HW4

110077443

3/7/2022

Please note that all code in this document is presented in a grey box and the output reflected below each box

- The below code allows lengthy lines of code to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

1) Spotting Malcious Apps:

a) The probability that a randomly chosen app from Google's app store will turn off the Verify security feature.

- Google refers to the DOI score as a Z-score.
- Thus, given the z-score, we can use the pnorm function in r to find the probability

```
pnorm(-3.7)
```

```
## [1] 0.0001077997
```

- The probability is 0.0001077997, which is less than 0.1% and a very low retention rate.

b) Number of apps on the Play Store Google expected to maliciously turn off the Verify feature.

```
2200000 * pnorm(-3.7)  # 237.1594 number of apps
```

```
## [1] 237.1594
```

```
round(2200000 * pnorm(-3.7))  # rounding the number of apps
```

```
## [1] 237
```

- There are about **237** apps on the Play Store Google expected to maliciously turn off the Verify feature.

2) Verify the claim that Verizon takes 7.6 minutes to repair phone services for its customers on average:

- Hypothesized mean claim:

```r
verizon_claim <- 7.6
```

- Import the data for our sample:

```r
verizon <- read.csv("verizon.csv", header = TRUE)
str(verizon)  # Checking structure for possible formatting
```

```
## 'data.frame':    1687 obs. of  2 variables:
##  $ Time : num  17.5 2.4 0 0.65 22.23 ...
##  $ Group: chr  "ILEC" "ILEC" "ILEC" "ILEC" ...
```

```r
table(verizon$Group)  # Checking how many observations in each 'Group'
```

```
##
## CLEC ILEC
##   23 1664
```

```r
verizon_sample <- verizon$Time  # Removing 'Group' variable since we only need time
sample_size <- length(verizon_sample)  # 1687
sample_mean <- mean(verizon_sample)  # 8.522009
sample_sd <- sd(verizon_sample)  # 14.78848
```

a) The Null distribution of t-values:

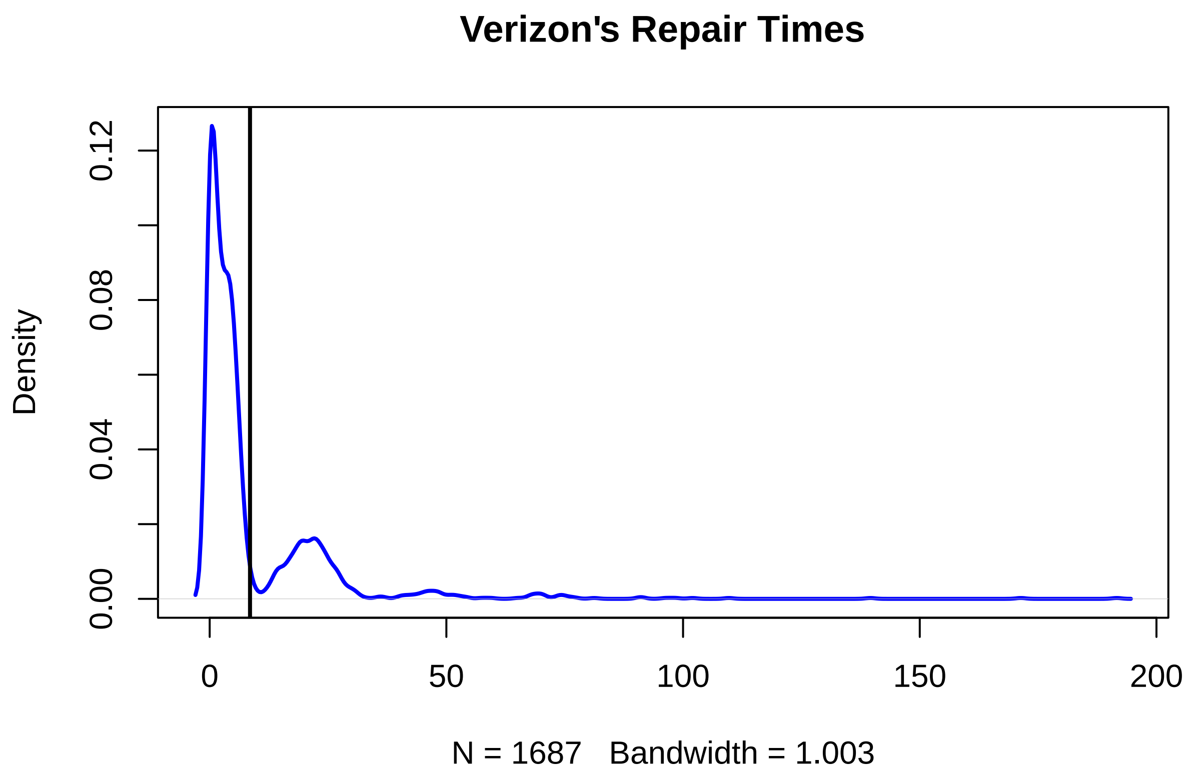
- Running a quick t-test on claim before breaking it down:

```r
t.test(verizon_sample, conf.level = 0.99, alternative = "two.sided",
    mu = 7.6)
```

```
##
##  One Sample t-test
##
## data:  verizon_sample
## t = 2.5608, df = 1686, p-value = 0.01053
## alternative hypothesis: true mean is not equal to 7.6
## 99 percent confidence interval:
##  7.593524 9.450495
## sample estimates:
## mean of x
##  8.522009
```

