BACS_HW_Week5

106071041

2021/3/27

Question 1) DOI App

a. Given the critical DOI score that Google uses to detect malicious apps (-3.7), what is the probability that a randomly chosen app from Google's app store will turn off the Verify security feature? (report a precise decimal fraction, not a percentage)

```
DOI_decimal <- pnorm(-3.7)
```

```
DOI_decimal
```

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

b. Assuming there were ~2.2 million apps when the article was written, what number of apps on the Play Store did Google expect would maliciously turn off the Verify feature once installed?

```
2.2 * (10**6) * DOI_decimal
```

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

Question 2) Verizon's repair time

a. The Null distribution of t-values:

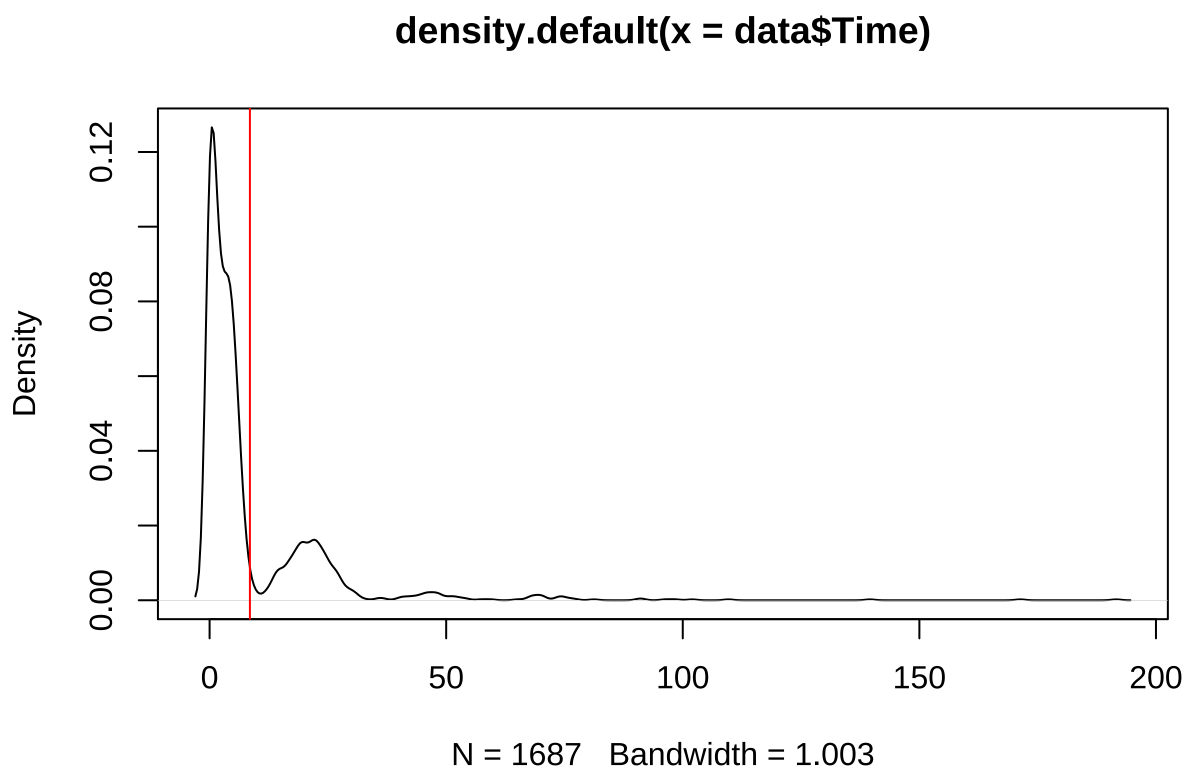
(i) Visualize the distribution of Verizon's repair times, marking the mean with a vertical line

```
data <- read.csv("verizon.csv", header = T)
```

```
verizon_dis <- plot(density(data$Time))
mean(data$Time)
```

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

```
abline(v=mean(data$Time),col="red")
```



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

Null Hypothesis: population mean <= Verizon's claim (7.6)
Alternative Hypothesis: population mean > Verizon's claim

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

```
# estimated population mean
mean(data$Time)
```

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

```
# 99%CI
quantile(data$Time, c(0.005, 0.995))
```

```
##    0.5%   99.5%
##  0.0000 86.8103
```

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

```r
# standard error of the mean
std_error <- sd(data$Time)/sqrt(length(data))
```

```r
# t score
t_score <- (mean(data$Time)-7.6)/std_error
```

```r
t_score
```

```
## [1] 0.0881712
```

```r
# p-value
1 - pt(t_score, df=length(data)-1)
```

```
## [1] 0.4720066
```

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

In right-tailed test with 99% confidence, if p-value less than 0.01, we will reject the hypothesis.

