Business Analytics using Statistical Modeling
Assignment 4

Question 1

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.)

```
pnorm(-3.7)
```

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

b) Assuming there are approximately 2.5 million apps on the appstore today, what number of apps on the Play Store does Google expect will maliciously turn off the Verify feature once installed?

```
pnorm(-3.7) * 2500000
```

```
## [1] 269.4993
```

Question 2

a) Use traditional statistical methods, on the statistics provided in the description, to set up a hypothesis:

i) How would you write your hypothesis?

H_0: $\mu = 90{,}000$ km vs. H_a: $\mu \neq 90{,}000$ km

ii) Estimate the population mean, and the 95% confidence interval (CI) of this estimate

```
sample_mean <- 85945.29
standard_deviation <- 14996.55
sample_size <- 360

standard_error <- standard_deviation / sqrt(sample_size)
standard_error
```

```
## [1] 790.3876
```

```
CI_95 <- c(sample_mean + c(-1.96, 1.96) * standard_error)
CI_95
```

```
## [1] 84396.13 87494.45
```

iii) What is the t-statistic of your test?

```
hyp_mean <- 90000
t <- (sample_mean - hyp_mean) / standard_error
t
```

```
## [1] -5.130027
```

iv) What is your conclusion about the advertising claim for this t-statistic, and why?

Since the t-statistic is less than -1.96, we can reject the null hypothesis, which claims that the tires can run an average of 90,000 km before needing to be replaced.

b) Let's use bootstrapping on the sample data itself (see `tires.csv` file)

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.2.5
```

```
tires <- fread('4-tires.csv')
```

i) Estimate the population mean, and the bootstrapped 95% CI of this estimate

```
num_bootstraps <- 2000
compute_sample_mean <- function(sample0){
  resample <- sample(sample0, length(sample0), replace = TRUE)
  return(mean(resample))
}

sample_means <- replicate(num_bootstraps, compute_sample_mean(tires$lifetime_km))
mean(sample_means)
```

```
## [1] 85931.75
```

```
sample_means_ci_95 <- quantile(sample_means, probs = c(0.025, 0.975))
sample_means_ci_95
```

```
##     2.5%    97.5%
## 84402.40 87533.93
```

ii) Bootstrap the difference between the population mean and the hypothesized mean: what is the mean bootstrapped difference, and its 95% CI?

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

mean_diffs <- replicate(num_bootstraps, boot_mean_diffs(tires$lifetime_km, hyp_mean))
mean(mean_diffs)
```

```
## [1] 4072.992
```

```
mean_diffs_ci_95 <- quantile(mean_diffs, probs = c(0.025, 0.975))
mean_diffs_ci_95
```

```
##     2.5%    97.5%
## 2479.385 5618.945
```

iii) Bootstrap the t-statistic: what is the mean bootstrapped t-statistic and its 95% CI?

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

t_boots <- replicate(num_bootstraps, boot_t_stat(tires$lifetime_km, hyp_mean))
mean(t_boots)
```

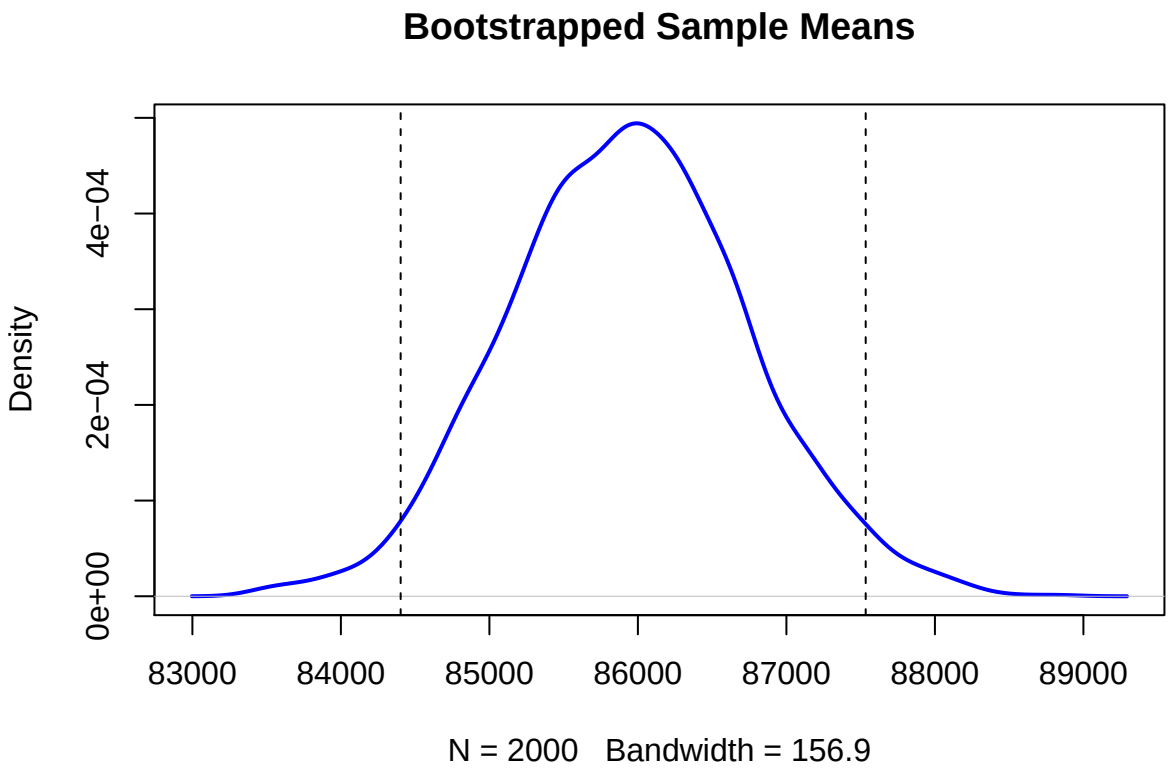
```
## [1] -5.141511
```

