Assignment 2

Kyle Ligon

September 30, 2018

A)

1. Explain how do the bootstrap methods work?
   Bootstrap methods take multiple runs with replacement of a sample. With the created distribution, you can map out confidence intervals of the statistic in question.
   2. and 3.

```
# Package(s) needed
library(boot)
library(tidyverse)
library(knitr)
ttSample <- c(0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

# Create a function f in R to compute the observed statistic
f <- function(DATA, i){
  return(mean(DATA[i]))
}

# bootstrapping with R replications
results <- boot(ttSample, f , R = 250)

# get confidence interval
boot.ci(results, conf = 0.90, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 250 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.9, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 90%   ( 0.7921,  1.0011 )   ( 0.8000,  1.0000 )
##
## Level     Percentile            BCa
## 90%   ( 0.80,  1.00 )   ( 0.70,  0.95 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.95, type="all")
```

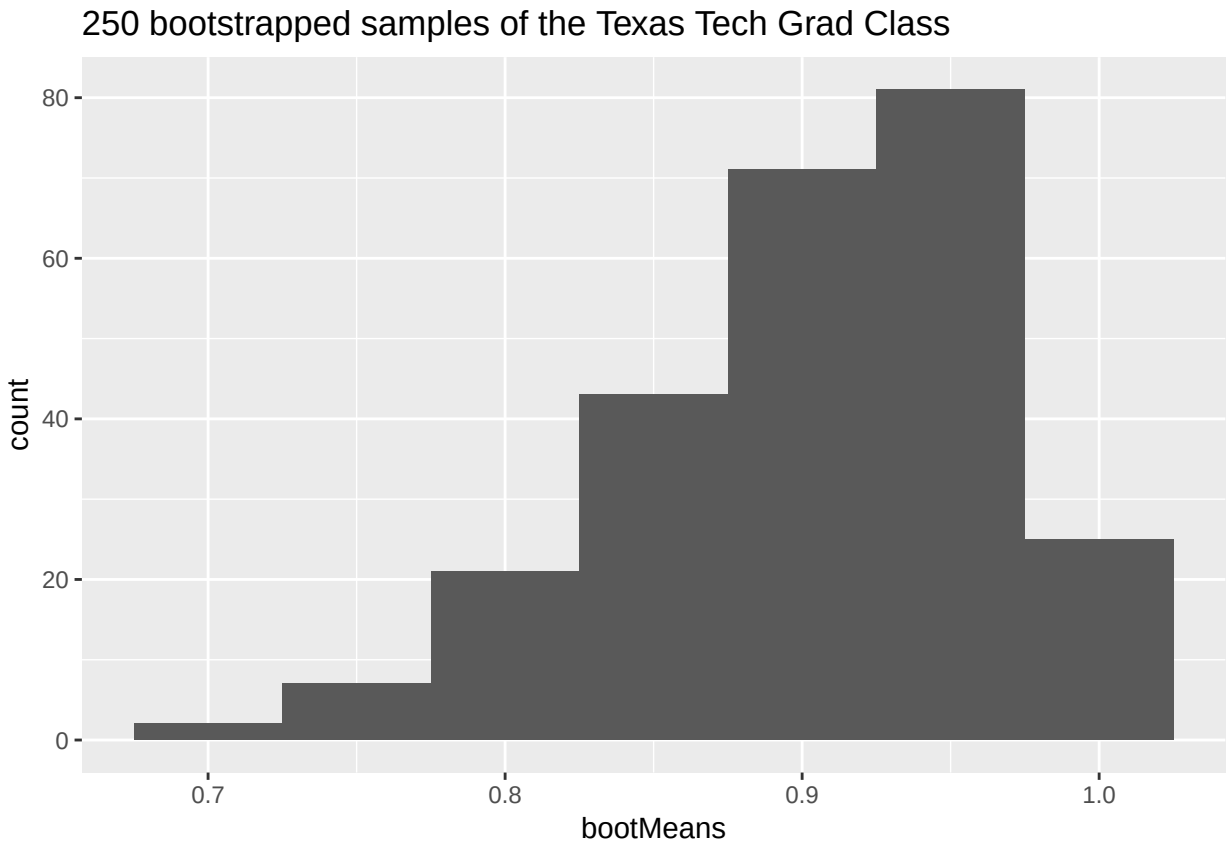
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 250 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.95, type = "all")
##
## Intervals :
```