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

PUC can believe Verizon's claim.
Since p-value is obviously larger than 0.01, that means 7.6 is in 99%CI so we accept the null hypothesis.

b. Let's use bootstrapping on the sample data to examine this problem:

```r
# Verizon's Claim
Verizon_hyp <- 7.6
```

```r
# PUC validation sample
set.seed(72342)
PUC_sample <- sample(data$Time, 200)
PUC_mean <- mean(PUC_sample)
PUC_sd <- sd(PUC_sample)
```

```r
PUC_sample
```

```
##   [1]   0.00   2.15 171.35  27.05  24.58   0.65   3.95   5.72   7.47  30.42
##  [11]   4.33  17.80   1.28  45.57  19.68  24.97   0.00   1.97   0.05   6.55
##  [21]   1.98   5.92   0.00   3.35  14.92   3.92   0.47   4.63   3.60   5.00
##  [31]   0.00   2.20   0.00   3.87   3.22   0.80   1.62   5.35  22.83  14.70
##  [41]  24.18   0.50   1.22   4.38  22.32   2.38   4.32   2.27   2.37   3.25
##  [51]   5.35  47.27   0.73  16.40   3.93   6.00   0.10  18.57  12.28   3.98
##  [61]   4.13   0.03   0.32   0.05  22.25   3.13  24.03   2.43   0.95   1.62
```

```
##  [71] 17.50  6.45  0.02 22.18  2.62  0.00  2.20  2.32  2.47  6.15
##  [81]  2.40  0.03  0.00 26.50 21.38  5.30 24.63  0.00  0.00  0.00
##  [91] 26.03  4.83  0.00  2.57  2.98  4.33  0.00 27.65  2.20  0.00
## [101]  5.00  5.05  0.00  3.98  0.00  3.00  5.20  4.87 14.33  1.70
## [111]  0.43  1.18  3.65 23.92  2.00  0.60  2.82 23.25  5.83  1.67
## [121]  4.58  0.00  1.88  4.05 21.77 26.38  1.65 21.25  0.00  4.97
## [131]  0.00  3.22  0.22  0.62  3.97  4.33 24.37 23.02  4.97  3.02
## [141]  3.00  4.32  1.65  6.37  1.37  3.90  2.62 26.70  0.00 20.90
## [151]  1.00  1.62 29.70  0.00 25.18  1.17  0.68  6.35  0.28  0.00
## [161] 21.38  4.95  2.13  4.17  0.17  4.80  0.00  0.00  5.18 27.60
## [171]  0.00 25.63  0.07  4.38  0.02  0.25  3.50  2.57  1.18  1.13
## [181]  4.43  5.00  0.00 17.97  4.18  2.37  2.22  5.58  2.70  0.53
## [191]  1.98 22.08  0.00  1.38  4.22  2.98  5.37  2.02 13.00  6.48
```

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

```r
# bootstrapping samples just like PUC did for 2000 times
boot_samples <- replicate(2000, sample(data$Time, 200))
```

```r
# For all 2000 samples, we find out their means
boot_means <- apply(boot_samples, 2, mean)
```

```r
quantile(boot_means, c(0.005, 0.995))
```

```
##      0.5%     99.5%
##  6.289891 11.296957
```

(ii) Bootstrapped Difference of Means

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

```r
set.seed(432342327)
num_boots <- 2000
mean_diffs <- replicate(num_boots, boot_mean_diffs(PUC_sample, Verizon_hyp))
```

```r
diff_ci_99 <- quantile(mean_diffs, probs=c(0.005, 0.995))
```

```r
diff_ci_99
```

```
##      0.5%     99.5%
## -2.070225  3.355324
```

(iii) Bootstrapped t-Interval:

What is 99% CI of the bootstrapped t-statistic?

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

```
set.seed(23123)
num_boots <- 2000
t_boots <- replicate(num_boots, boot_t_stat(PUC_sample, Verizon_hyp))
```

```
quantile(t_boots, c(0.005, 0.995))
```