```
t_boots_ci_95 <- quantile(t_boots, probs = c(0.025, 0.975))
t_boots_ci_95
```

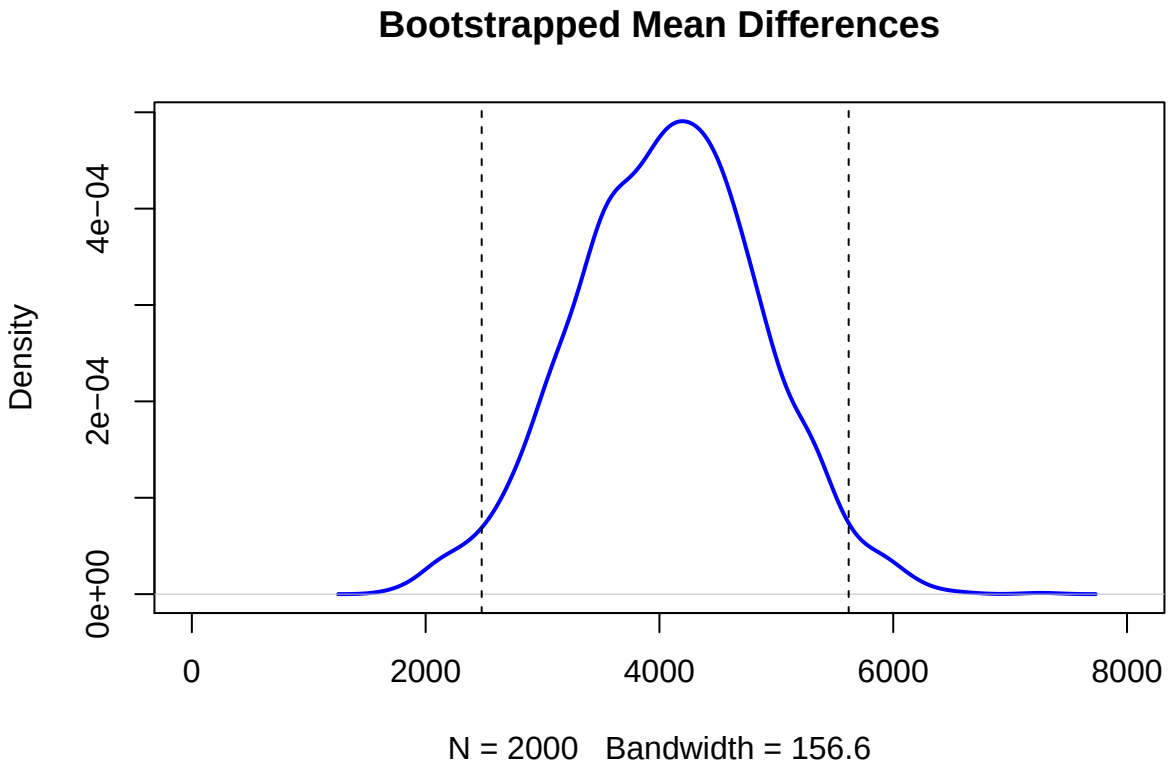
```
##      2.5%     97.5%
## -7.188286 -3.219681
```

iv) Plot the density curve of all 3 bootstraps above

```
plot(density(sample_means), col = 'blue', lwd = 2, main = 'Bootstrapped Sample Means')
abline(v = sample_means_ci_95, lty = 2)
```



```r
plot(density(mean_diffs), col = 'blue', lwd = 2, xlim = c(0, 8000),
     main = 'Bootstrapped Mean Differences')
abline(v = mean_diffs_ci_95, lty = 2)
```



```r
plot(density(t_boots), col = 'blue', lwd = 2, xlim = c(-10, 0),
     main = 'Bootstrapped t-statistics')
abline(v = t_boots_ci_95, lty = 2)
```