```
## Level      Normal              Basic
## 95%   ( 0.7721,  1.0211 )   ( 0.8000,  1.0500 )
##
## Level     Percentile            BCa
## 95%   ( 0.75,  1.00 )   ( 0.70,  0.95 )
## Calculations and Intervals on Original Scale
## Some basic intervals may be unstable
## Some percentile intervals may be unstable
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.99, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 250 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.99, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 99%   ( 0.7330,  1.0602 )   ( 0.8000,  1.1000 )
##
## Level     Percentile            BCa
## 99%   ( 0.70,  1.00 )   ( 0.70,  0.95 )
## Calculations and Intervals on Original Scale
## Some basic intervals may be unstable
## Some percentile intervals may be unstable
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
r <- as.data.frame(bootMeans <- results$t)
#plotting the histogram
ggplot(r, aes(x = bootMeans)) +
  geom_histogram(binwidth = 0.05, bins = round(sqrt(nrow(r)), 0)) +
  ggtitle(label = paste0(results$R, " bootstrapped samples of the Texas Tech Grad Class"))
```




```
results <- boot(ttSample, f , R = 1000)
boot.ci(results, conf = 0.90, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.9, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 90%   ( 0.7913,  1.0011 )   ( 0.8000,  1.0000 )
##
## Level     Percentile            BCa
## 90%   ( 0.80,  1.00 )   ( 0.70,  0.95 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.95, type="all")
```