i) Visualize the distribution of Verizon's repair times and marking the mean

```r
# Plot Verizon's repair times
plot(density(verizon_sample), col = "blue", lwd = 2, main = "Verizon's Repair Times")

# Plot the mean with a vertical line
abline(v = mean(verizon_sample), lwd = 2)
```



ii) PUC Hypothesis (two-tailed test)

- H0: mu - 7.6 = 0
- Ha: mu - 7.6 != 0

iii) Estimate the population mean, and the 99% CI of this estimate

```r
# Population mean
sample_mean <- mean(verizon_sample)  # 8.522009
sample_mean  # Print
```

```
## [1] 8.522009
```

```r
# Compute Standard Error
sample_se <- sample_sd/(sqrt(sample_size))  # 0.3600527
sample_se  # Print
```

```
## [1] 0.3600527
```

```
# Compute 99% confidence interval for this estimate
verizon_ci99 <- sample_mean + c(-2.576, 2.576) * sample_se  #99% CI
verizon_ci99  # Print
```

```
## [1] 7.594514 9.449505
```

- The estimated population mean is **8.522009**, and we are 99% confident that this estimate is between **7.594514 and 9.449505**

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

```
# t-statistic
t_stat <- (sample_mean - verizon_claim)/sample_se  # 2.560762
t_stat  # Print
```

```
## [1] 2.560762
```

```
# p-value
df <- sample_size - 1  # Degrees of freedom
p_value <- pt(t_stat, df, lower.tail = FALSE) * 2  # 0.01053068
p_value  # Print
```

```
## [1] 0.01053068
```

v) Briefly describe how these values relate to the Null distribution of t

- T-statistic: Gives us the standardized difference our sample mean is away from the hypothesized mean
- P-value: Tells us how likely the data observed is to have occurred under the null hypothesis

vi) Conclusion about the advertising claim from this t-statistic, and why.

