DATA 602 Assignment 2

Michael Ellsworth, ID 30101253

September 23, 2019

Question 1

Refer to Question 11 from Assignment 1:

a.

Compute the probability that another random sample of the same size will produce a sample mean that is at least the same value as this particular sample produces. $P(\overline{X} > 5.6875)$

```
1 - pnorm(5.6875, mean = 5, sd = 1.5/12**0.5)
```

```
## [1] 0.0561756
```

b.

Observe the value of the sample standard deviation S (which you computed in Exercise 11 of Assignment 1): Compute the probability that another random sample (again, of the same size) will yield a sample standard deviation that is between 0.5 hours and 1 hour. $P(0.5 < S < 1)$

$$
\begin{aligned}
P(0.5 \le S \le 1) &= P(0.5^2 \le S^2 \le 1^2) \\
&= P\left(\frac{(n-1)*0.5^2}{\sigma^2} \le \frac{(n-1)*S^2}{\sigma^2} \le \frac{(n-1)*1^2}{\sigma^2}\right) \\
&= P\left(\frac{(12-1)*0.5^2}{1.5^2} \le \chi^2_{12-1} \le \frac{(12-1)*1^2}{1.5^2}\right) \\
&= P(1.22 \le \chi^2_{11} \le 4.89) \\
&= P(\chi^2_{11} \le 4.89) - P(\chi^2_{11} \le 1.22) \\
&\quad \overset{\text{pchisq}(4.89,11)}{\qquad} \overset{\text{pchisq}(1.2,11)}{\qquad} \\
&= 0.06343368 \\
&\approx 0.0634
\end{aligned}
$$

```
pchisq((12-1)*1**2/1.5**2, 11) - pchisq((12-1)*0.5**2/1.5**2, 11)
```

```
## [1] 0.06343368
```

Question 2

A recent poll[1] found that 4 in 5 Canadians, or 80%, support "Canada's Foreign Minister's decision to call for the release of human-rights activists detained in Saudi Arabia".

a.

State the mean and standard deviation of the disribution of $\hat{p}$, based on a random sample of $n = 500$ Canadians, presuming $p = 0.80$.

The mean and standard deviation of the distribution of $\hat{p}$ is:

$$\mu_{\hat{p}} = 0.80$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.80(1 - 0.80)}{500}} \approx 0.0179$$

b.

A sample of $n = 500$ Canadians revealed that 374, or $\hat{p} = \frac{374}{500} = 0.748$, supported the Foreign Minister's decision to call for Saudi Arabia to release human-rights activists detained. How likely is this outcome? Compute the probability of a sample of $n = 500$ producing a sample proportion that is less than or equal to what was observed.

```
pnorm(374/500, mean = 0.8, sd = (0.8*(1 - 0.8)/500)**0.5)
```

```
## [1] 0.001825217
```

$P(\hat{p} \leq 0.748) = 0.001825217 \approx 0.002$

c.

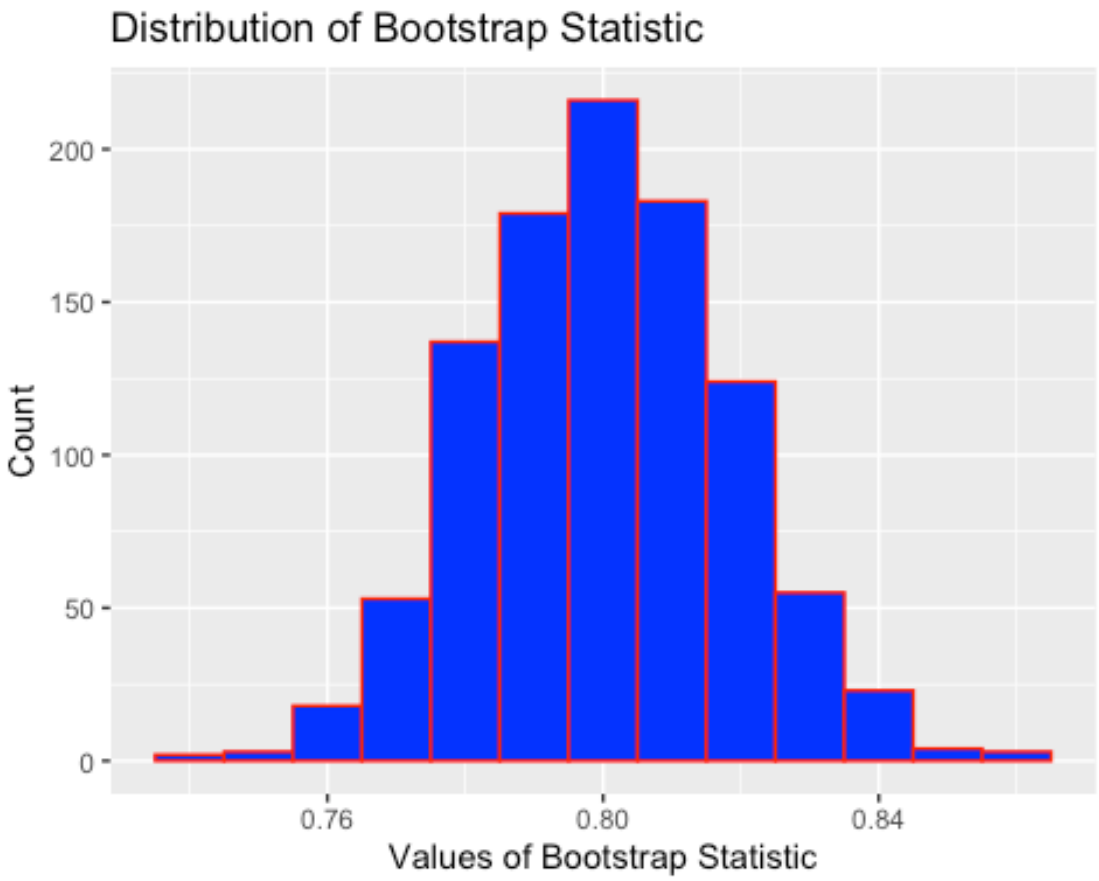
Consider the steps and associated R Code required to generate a distribution of the sample proportion, $\hat{p}$, when sampling $n = 500$ Canadians to determine national support for the Foreign Minister's decision to call for the release of human-rights activists detained in Saudi Arabia. Create, then run your code to determine the proportion of your $\hat{p}$s that are less than or equal to 0.748. Provide this proportion. (Hint: You are computing an empirical probability.)

```
ntimes_2c <- 1000
downsize_vector <- c(rep(0, 20), rep(1, 80))
bootstrap_2c = do(ntimes_2c) * mean(sample(downsize_vector, size = 500,
replace=TRUE))
bootstrap_2c %>%
  ggplot(aes(mean)) +
```