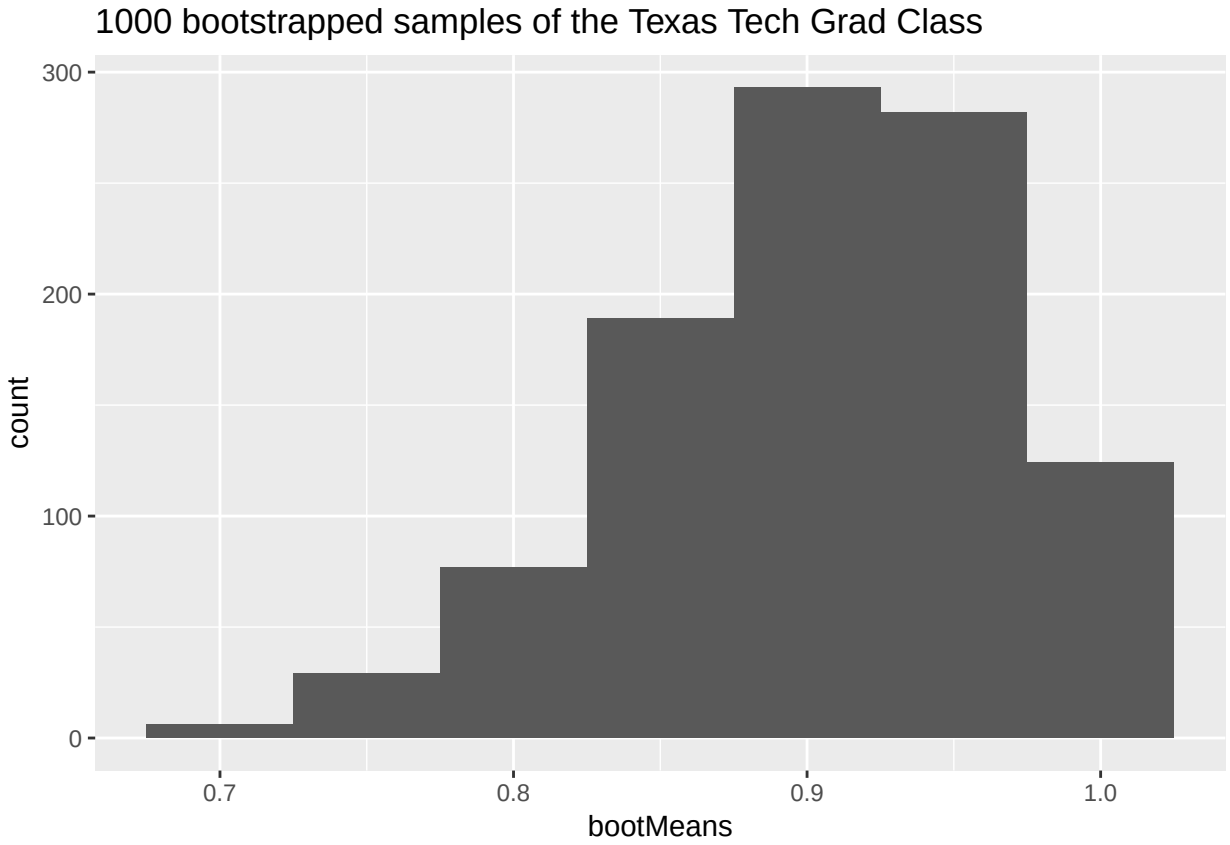
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.95, type = "all")
##
```

```
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.7712,  1.0212 )   ( 0.8000,  1.0500 )
##
## Level     Percentile            BCa
## 95%   ( 0.75,  1.00 )   ( 0.70,  0.95 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.99, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.99, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 99%   ( 0.7319,  1.0605 )   ( 0.8000,  1.1000 )
##
## Level     Percentile            BCa
## 99%   ( 0.7,  1.0 )   ( 0.7,  1.0 )
## Calculations and Intervals on Original Scale
## Some basic intervals may be unstable
## Some percentile intervals may be unstable
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
r <- as.data.frame(bootMeans <- results$t)
ggplot(r, aes(x = bootMeans)) +
  geom_histogram(binwidth = 0.05) +
  ggtitle(label = paste0(results$R, " bootstrapped samples of the Texas Tech Grad Class"))
```




```
results <- boot(ttSample, f , R = 5000)
boot.ci(results, conf = 0.90, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.9, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 90%   ( 0.7885,  1.0112 )   ( 0.8000,  1.0000 )
##
## Level     Percentile            BCa
## 90%   ( 0.80,  1.00 )   ( 0.65,  0.95 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.95, type="all")
```