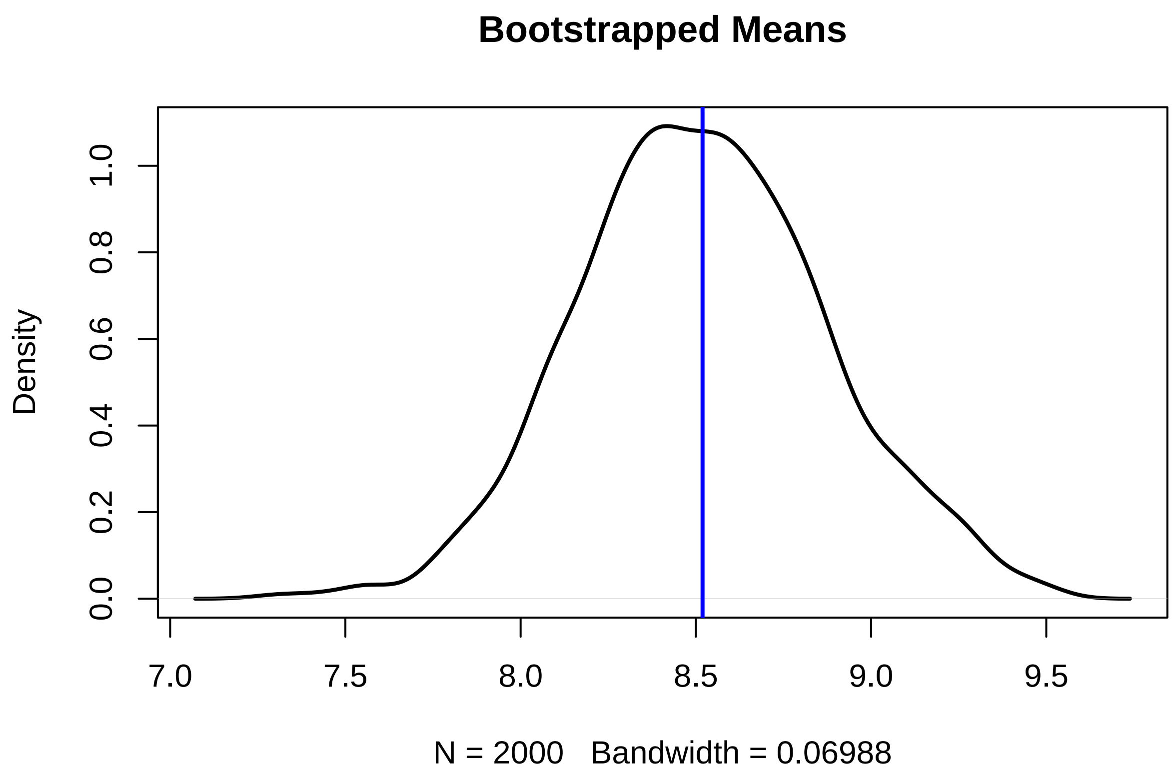
- The advertising claim may be correct based on the t-statistic.
- The two-sided test with a significance level of 0.01 revealed the p-value of **0.01053068** is slightly greater than the significance level. Thus, **we cannot reject the NULL hypothesis**.

b) Bootstrapping the sample data to examine this problem:

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

```
# Set seed
set.seed(3893)  # For reproducibility

# Let's bootstrap
num_boots <- 2000  # Number of bootstrap samples
verizon_sample <- verizon$Time  # Variable we will resample from
```

```r
# Bootstrap function
sample_statistic <- function(stat_function, sample0) {
    resample <- sample(sample0, length(sample0), replace = TRUE)
    stat_function(resample)
}

# Bootstrapped means
boot_means <- replicate(num_boots, sample_statistic(mean, verizon_sample))
plot(density(boot_means), lwd = 2, main = "Bootstrapped Means")  # Visualize

# Bootstrapped estimated mean
boot_estimated_mean <- mean(boot_means)  #8.519121
abline(v = mean(boot_estimated_mean), lwd = 2, col = "blue")
```




```r
boot_estimated_mean  # Print
```

```
## [1] 8.519121
```

```r
# 99% CI of estimated mean
boot_mean99 <- quantile(boot_means, probs = c(0.005, 0.995))  # 99% CI
boot_mean99
```

```
##     0.5%    99.5%
## 7.564853 9.413598
```