---

[1] https://www.ipsos.com/en-ca/news-polls/HomeEquity-Bank-Downsizing-Poll-Sept-19-2018

```
  geom_histogram(col='red', fill='blue', binwidth=0.01) +
  xlab("Values of Bootstrap Statistic") +
  ylab("Count") +
  ggtitle("Distribution of Bootstrap Statistic")
```




```
proportion_2c <- bootstrap_2c %>%
  filter(mean <= 374/500) %>%
  count()/ntimes_2c
pull(proportion_2c, n)
```

```
## [1] 0.003
```

Question 3

Billy plays purchases one 6-49 lottery ticket every week and keeps track of the number of "matches"" he has on each of his tickets. To be clear, a "match" will occur when a number on his ticket matches a number that appears in the winning combination. A random variable X that keeps track of the number of matching numbers Billy experiences per week has the probability distribution function with a mean and standard deviation of

$$P(X = x) = \frac{\binom{6}{x}\binom{43}{6-x}}{\binom{49}{6}} \quad x = 0,1,2,3,4,5,6.$$

$$E(X) = \mu_X = \frac{36}{49} = 0.7347 \quad \text{and} \quad SD(X) = \sigma_X = 0.4179$$

Billy claims that in a year (52 weeks), on average, he manages to have at least one matching number on his 6-49 ticket. What do you think about Billy's claim? Provide a brief commentary about Billy's claim using your current knowledge of statistics and probability theory.

```
1 - pnorm(1, mean = 0.7347, sd = 0.4179/(52**0.5))
```

```
## [1] 2.348306e-06
```

$P(\overline{X} \geq 1) = 0.000002348 \approx 0.000002$

Billy is either extraordinarily lucky or he is full of it. His claim is highly improbable.

Question 4

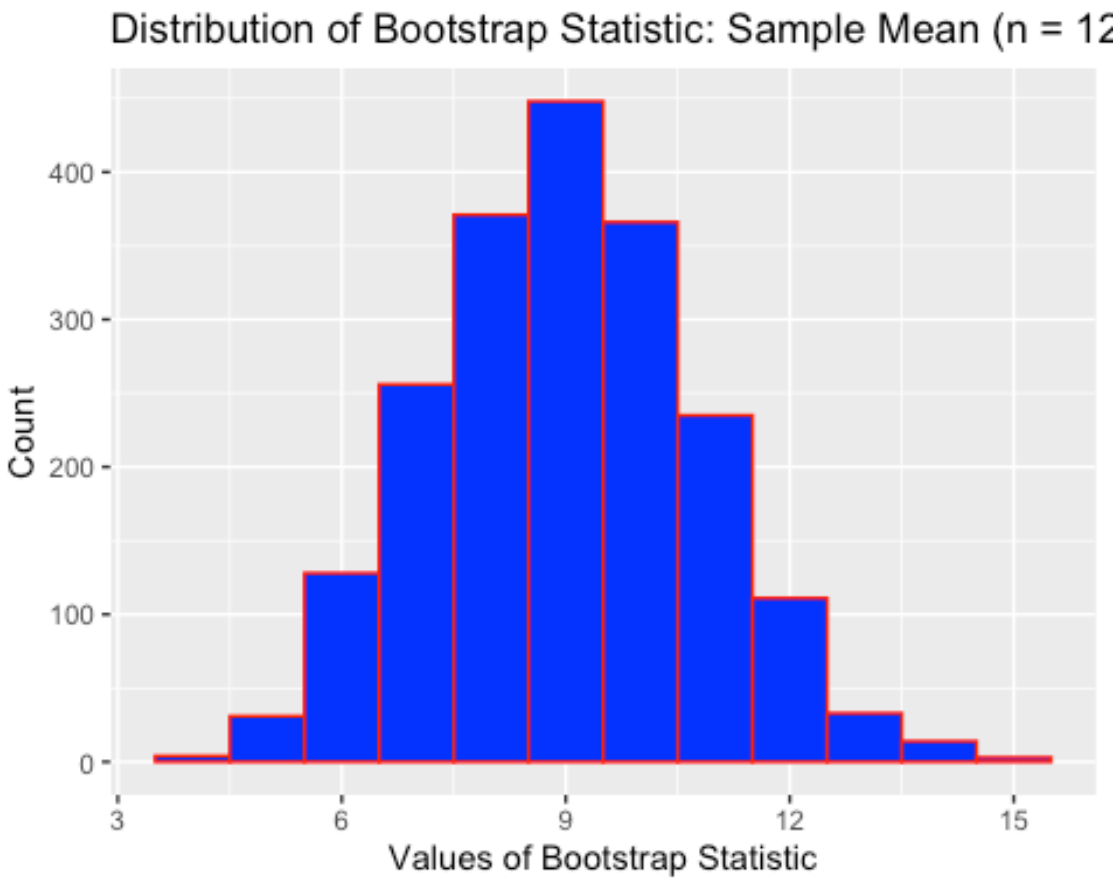
A common measure of toxicity for any pollutant is the concentration of the pollutant that will kill half of the test species in a given amount of time (usually about 96 hours for the fish species). This measurement is called the LC50, which refers to the lethal concentration killing 50% of the test species). The Environmental Protection Agency has collected data on LC50 measurements for certain chemicals likely to be found in freshwater and lakes. For a certain species of fish, the LC50 measurements (in parts per million) for DDT in 12 experiments to determine the LC50 "dose" are

16,5,21,19,10,5,8,2,7,2,4,9

a.

Use R studio to create the bootstrap distribution of the sample mean $\overline{X}$. Use 2000 resamples in your work.

```
ntimes_4a = 2000  #number of times to resample
nsize_4a = 12 #sample size
LC50_sample = c(16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9)
bootstrap_4a = do(ntimes_4a) * mean(sample(LC50_sample, size = nsize_4a,
replace=TRUE))
bootstrap_4a %>%
  ggplot(aes(mean)) +
  geom_histogram(col='red', fill='blue', binwidth=1) +
  xlab("Values of Bootstrap Statistic") +
  ylab("Count") +
  ggtitle("Distribution of Bootstrap Statistic: Sample Mean (n = 12)")
```



b.

From your result in (a), find a 95% confidence interval for μ, the mean LC50 measurement for DDT. Interpret the meaning of your interval in the ***context of these data****.*

```
LC50_95 <- qdata(~mean, c(0.025, 0.975), data = bootstrap_4a)
cat("The lower bound of the 95% confidence interval is",
    round(LC50_95$quantile[1], digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(LC50_95$quantile[2], digits = 2),
    "")
```

```
## The lower bound of the 95% confidence interval is 5.67 and the upper bound
of the 95% confidence interval is 12.5
```

The 95% confidence interval for μ is: $5.58 \leq \mu \leq 12.59$ (from this particular run of the bootstrap). This is the 95% confidence interval of the mean concentration of DDT that kills half of a species following 96 hours of exposure.

c.