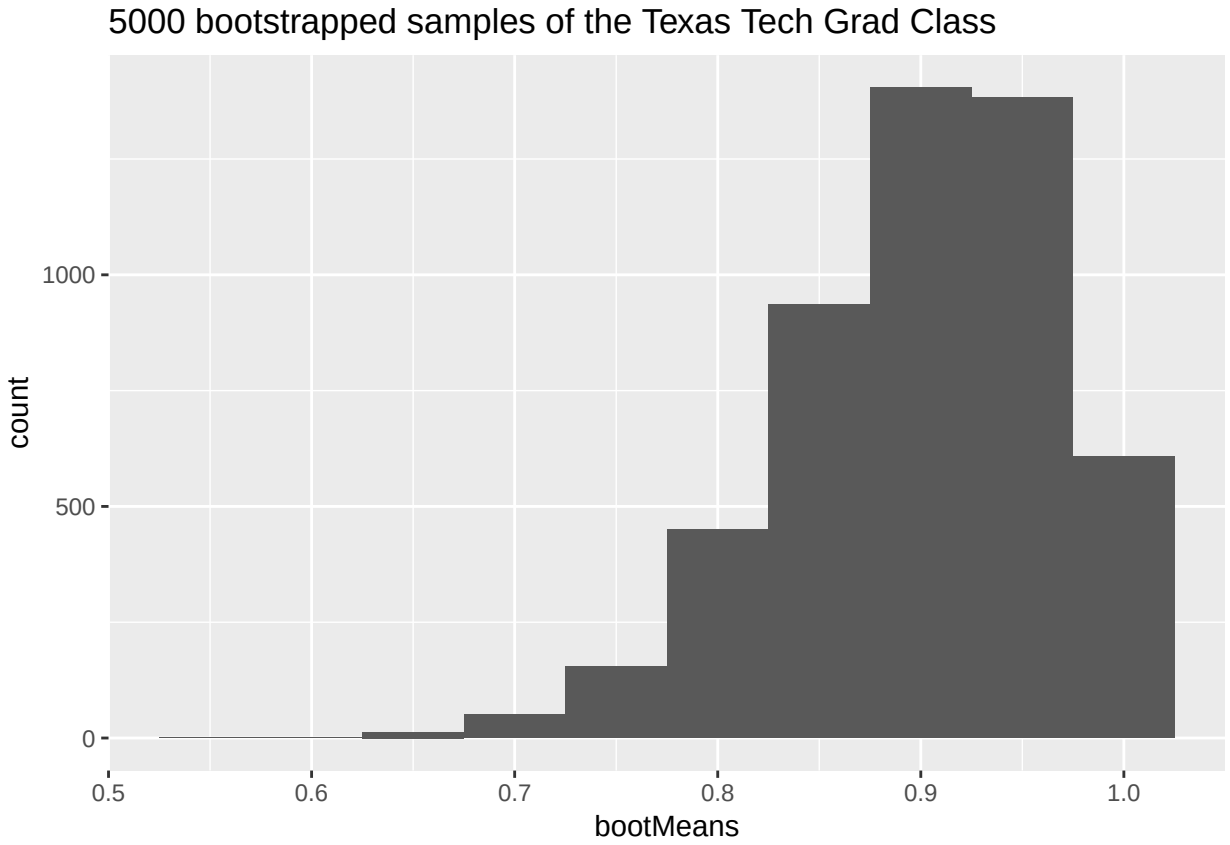
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.95, type = "all")
##
## Intervals :
```

```
## Level      Normal              Basic
## 95%   ( 0.7672,  1.0325 )   ( 0.8000,  1.0500 )
##
## Level     Percentile            BCa
## 95%   ( 0.75,  1.00 )   ( 0.55,  0.95 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.99, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.99, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 99%   ( 0.7255,  1.0742 )   ( 0.8000,  1.1000 )
##
## Level     Percentile            BCa
## 99%   ( 0.70,  1.00 )   ( 0.55,  1.00 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
r <- as.data.frame(bootMeans <- results$t)
ggplot(r, aes(x = bootMeans)) +
  geom_histogram(binwidth = 0.05) +
  ggtitle(label = paste0(results$R, " bootstrapped samples of the Texas Tech Grad Class"))
```