- We are 99% confident that the mean is between **7.564853 and 9.413598.**

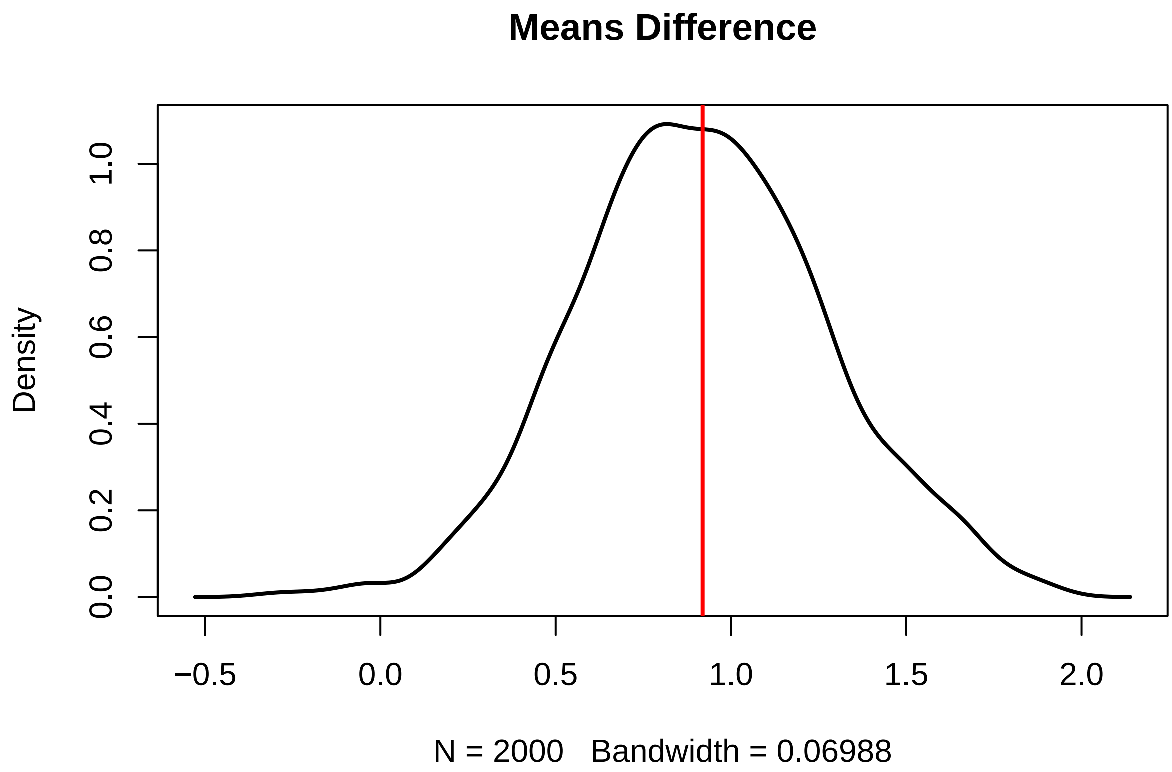
ii) Bootstrapped Difference of Means

- The 99% CI of the bootstrapped difference between the population mean and the hypothesized mean:

```
set.seed(3893)  # For reproducibility

verizon_claim <- 7.6  # Hypothesized mean
mean(verizon_sample)  # 8.522009
```

```
## [1] 8.522009
```

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

# Bootstrapping
mean_diffs <- replicate(num_boots, boot_mean_diffs(verizon_sample,
    verizon_claim))
plot(density(mean_diffs), lwd = 2, main = "Means Difference")  # Visualize

# Bootstrap mean difference
boot_mean_diff <- mean(mean_diffs)  # 0.920083
abline(v = mean(mean_diffs), lwd = 2, col = "red")
```




```
boot_mean_diff  # Print
```

```
## [1] 0.9191212
```

```
# 99% CI of estimated mean difference
boot_means_diff99 <- quantile(mean_diffs, probs = c(0.005, 0.995))  # 99% CI
boot_means_diff99
```