Compute the 95% confidence interval for μ using the t-version of confidence interval. Ensure you appropriately present your finding/result.

```
LC50_df = data.frame(LC50_sample = c(16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9))
LC50_mean = favstats(~LC50_sample, data = LC50_df)$mean  #assigns sample mean
to meanbill
LC50_sd = favstats(~LC50_sample, data = LC50_df)$sd  #assigns sample standard
deviation to sdbill
LC50_lbave = LC50_mean - (qt(0.975, nsize_4a - 1)*(LC50_sd)/sqrt(nsize_4a))
LC50_ubave = LC50_mean + (qt(0.975, nsize_4a - 1)*(LC50_sd)/sqrt(nsize_4a))
cat("The lower bound of the 95% confidence interval is",
    round(LC50_lbave, digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(LC50_ubave, digits = 2),
    "")
```

```
## The lower bound of the 95% confidence interval is 4.92 and the upper bound
of the 95% confidence interval is 13.08
```

The 95% confidence interval for μ is: $4.92 \leq \mu \leq 13.08$

d.

Compare your results in parts (b) and (c). If you were to report one of these confidence intervals, which would you report? Explain your answer.

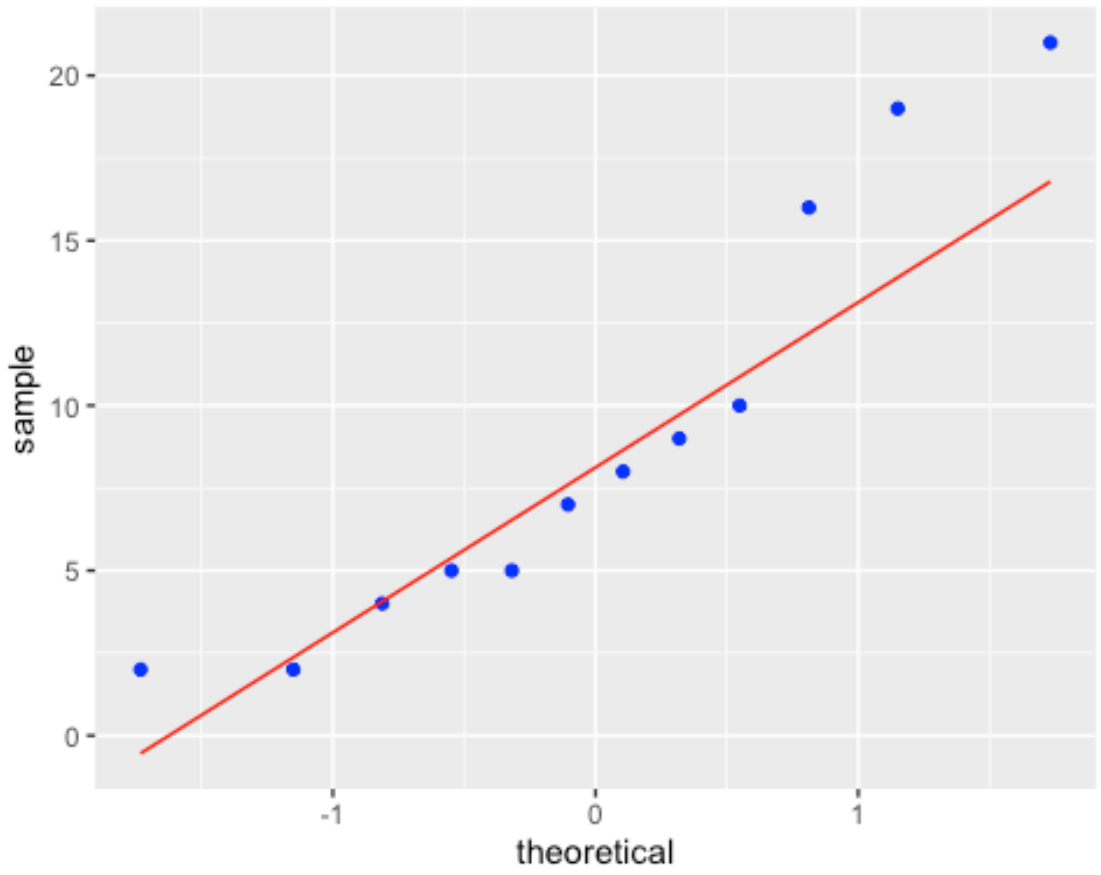
The reported confidence interval should be the bootstrap interval calculated in part (b). This methodology does not require the the data to follow a normal distribution, which is required in the t version of the confidence interval.

e.

The confidence interval you computed in part (c) is valid provided a certain condition holds. Use ggplot() to create a graph that is used to check this condition. From your plot, can you infer that this condition is satisfied? Explain.

Since the majority of the points on the following Normal Probability Plot fall along a straight line, we can say that this data conforms to a Normal probability model.

```
LC50_df %>%
  ggplot(aes(sample = LC50_sample)) +
  stat_qq(col = 'blue') +
  stat_qqline(col = 'red')
```



Question 5

Ipsos Reid recently reported[2] *on a survey conducted on "Baby-Boomer" Canadians (Canadians aged 55 or older) homeowners and found that of* $n = 1866$ *who have either downsized their home or plan to downsize their home, 571 indicates they either downsized or plan to downsize to take the equity out of their home to live comfortably in retirement.*

a.

Compute a 95% confidence interval for p, the proportion of all Canadians aged 55 years or older homeowners who have either downsized or plan to downsize to take equity out of their home to live comfortably in retirement.

```
downsize_stats <- binom.confint(571, 1866, conf.level = 0.95,
method="agresti-coull")
cat("The lower bound of the 95% confidence interval is",
    round(downsize_stats$lower, digits = 2),
    "and the upper bound of the 95% confidence interval is",
```

[2] https://www.ipsos.com/en-ca/news-polls/HomeEquity-Bank-Downsizing-Poll-Sept-19-2018

```
    round(downsize_stats$upper, digits = 2),
    "")
```

