.

```
#iii esitimate population mean
sample_mean <- mean(verizon$Time)
standard_error <- sd(verizon$Time)/ sqrt(length(verizon$Time))
CI_99 <-
c(sample_mean-2.58*standard_error,sample_mean+2.58*standard_error)
[1] 7.593073 9.450946
```

iv. Using the traditional statistical testing methods we saw in class, find the t-statistic and p-value of the test

The t-statistic and p-value definition:

t-statistic : How many standard errors the sample mean is away from the hypothesized population mean

p-value : Probability of t

```
#iv t-statistic and p-value
claim <- 7.6
t <- (sample_mean-claim)/standard_error
#the distance of mean and hypothesized mean divide by se
> t
[1] 2.560762
df <- length(verizon$Time)-1
#degree of freedom used to calculate probability
p <- 1-pt(t, df)
> p
[1] 0.005265342
```

v. Briefly describe how these values relate to the Null distribution of t (not graded)

The p-value is less than 0.01, it means that the values and the Null distribution of t has significant difference.

vi. What is your conclusion about the advertising claim from this t-statistic, and why?

We found that the advertising claim average(7.6) is included in the 99% CI of mean estimated by sample, but it is very close to the lower bound(7.59).

And the t-statistic and p-value tell us that the sample and claim has large difference, by p-value < 0.01.

So the advertising claim is not true.

(b) use bootstrapping <u>on the sample data</u> to examine this problem:

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

```
set.seed(1266875)
#the bootstrap fcn, returns the resample mean
boot_mean <- function(data){
  resample <- sample(data, length(data), replace = TRUE)
  return(mean(resample))
}
#bootstrap 2000 times
mean_boots <- replicate(2000, boot_mean(verizon$Time))
#the middle 99% of bootstrapped mean represented the 99% CI
quantile(mean_boots, c(0.005, 0.995))
    0.5%    99.5%
7.615115 9.483419
```

ii. Bootstrapped Difference of Means: What is the 99% CI of the bootstrapped difference between the population mean and the hypothesized mean?

```
set.seed(4564258)
#the bootstrap fcn, returns the resample mean - claim
boot_mean_diffs <- function(data, hyp){
  resample <- sample(data, length(data), replace = TRUE)
  return(mean(resample)-hyp)
}
#bootstrap 2000 times
mean_diffs <- replicate(2000, boot_mean_diffs(verizon$Time,
claim))
#the middle 99% of bootstrapped mean represented the 99% CI
quantile(mean_diffs, c(0.005, 0.995))
      0.5%      99.5%
0.03660006 1.93486852
```

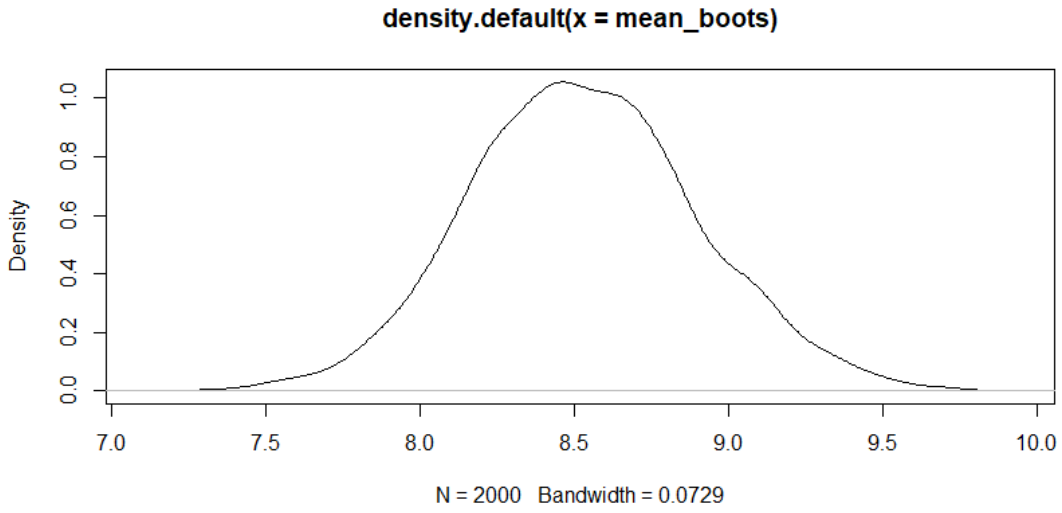
iii. Bootstrapped t-Interval: What is 99% CI of the bootstrapped t-statistic?