```
##        0.5%       99.5%
## -0.03514713  1.81359825
```

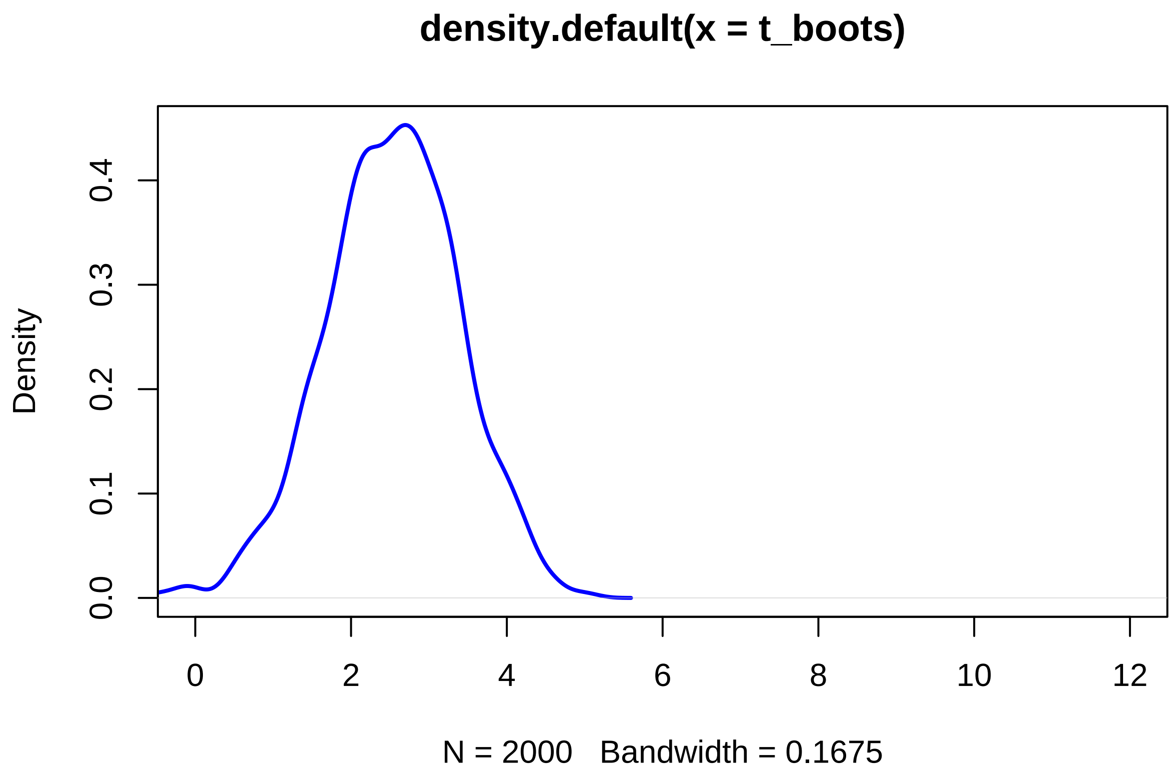
- We are 99% confident that the mean difference is between **-0.03514713 and 1.81359825.**

iii) Bootrsapped t-interval

- is 99% CI of the bootstrapped t-statistic

```
set.seed(3893)  # For reproducibility

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))
```

```r
    return(diff/se)
}

# Bootstrap standardized difference
t_boots <- replicate(num_boots, boot_t_stat(verizon_sample, verizon_claim))
plot(density(t_boots), xlim = c(0, 12), col = "blue", lwd = 2)
```