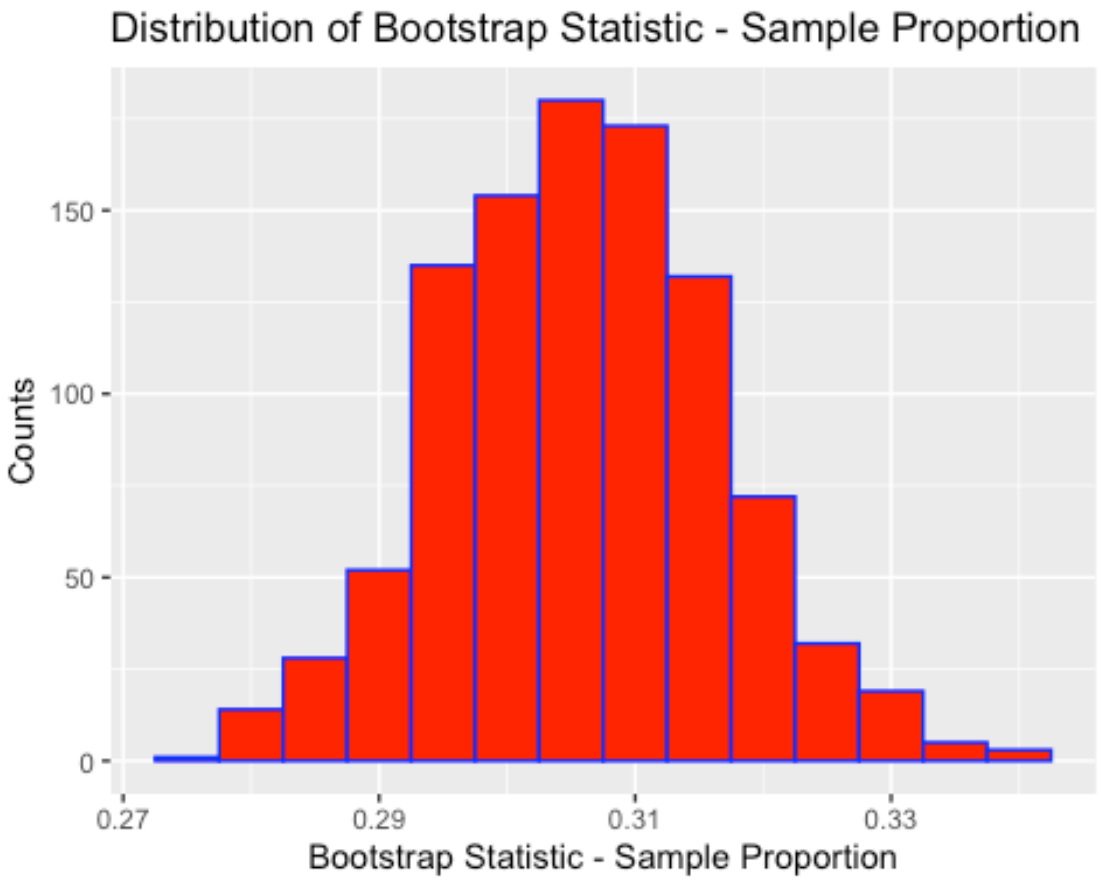
```
## The lower bound of the 95% confidence interval is 0.29 and the upper bound
of the 95% confidence interval is 0.33
```

The 95% confidence interval for p is: $0.29 \leq \mu \leq 0.33$

b.

Similar to your work in Question 4(b), create the distribution of the bootstrap statistic $\hat{p}$.

```
downsize_vector = c(rep(0, 1866 - 571), rep(1, 571))
nsize = length(downsize_vector)
bootstrap_5b = do(1000) * mean(resample(downsize_vector, nsize))
bootstrap_5b %>%
  ggplot(aes(x = mean)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.005) +
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```



c.

From your result in (b), compute the 95% confidence interval for p.

```
downsize_95 <- qdata(~mean, c(0.025, 0.975), data = bootstrap_5b)
cat("The lower bound of the 95% confidence interval is",
    round(downsize_95$quantile[1], digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(downsize_95$quantile[2], digits = 2),
    "")
```

```
## The lower bound of the 95% confidence interval is 0.28 and the upper bound
of the 95% confidence interval is 0.33
```

The 95% confidence interval for p is: $0.29 \leq \mu \leq 0.33$

d.

compare your results in (a) and (c). which interval should you report? Report the interval and interpret its meaning on the context of these data.

Since the calculated intervals are very similar in (a) and (c) (equal to two decimal points), either could be reported. In this case, it can be said with 95% accuracy that the proportion of all Canadians aged 55 years or older homeowners who have either downsized or plan to downsize to take equity out of their home to live comfortably in retirement, is between 29% and 33%.

$$0.29 \leq p \leq 0.33$$

Question 6

Does one's educational level influence their opinion about vaccinations? A recent Angus Reid[3] *survey was taken. Each person sampled was asked to respond to the statement "The science around vaccinations isn't clear." Respondents either "strongly agree", "moderately agree", "moderately disagree", or "strongly disagree". The sample was partitioned by level of education.*

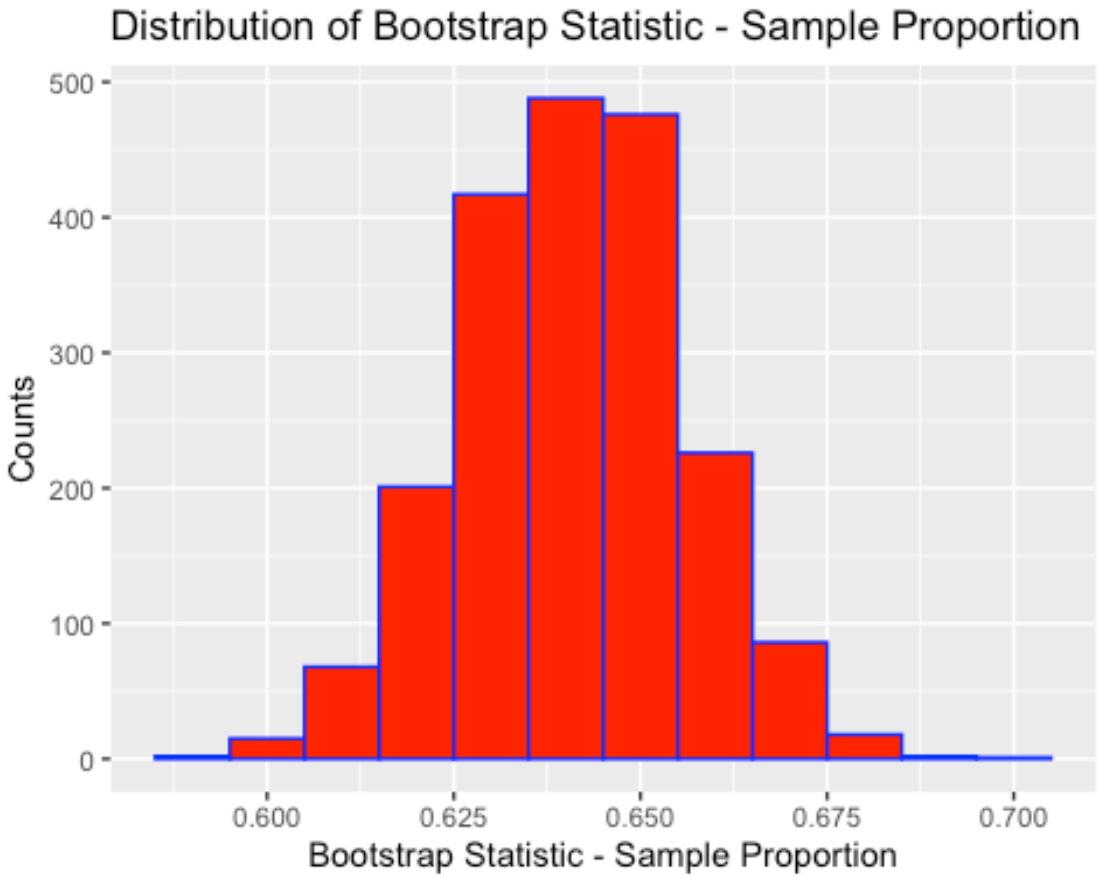
There were $n = 670$ *respondants who's highest level of education was high school or less, of which 348 "disagreed" (moderately disagree or stongly disagree). There were also* $n = 376$ *who's highest level of education was at least an undergraduate university eduction. Of these, 274 diagreed.*

[3] http://angusreid.org/wp-content/uploads/2015/02/2015.02.13-Vaccinations.pdf

a.

Consider the population consisting of all persons, who's highest level of education was high school or less and the bootstrap statistic $\hat{p}_{HS}$. Using 2000 iterations/replications, create a bootstrap distribution of $\hat{p}_{HS}$. Display your distribution.