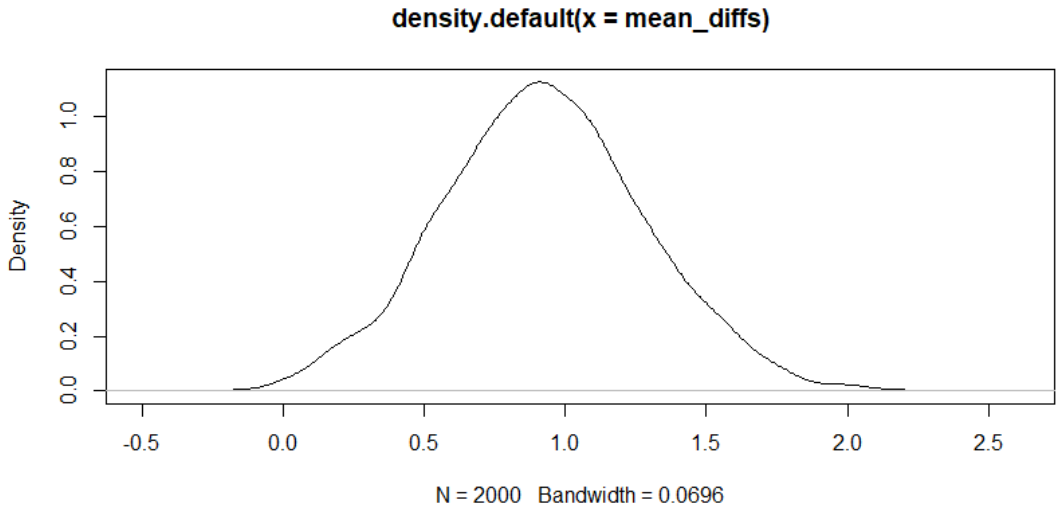
```
set.seed(2346786)
#the bootstrap fcn, returns the resample t-statistic
boot_t_stat <- function(data, hyp){
  resample <- sample(data, length(data), replace = TRUE)
  diff <- mean(resample)-hyp
  se <- sd(resample)/sqrt(length(resample))
  return(diff/se)
}
#bootstrap 2000 times
t_boots <- replicate(2000, boot_t_stat(verizon$Time, claim))
#the middle 99% of bootstrapped mean represented the 99% CI
quantile(t_boots, c(0.005, 0.995))
     0.5%     99.5%
0.2434266 4.6637516
```

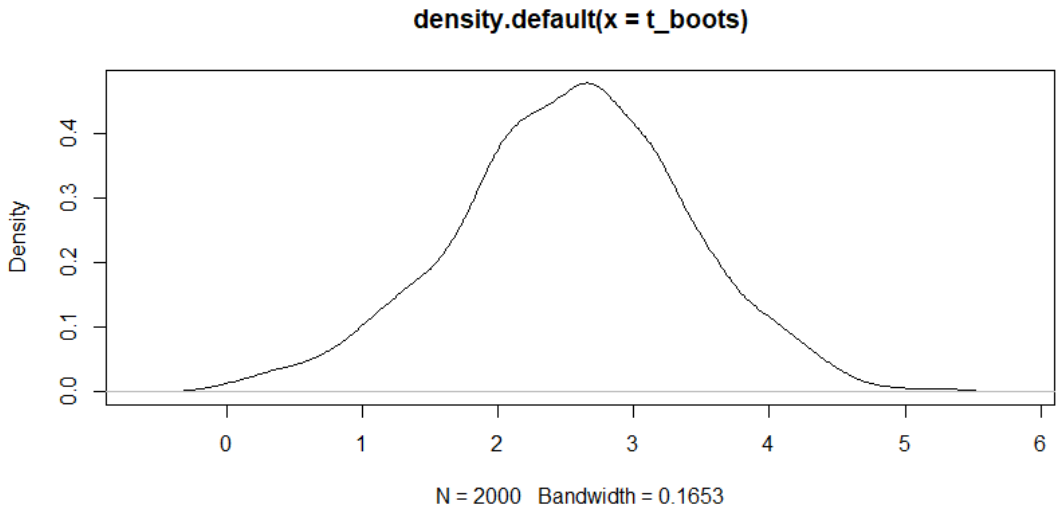
iv. Plot separate distributions of all three bootstraps above

(for ii and iii make sure to include zero on the x-axis)

```
#plot the density plot for three bootstrap result
plot(density(mean_boots))
plot(density(mean_diffs))
plot(density(t_boots))
```







(C) Do the four methodsagree with each other on the test?

traditional test: the p-value < 0.01

bootstrapped percentile: the 99% CI not included claim

bootstrapped difference of means: the 99% CI not included zero.

bootstrapped t-Interval : the 99% CI not included zero.

The four test method does support each other.