HW4

Black text are answers

Question 1)

Read the article Google explains how it spots malicious Android apps (malware) on its online store for mobile apps. The article describes how some malware suspiciously turn off a security feature (called Verify) when they are installed on an Android device. Note that there are other reasons why the Verify feature gets turned off. Google computes a “DOI score” for each app. The distribution of DOI scores is binomial, which Google approximates as a normal distribution.

(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; hint: we saw a function in class that can do this)

```
pnorm(-3.7)

## [1] 0.0001077997
```

In the article, Google calls the Z-score of an app's retention rate a DOI score; that is DOI score is Z-score in statistics. Therefore, we can use the function pnorm to convert DOI score into probability. The answer is 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?

```
2500000*pnorm(-3.7)

## [1] 269.4993
```

About 270 apps on the Play Store Google expects will maliciously turn off the Verify feature once installed.

Question 2)

The Hsinchu Rubber Company is in financial trouble because of a reputation of poor tire quality.They have launched a new advertising campaign to change their image. In their new advertisement, they claim that their tires can run an average of 90,000 km before needing to be replaced. The editors of a Taiwanese consumer magazine are skeptical and wanted to test the claim. They collected data from 360 drivers who use these tires and recorded how long the tires lasted. On average, the tires in this sample lasted 85945.29 km with a standard deviation of 14996.55 km. Is the company's new advertising claim to be believed?

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

i.How would you write your hypothesis? (not graded)

$H_0: \mu = 90000$

$H_1: \mu \neq 90000$

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

```
#95% confidence interval
right <- 85945.29 - 1.96*(14996.55/(360**0.5))
left <- 85945.29 + 1.96*(14996.55/(360**0.5))
cat ("95% confidence:,",right,"-",left)

## 95% confidence:, 84396.13 - 87494.45
```

The estimated population mean is 85945.29 and the 95 confidence interval of this estimate is from 84396.13 to 87494.45.

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

```
t = (85945.29-90000)/(14996.55/(360**0.5))
cat("t-statistic:",t)

## t-statistic: -5.130027
```

The t-statistic of my test is -5.130027.

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

```
p_value <- pt(t,df=299)*2
cat("\np value:", p_value)
```

```
##
## p value: 5.219607e-07
```

Here we conduct a two-sided t-test and set the significant level to 0.05. From the above calculation, we find out that the p value is smaller than the significant level. Therefore, we reject the null hypothesis, $\mu = 90000$;that is, the advertising claim might be incorrect.

(b)Let's use bootstrapping on the sample data itself (see tires.csv file) to examine this problem:

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

```
#Load data
tires <- c(read.csv("tires.csv"))$lifetime_km

set.seed(38)
number_boots <- 2700
#bootstrap function
boot_mean<-function(sample0) {
  resample <-sample(sample0, length(sample0), replace=TRUE)
  return(mean(resample))}
means <- replicate(number_boots, boot_mean(tires))
cat("bootstrap estimated mean:",mean(means),"\n")

## bootstrap estimated mean: 85951.5

cat("95% CI of estimated mean:",quantile(means,c(0.025,0.975)))

## 95% CI of estimated mean: 84432.81 87442.28
```

bootstrap estimated mean: 85951.5

95% CI of estimated mean: (84432.81,87442.28)

ii. Bootstrap the difference between the population mean and the hypothesized mean: what is the mean bootstrapped difference, and its 95% CI? (in each bootstrap iteration, find the difference between the mean of the bootstrapped sample and the hypothesized mean)

```
set.seed(38)
boot_mean_diffs<-function(sample0, mean_hyp) {
  resample <-sample(sample0, length(sample0), replace=TRUE)
  return( mean(resample) -mean_hyp)}
manager_hyp = 90000
mean_diffs<-replicate(number_boots, boot_mean_diffs(tires, manager_hyp))
cat("bootstrap mean difference:",mean(mean_diffs),"\n")

## bootstrap mean difference: -4048.496

cat("95% CI of estimated mean difference:",quantile(mean_diffs,c(0.025,
0.975)))
```

```
## 95% CI of estimated mean difference: -5567.192 -2557.717
```

bootstrap mean difference: -4048.496

95% CI of estimated mean difference: (-5567.192,-2557.717)

iii. the t-statistic: what is the mean bootstrapped t-statistic and its 95% CI? (in each bootstrap iteration, divide the difference between the mean of the bootstrapped sample and the hypothesized mean by the standard error of that bootstrapped sample – as you can see, the t-statistic is just a standardized difference of means)

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

cat("mean bootstrapped t-statistic:",mean(t_boots),"\n")