```
hs_vector <- c(rep(0, 376), rep(1, 670))
bootstrap_6a = do(2000) * mean(resample(hs_vector, 376 + 670))
bootstrap_6a %>%
  ggplot(aes(x = mean)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.01) +
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```

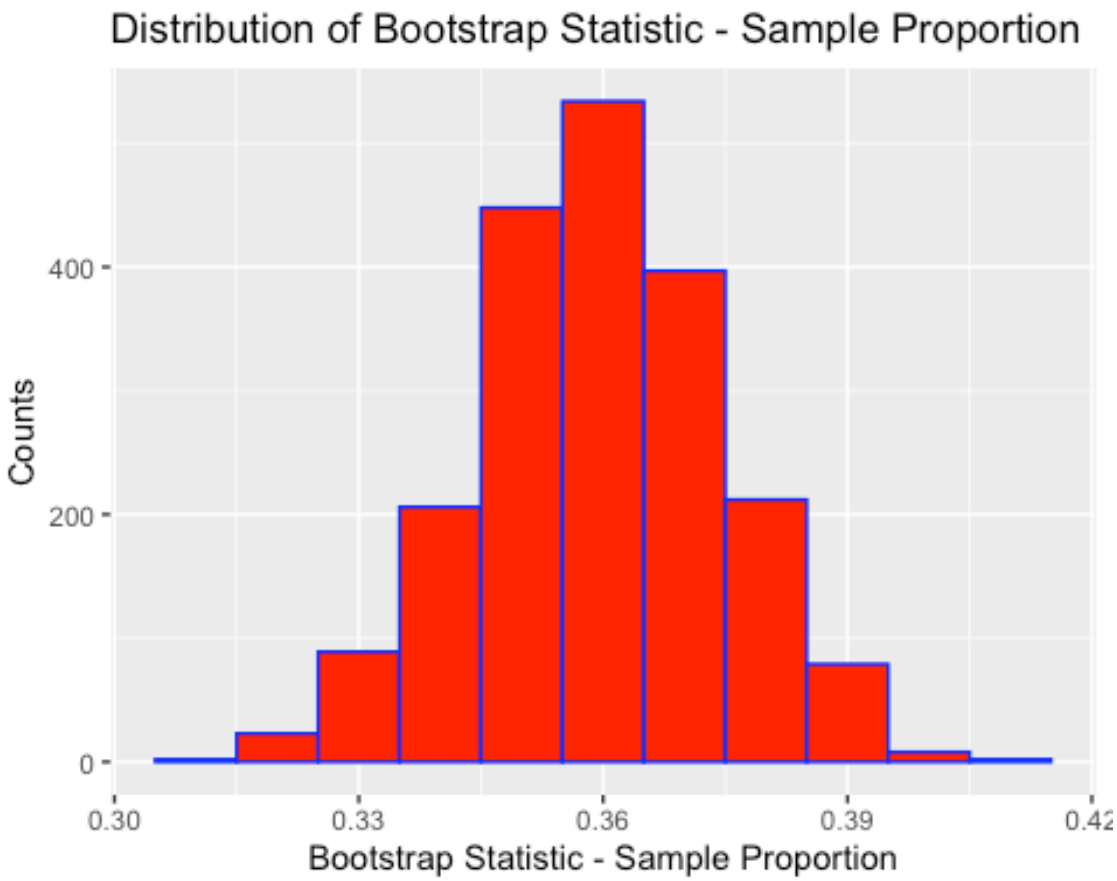


b.