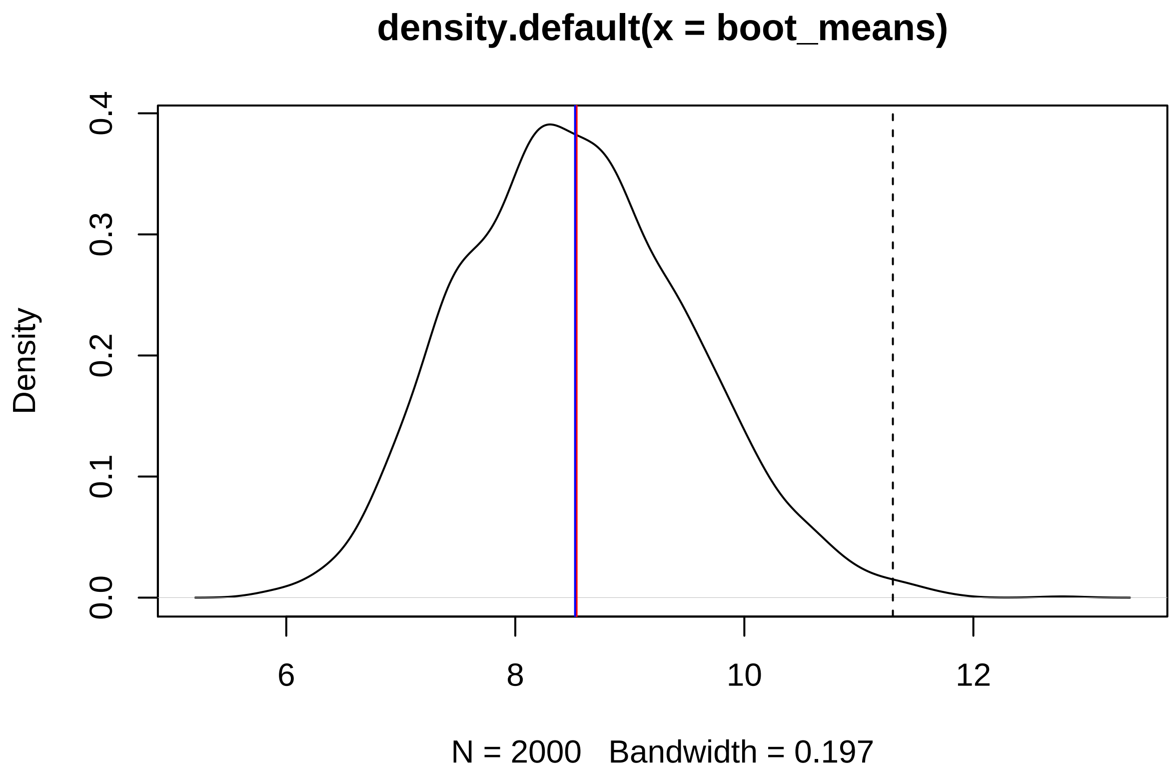
```
##      0.5%     99.5%
## -3.561816  1.970135
```

(iv) Plot separate distributions of all three bootstraps above

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

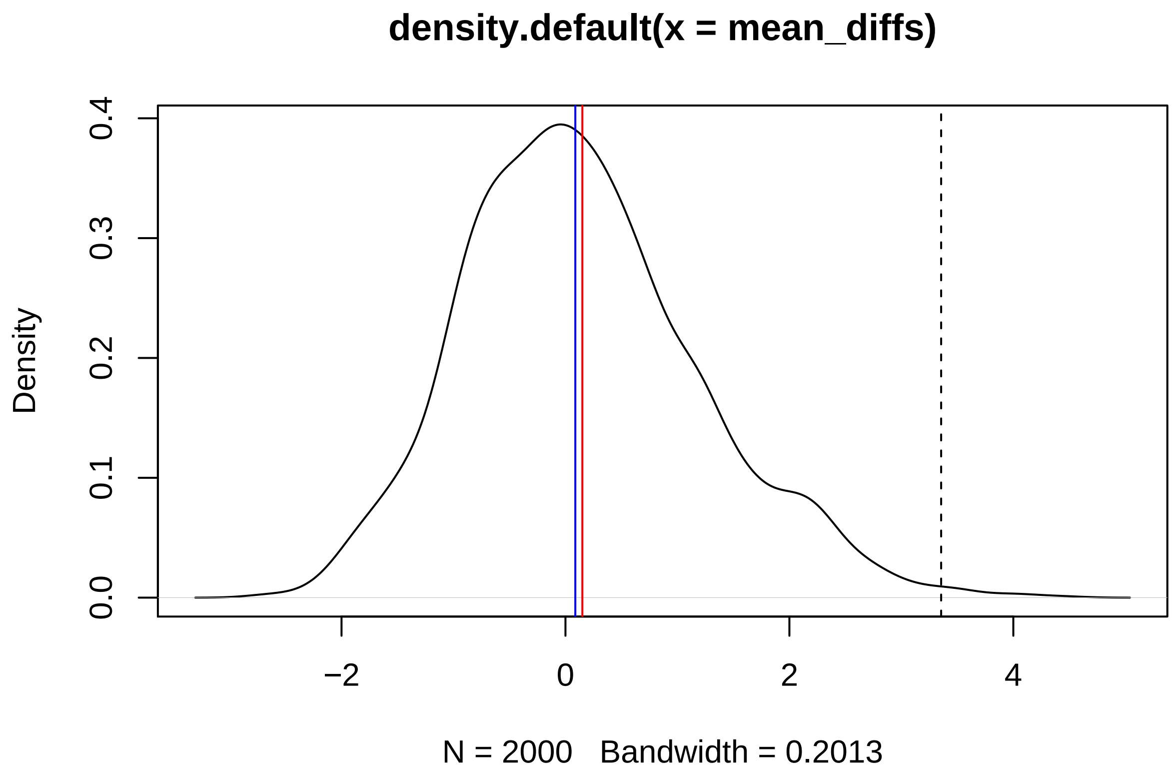
```
# for (i) means
plot(density(boot_means))
abline(v=mean(boot_means), col="red")
abline(v=quantile(boot_means, 0.995), lty="dashed")

# sample0
abline(v=mean(data$Time), col="blue")
```