## mean bootstrapped t-statistic: -5.149735

cat("95% CI of mean bootstrapped t-statistic:",quantile(t_boots,c(0.025,
0.975)))

## 95% CI of mean bootstrapped t-statistic: -7.216829 -3.139113
```

mean bootstrapped t-statistic: -5.149811

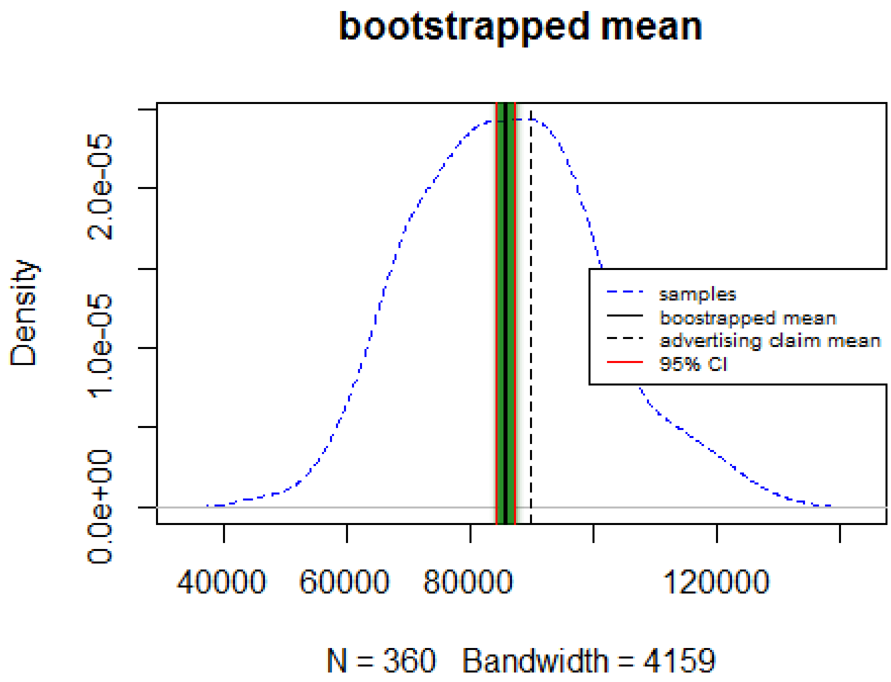
95% CI of mean bootstrapped t-statistic: (-7.216829,-3.139113)

Here we can find out that the mean bootstrapped t-statistic is similar to the one obtained from traditional way.

iv. Plot the density curve of all three bootstraps above.(for ii and iii make sure to include zero on the x-axis)

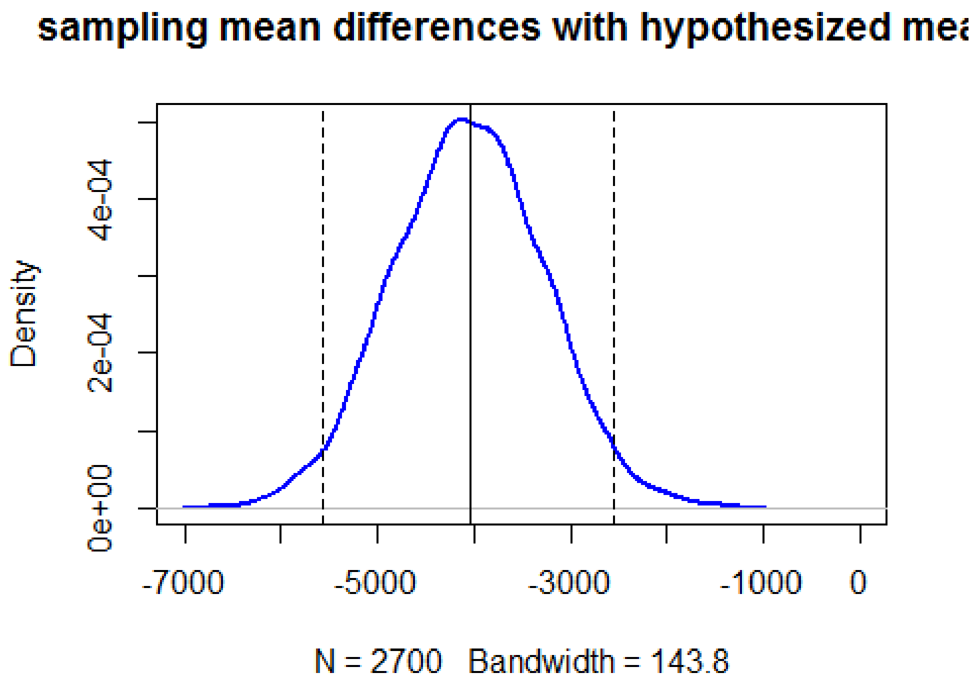
```
#bootstrapped mean
set.seed(38)
plot(density(tires), col="blue", lty="dashed", main="bootstrapped mean")
resamples <-replicate(number_boots, sample(tires, length(tires), replac
e=TRUE))
plot_resample_mean<-function(sample_i) {abline(v=mean(sample_i), col=rg
b(0.0, 0.4, 0.0, 0.01))
  return(mean(sample_i))}
sample_means<-apply(resamples, 2, FUN=plot_resample_mean)
sample_means_mean <- mean(sample_means)
abline(v=sample_means_mean, lwd=2)
abline(v=manager_hyp, lty="dashed")
abline(v=quantile(means,c(0.025,0.975)),col = "red")
legend(99500,0.000015, # places a legend at the appropriate place
       c("samples","boostrapped mean","advertising claim mean","95% CI
"), # puts text in the legend
```