*Now consider a **different** population that consits of all persons who's highest level of education was at least an undergraduate degree. Repeat part (a), creating a bootstrap distribution for $\hat{p}_{Uni}$. (Again, display your distribution).*

```
hs_vector2 <- c(rep(0, 670), rep(1, 376))
bootstrap_6b = do(2000) * mean(resample(hs_vector2, 376 + 670))
```

```r
bootstrap_6b %>%
  ggplot(aes(x = mean)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.01) +
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```

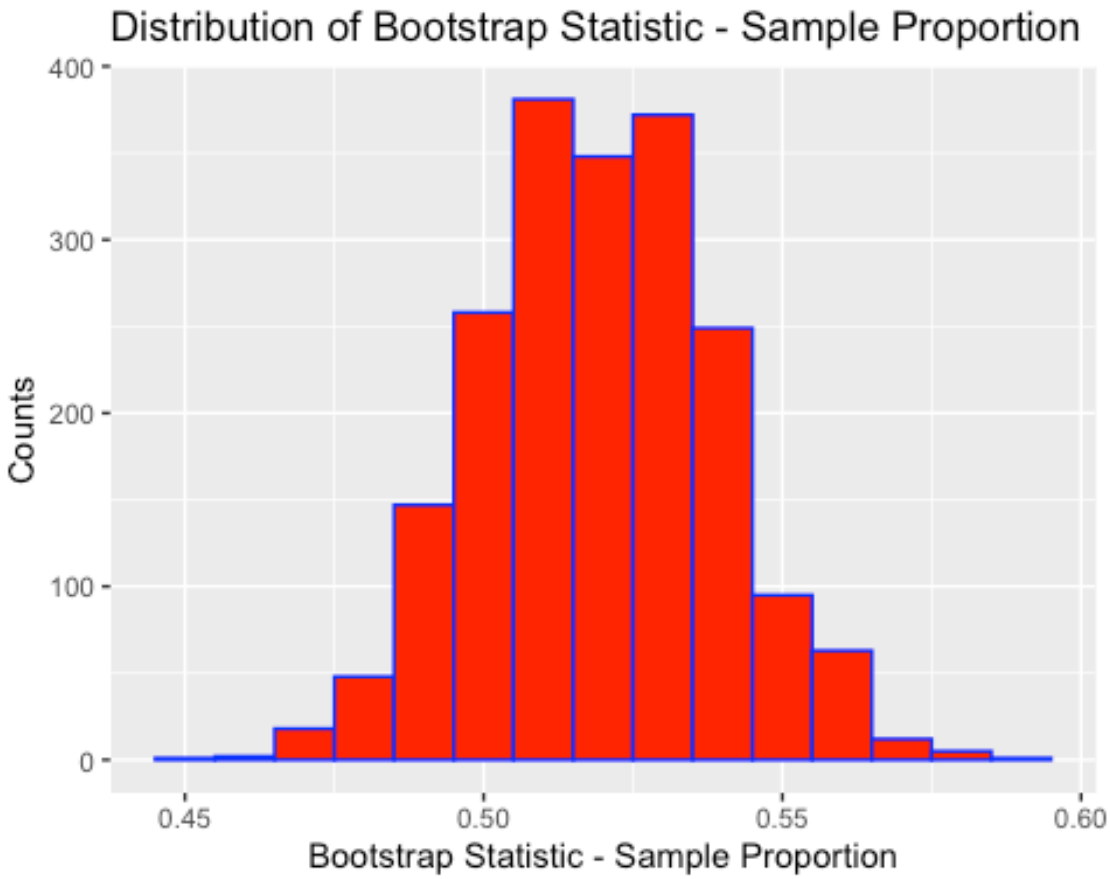


c.

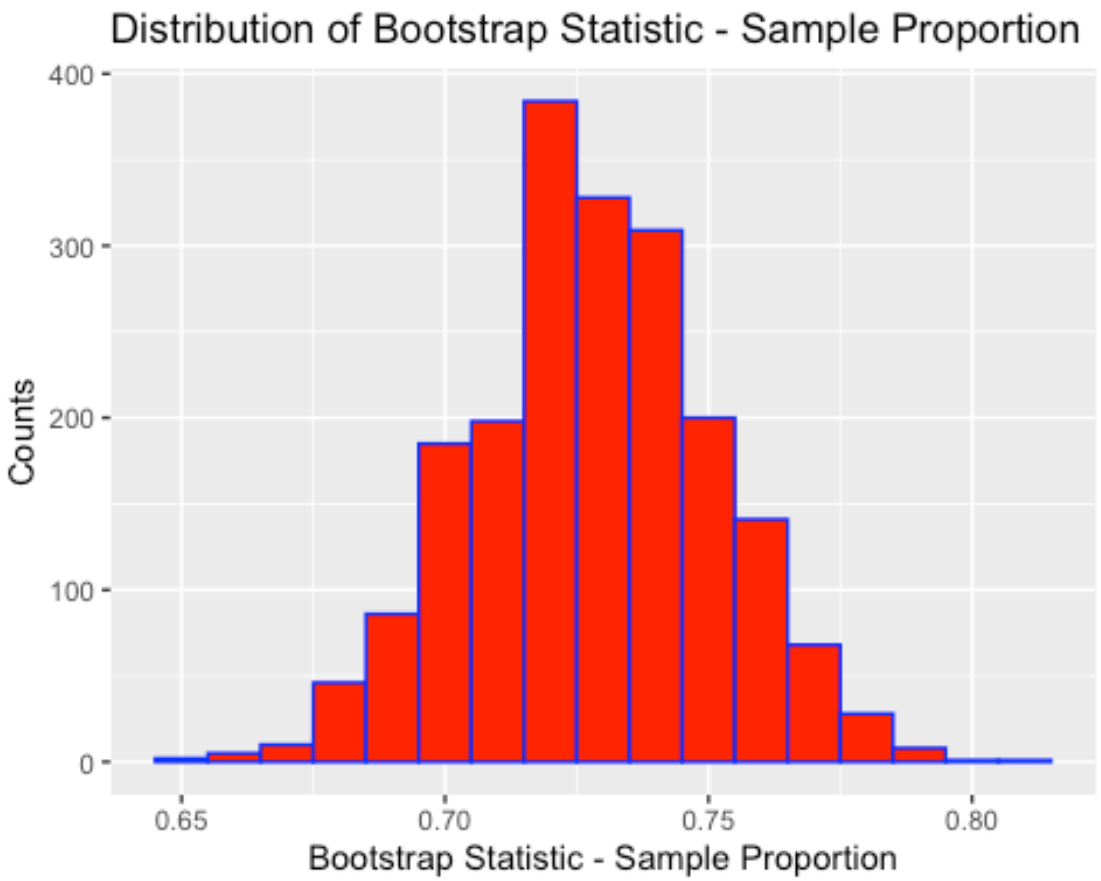
You wish to estimate $p_{Uni} - p_{HS}$, the difference between the proportion of all university-educated Canadians who disagree that the science of vaccinations isn't clear and the proportion of all Canadians who's highest level of completed education is high school who believe the same. You wish to have 95% confidence in your result. Think about the code you created to generate the bootstrap distributions on parts (a) and (b). Modify your code to create a distribution of the bootstrap statisic $\hat{p}_{HS} - \hat{p}_{Uni}$.

```r
# Distribution of the proportion of high school educated respondents who
disagree
hs_vector3 <- c(rep(0, 670 - 348), rep(1, 348))
bootstrap_6c = do(2000) * mean(resample(hs_vector3, 670))
bootstrap_6c %>%
  ggplot(aes(x = mean)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.01) +
```

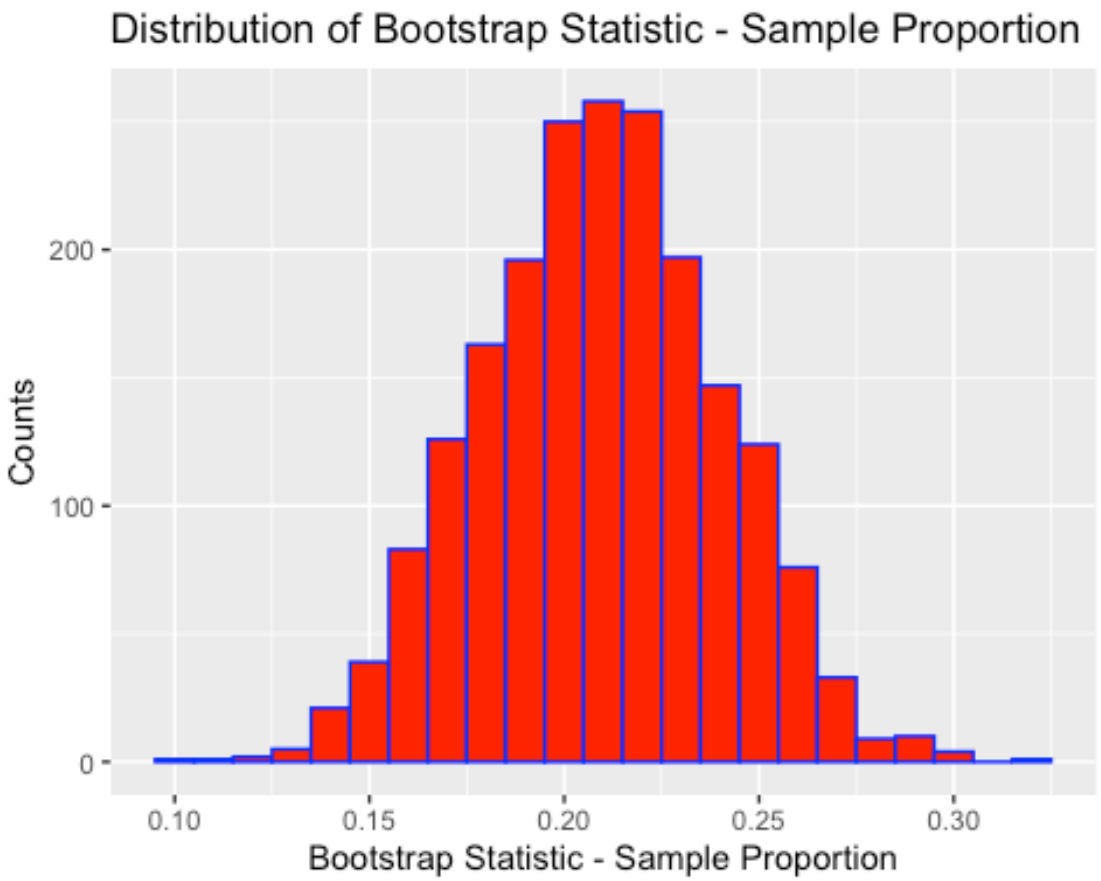
```
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```




```
# Distribution of the proportion of university educated respondents who
disagree
hs_vector4 <- c(rep(0, 376 - 274), rep(1, 274))
bootstrap_6c2 = do(2000) * mean(resample(hs_vector4, 376))
bootstrap_6c2 %>%
  ggplot(aes(x = mean)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.01) +
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```