```r
# mean
boot_mean_t <- mean(t_boots)
boot_mean_t  # Print
```

```
## [1] 2.522641
```

```r
# 99% CI of bootstrapped t.
t_stat99 <- quantile(t_boots, probs = c(0.005, 0.995))  # 99% CI
t_stat99  # Print
```

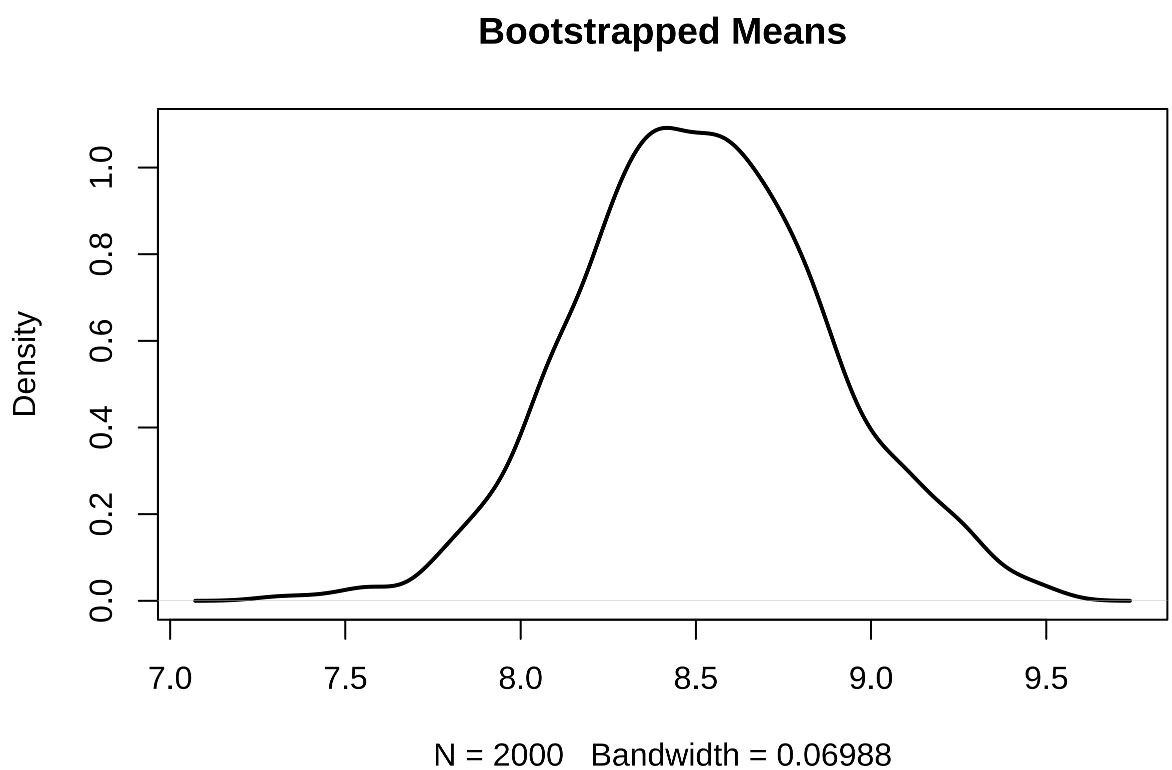
```
##       0.5%      99.5%
## -0.1196324  4.5730370
```

- We are 99% confident that the bootstrapped t-interval is between **-0.1196324 and 4.5730370.**

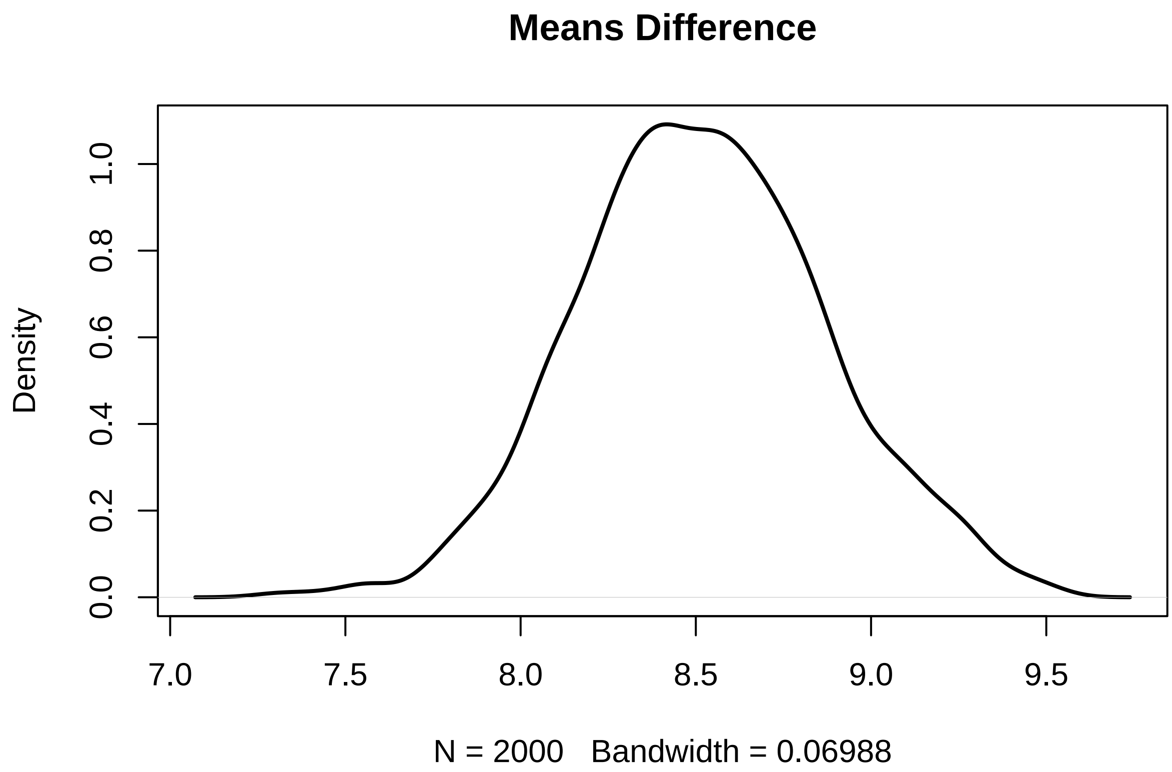
iv) Plot separate distributions of all three bootstraps above

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