```
results <- boot(ttSample, f , R = 10000)
boot.ci(results, conf = 0.90, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.9, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 90%   ( 0.7910,  1.0103 )   ( 0.8000,  1.0000 )
##
## Level     Percentile            BCa
## 90%   ( 0.80,  1.00 )   ( 0.65,  0.95 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.95, type="all")
```

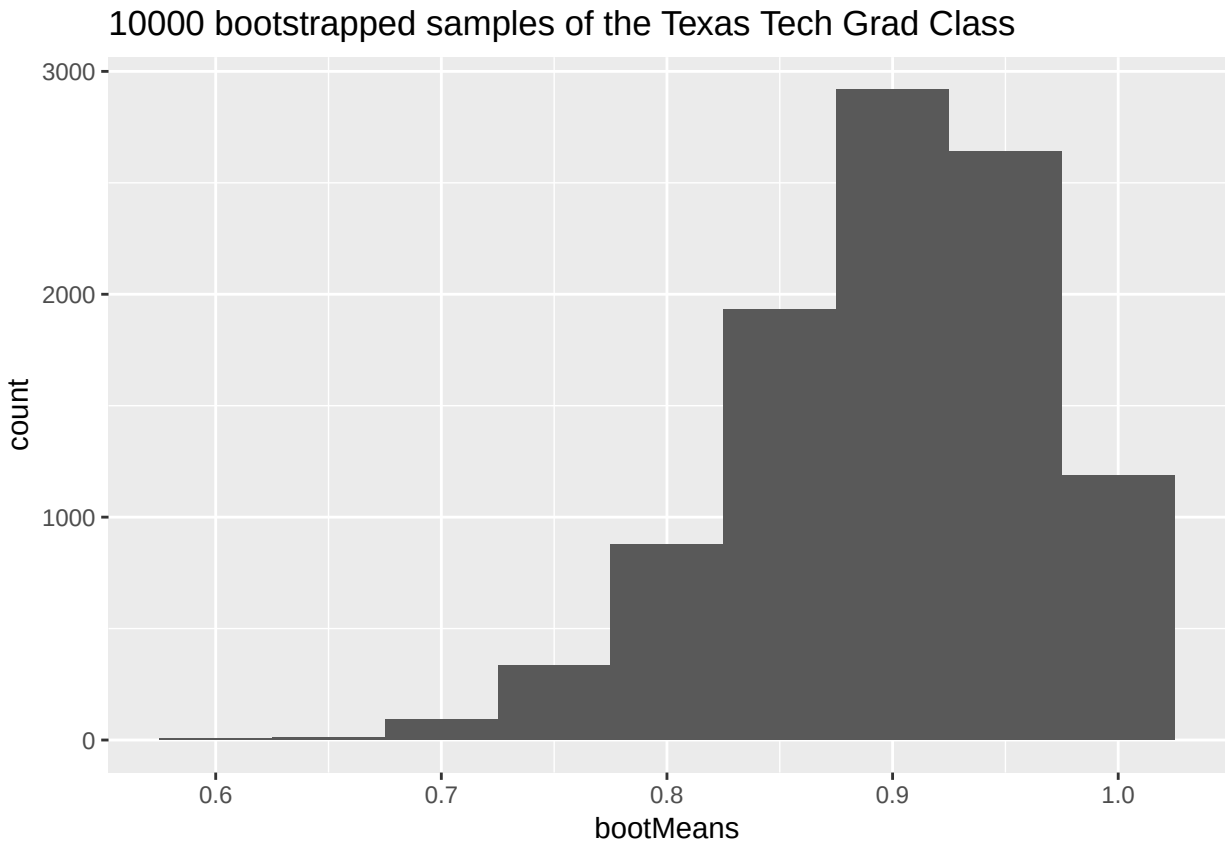
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.95, type = "all")
##
## Intervals :
```

```
## Level      Normal              Basic
## 95%   ( 0.7700,  1.0313 )   ( 0.8000,  1.0500 )
##
## Level     Percentile            BCa
## 95%   ( 0.75,  1.00 )   ( 0.60,  0.95 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```
boot.ci(results, conf = 0.99, type="all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.99, type = "all")
##
## Intervals :
## Level      Normal              Basic
## 99%   ( 0.7289,  1.0724 )   ( 0.8000,  1.1000 )
##
## Level     Percentile            BCa
## 99%   ( 0.7,  1.0 )   ( 0.6,  1.0 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
r <- as.data.frame(bootMeans <- results$t)
ggplot(r, aes(x = bootMeans)) +
  geom_histogram(binwidth = 0.05) +
  ggtitle(label = paste0(results$R, " bootstrapped samples of the Texas Tech Grad Class"))
```



4. Discuss and comment on the results.
   As our sample size increases, we see the distributions center more closely around p_hat = 0.90. Additionally, the standard errors tighten up as the sample size increases. finally, the confidence intervals shrink accordingly as the bootstrap samples increase.