```
  lty=c(2,1,2,1), # gives the legend appropriate symbols (lines)
  col=c("blue","black","black","red"),
  cex=0.6)
```



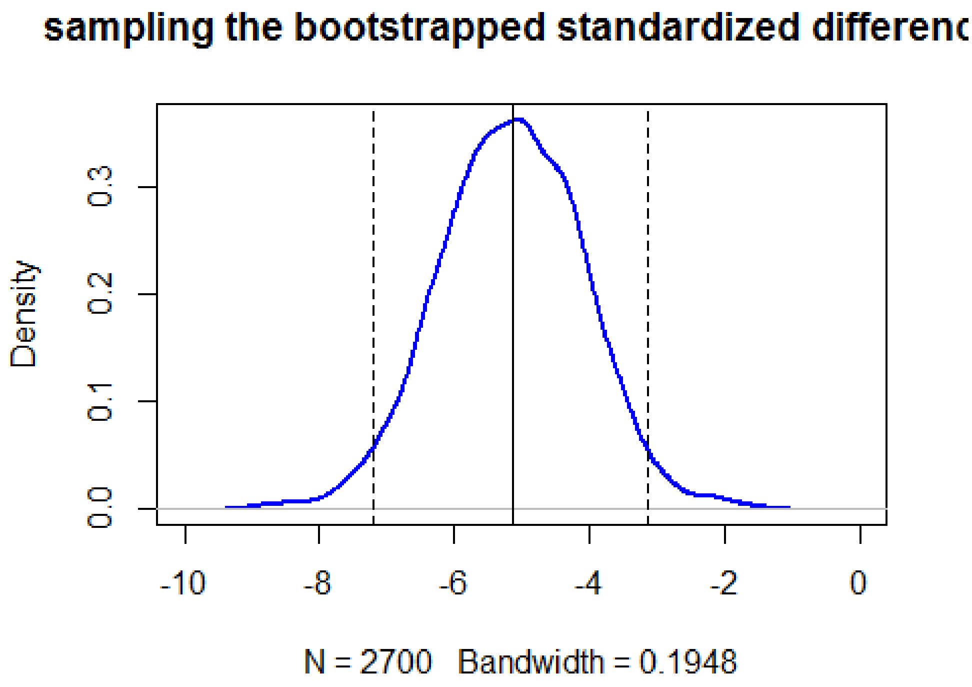
By visualizing the bootstrapped mean, it is clearly that the advertising mean is a little bit far from the bootstrapped mean, and out of the 95% confidece interval!

```
set.seed(38)
plot(density(mean_diffs),xlim=c(-7000,0),
     main = "sampling mean differences with hypothesized mean",
     lwd = 2,
     col="blue")
# 95% CI of difference b/w manager and auditor
diff_ci_95 <-quantile(mean_diffs, probs=c(0.025, 0.975))
abline(v=mean(mean_diffs))
abline(v=diff_ci_95, lty="dashed")
```



Here we visualize the sampling mean difference with advertising one. Ideally, the difference should be close to zero if two means are similar. However, from the above picture, the mean of difference is about -4000, which indicates that their are different in value.

```
set.seed(38)
plot(density(t_boots),
     main = "sampling the bootstrapped standardized difference",
     xlim=c(-10,0),
     lwd = 2,
     col="blue")
# 95% CI of difference b/w manager and auditor
diff_ci_95 <-quantile(t_boots, probs=c(0.025, 0.975))
abline(v=mean(t_boots))
abline(v=diff_ci_95, lty="dashed")
```



The above figure shows the distribution of the bootstrapped standardized difference. Basically, the shape of this plot is similar to the former one since the t-statistic is derived from standardization of mean difference.