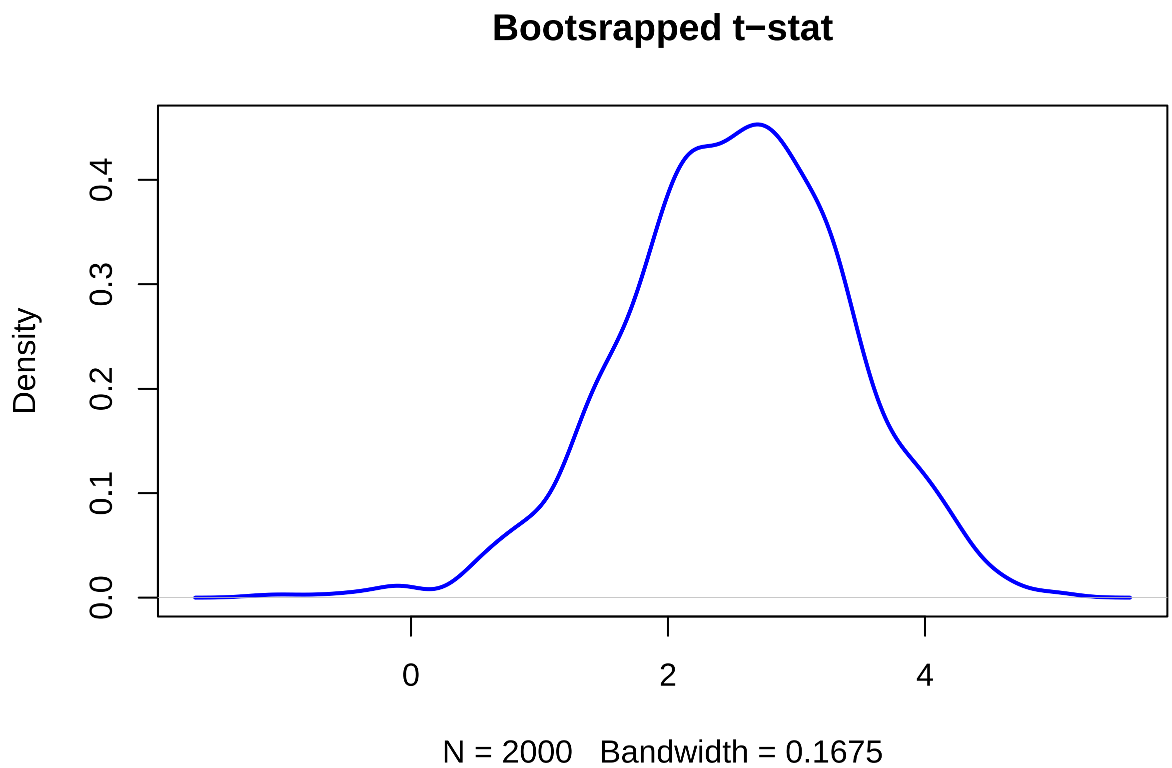
```
# Bootstrapped means:
plot(density(boot_means), lwd = 2, main = "Bootstrapped Means")  # Visualize
```




```
# Means Difference:
plot(density(boot_means), lwd = 2, main = "Means Difference")  # Visualize
```




```r
# Bootrapped t-stat:
plot(density(t_boots), col = "blue", lwd = 2, "Bootsrapped t-stat")  # Visualize
```



c) Do the four methods agree with each other on the test?

- The four different methods agree because all the resulting statisics fall within the 99% confidence interval. Thus, all four different methods **do not reject the hypothesis.**