```r
# Combining these together to create phat uni - phat hs
bootstrap_6c3 = do(2000) * (mean(resample(hs_vector4, 376)) -
mean(resample(hs_vector3, 670)))
bootstrap_6c3 %>%
  ggplot(aes(x = result)) +
  geom_histogram(col = 'blue', fill = 'red', binwidth = 0.01) +
  xlab("Bootstrap Statistic - Sample Proportion") +
  ylab("Counts") +
  ggtitle("Distribution of Bootstrap Statistic - Sample Proportion")
```



d.

Consider your finding in part (c). Compute the 95% bootstrap interval for $p_{HS} - p_{Uni}$. What can you infer from your result? Does a proportion of person with at most a high school education who disagree the science around vaccinations isn't clear greater than the similar proportion of persons with at least an undergraduate university degree? Justify your conclusion.

```
hs_uni_disagree_95 <- qdata(~result, c(0.025, 0.975), data = bootstrap_6c3)
cat("The lower bound of the 95% confidence interval is",
    round(hs_uni_disagree_95$quantile[1], digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(hs_uni_disagree_95$quantile[2], digits = 2),
    "")
```

```
## The lower bound of the 95% confidence interval is 0.15 and the upper bound
of the 95% confidence interval is 0.27
```

$$0.15 \leq p_{Uni} - p_{HS} \leq 0.27$$

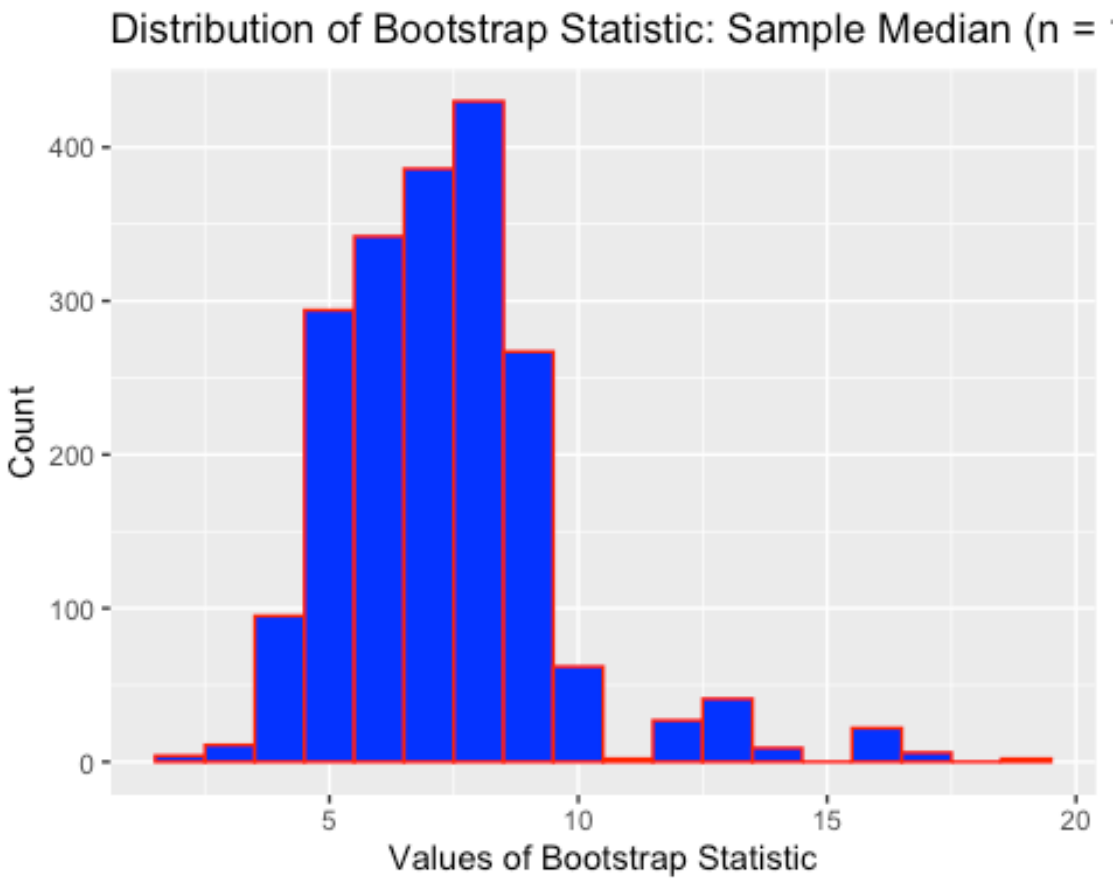
From the result, we can infer that the probability of people with a university education responding that they disagree that the science isn't clear around vaccinations is higher than the probability of people with a high school education. This is because the lower bound of

the 95% confidence interval calculated above is higher than 0. In addition, we can say the probability that people disagree that the science isn't around vaccinations is 15% to 27% higher in university students than high school students.

Question 7

Refer to the data encountered in Question 4 of this assignment. Create a bootstrap distribution of the sample median $\tilde{X}$, using the same number of replications as you did in Question 4. From this find a 99% confidence interval for the population median, $\tilde{\mu}$. Interpret your finding in the context of these data.