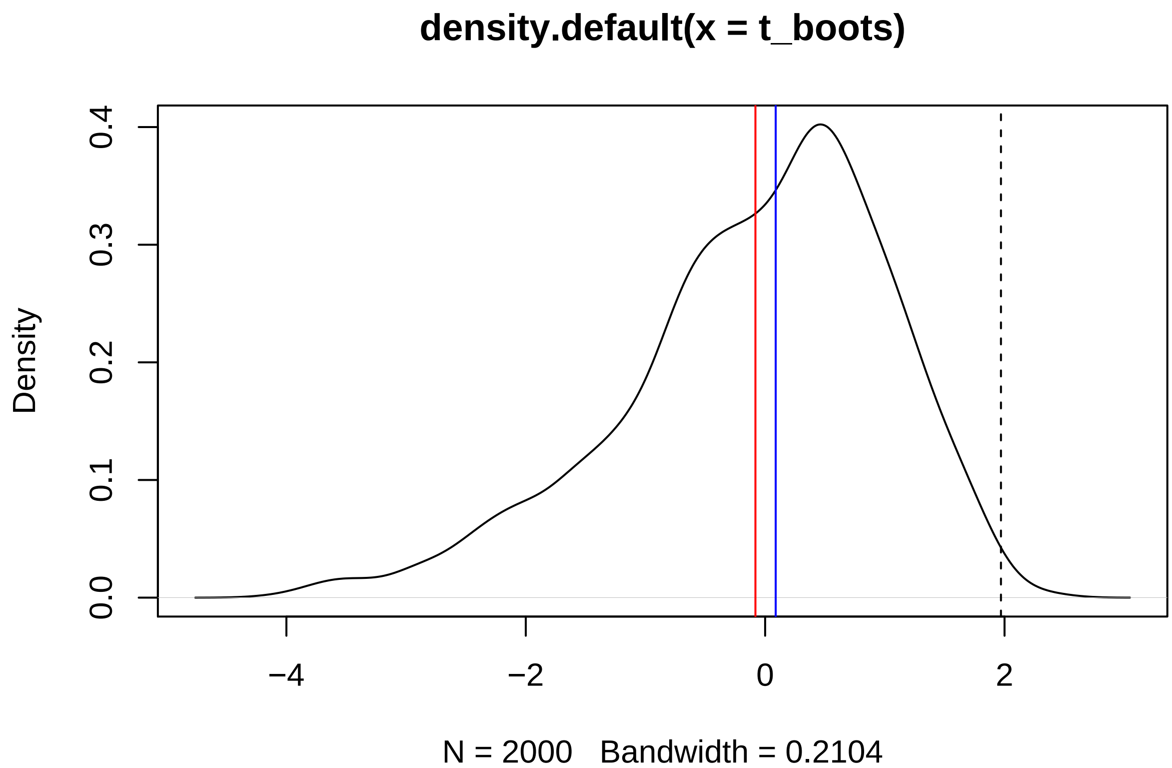

```r
# for (ii) the difference of means
plot(density(mean_diffs))
abline(v=mean(mean_diffs), col="red")
abline(v=quantile(mean_diffs, 0.995), lty="dashed")

# sample0
abline(v=t_score, col="blue")
```




```r
# for (iii) t-intervals
plot(density(t_boots))
abline(v=mean(t_boots), col="red")
abline(v=quantile(t_boots, 0.995), lty="dashed")

# sample0
abline(v=t_score, col="blue")
```



c. Do the four methods (traditional test, bootstrapped percentile, bootstrapped difference of means, bootstrapped t-Interval) agree with each other on the test?

They all agree with each other.