B) 1. Generate n in {5, 10, 30} samples from the standard normal distribution. Then, calculate the confidence intervals from each.

```
# 5 samples from standard normal distribution
set.seed(1)
n_5 <- rnorm(5)

confidence_int <- function(vector){
  #high value of the confidence interval
  high_val <- round(mean(vector) + qnorm(0.975) * sd(vector) / sqrt(length(vector)), 4)
  #low value of the confidence interval
  low_val <- round(mean(vector) + qnorm(0.025) * sd(vector) / sqrt(length(vector)), 4)
  #ouput statement
  cat(paste0("The two tail 95% confidence interval, of sample size ", length(vector),
      ", \nis between the values of ", low_val, " and ", high_val))
}
# calling the function
confidence_int(n_5)
```

```
## The two tail 95% confidence interval, of sample size 5,
## is between the values of -0.7131 and 0.9716
```

```r
set.seed(1)
n_10 <- rnorm(10)

confidence_int(n_10)
```

```
## The two tail 95% confidence interval, of sample size 10,
## is between the values of -0.3516 and 0.616
```

```r
set.seed(1)
n_30 <- rnorm(30)

confidence_int(n_30)
```

```
## The two tail 95% confidence interval, of sample size 30,
## is between the values of -0.2482 and 0.4131
```

2) Calculate the confidence intervals for the bootstrap samples where n in 1000, 5000, 10000, 100000. I'm going to use a sample of 10 random normal variables whose mean and standard deviation approaches 0 and 1 as n approaches infinity respectively.

```r
set.seed(1)
normVec <- rnorm(10)

results <- boot(normVec, f , R = 1000)
boot.ci(results)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.3180,  0.5941 )   (-0.3419,  0.5719 )
##
## Level     Percentile            BCa
## 95%   (-0.3075,  0.6063 )   (-0.2630,  0.6567 )
## Calculations and Intervals on Original Scale
```

```r
results <- boot(normVec, f, R = 5000)
boot.ci(results)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.3247,  0.5932 )   (-0.3339,  0.5860 )
##
## Level     Percentile            BCa
## 95%   (-0.3216,  0.5983 )   (-0.2900,  0.6259 )
## Calculations and Intervals on Original Scale
```

```
results <- boot(normVec, f, R = 10000)
boot.ci(results)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.3277,  0.5915 )   (-0.3376,  0.5825 )
##
## Level     Percentile            BCa
## 95%   (-0.3181,  0.6020 )   (-0.3020,  0.6251 )
## Calculations and Intervals on Original Scale
```

```
results <- boot(normVec, f, R = 100000)
boot.ci(results)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 100000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.3249,  0.5903 )   (-0.3328,  0.5767 )
##
## Level     Percentile            BCa
## 95%   (-0.3123,  0.5972 )   (-0.2924,  0.6221 )
## Calculations and Intervals on Original Scale
```

3) Determe the smallest sample size n in N* that would provide the best approximation compared to the theory.
According to Efron and Tibshirane(1986), a minimum threshold of 250 bootstrap samples is necessary to be accurate. With that in mind, they also say 1000 is a safe bet to land on an approximate estimation. With this previous work done(looking at the difference between the Normal Group and the Basic Group), I would look back to problem 2 and do 1000 bootstrap samples as my barebone minimum value for the number of replications as it provides a decent cut point for while balancing distance from the normal bootstrap method.

4) Discuss and comment on the results.
As can be seen above, our biggest gains come from 1000 bootstrap samples up to 5000 and then we hit dimishing returns for 4 decimal places. As an investigation, I'd like to see the barest minimum 250 bootstrap samples as recommended by Efron and Tibshirane(1986).

```
results <- boot(normVec, f, R = 250)
boot.ci(results, type = "norm")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 250 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "norm")
```

```
##
## Intervals :
## Level      Normal
## 95%   (-0.3340,  0.5657 )
## Calculations and Intervals on Original Scale
```

Surprisingly, the bootstrap group of 250 does come close to approximating the confidence interval. That being said, as long as compute stays cheap. I would rather push for more bootstraps in my confidences intervals. So 1000, would be amount that I lean on going forward.