```
bootstrap_7 = do(ntimes_4a) * median(resample(LC50_sample, n = nsize_4a))
bootstrap_7 %>%
  ggplot(aes(median)) +
  geom_histogram(col='red', fill='blue', binwidth=1) +
  xlab("Values of Bootstrap Statistic") +
  ylab("Count") +
  ggtitle("Distribution of Bootstrap Statistic: Sample Median (n = 12)")
```




```
LC50_median_99 <- qdata(~median, c(0.005, 0.995), data = bootstrap_7)
cat("The lower bound of the 99% confidence interval is",
```

```
    round(LC50_median_99$quantile[1], digits = 2),
    "and the upper bound of the 99% confidence interval is",
    round(LC50_median_99$quantile[2], digits = 2),
    "")
```

```
## The lower bound of the 99% confidence interval is 3 and the upper bound of
the 99% confidence interval is 16
```

The 99% bootstrap interval for $\tilde{\mu}$ is: $4 \leq \tilde{\mu} \leq 16$

In order to have 99% confidence in the population median $\tilde{\mu}$, the range needs to be relatively large. This is likely due to the small sample size.

Question 8

The most recent poll taken of the Canadian voting preference of their local MP for the various national political parties found that of $n = 1003$ *randomly chosen* ***decided*** *Canadian voters, 346 would support their Conservative Party MP-candidate, 341 would support their Liberal Party MP-candidate, 126 would support their NDP MP-candidate, 106 would support their Green Party MP-candidate, and 84 would support their Bloc Quebecois MP-candidate.*

a.

Compute the 95% confidence interval for p, the proportion of all Canadians that will vote for their respective NDP MP-candidate.

```
CPC <- 346
LPC <- 341
NDP <- 126
GPC <- 106
BQ <- 84
total <- 1003

NDP_stats <- binom.confint(NDP, total, conf.level = 0.95, method = "agresti-
coull")
cat("The lower bound of the 95% confidence interval is",
    round(NDP_stats$lower, digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(NDP_stats$upper, digits = 2),
    "")
```

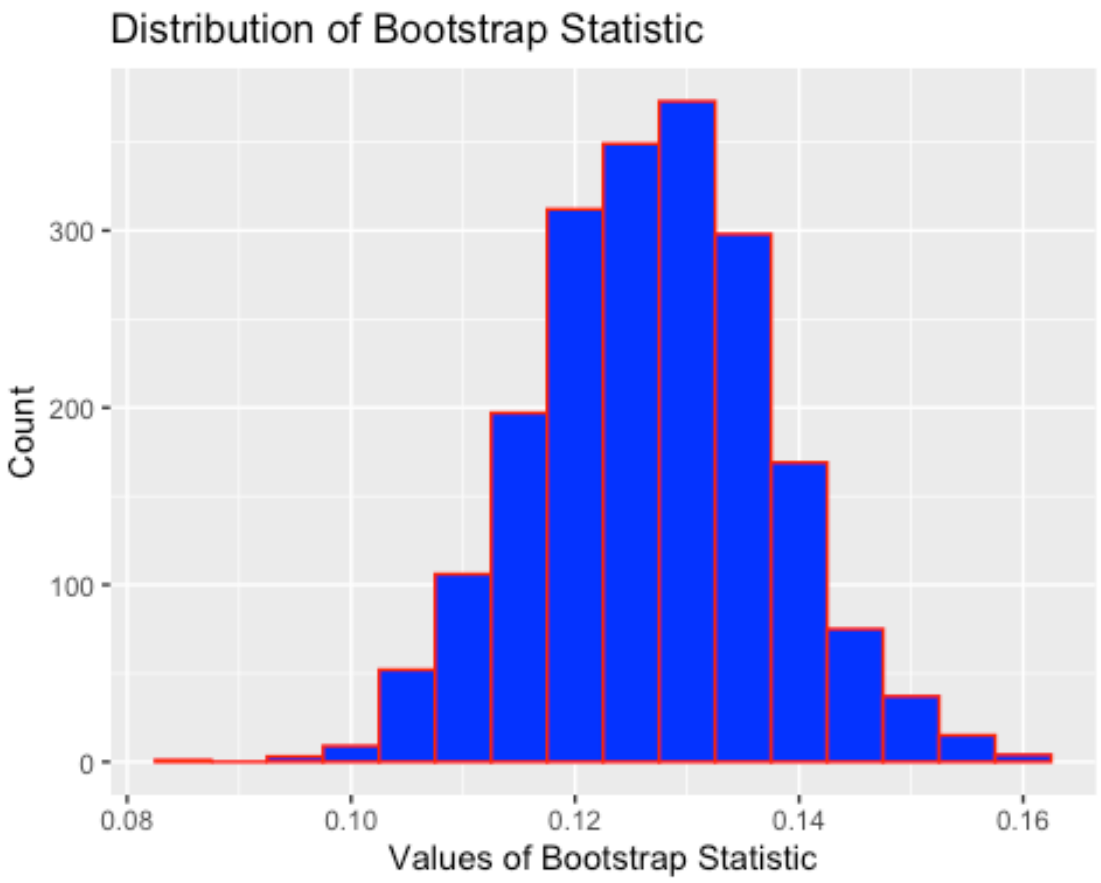
```
## The lower bound of the 95% confidence interval is 0.11 and the upper bound
of the 95% confidence interval is 0.15
```

$$0.11 \leq p_{NDP} \leq 0.15$$

b.

Consider the bootstrap statistic $\tilde{p} = \frac{X_{NDP}+2}{n+4}$. *Write the R code that will generate a bootstrap distribution for* $\tilde{p}$. *Use 2000 as the number of replications/iterations.*

```
ntimes_8b <- 2000
ptilde <- function(x){
  (x + 2) / (total + 4)
}
NDP_vector <- c(rep(0, (total - NDP)), rep(1, NDP))
bootstrap_8b <- do(ntimes_8b) * ptilde(sum(resample(NDP_vector, n = total)))
bootstrap_8b %>%
  ggplot(aes(ptilde)) +
  geom_histogram(col='red', fill='blue', binwidth=0.005) +
  xlab("Values of Bootstrap Statistic") +
  ylab("Count") +
  ggtitle("Distribution of Bootstrap Statistic")
```



c.

From your result in part (b), compute a 95% confidence interval for p.

```r
NDP_ptilde_95 <- qdata(~ptilde, c(0.025, 0.975), data = bootstrap_8b)
cat("The lower bound of the 95% confidence interval is",
    round(NDP_ptilde_95$quantile[1], digits = 2),
    "and the upper bound of the 95% confidence interval is",
    round(NDP_ptilde_95$quantile[2], digits = 2),
    "")
```

```
## The lower bound of the 95% confidence interval is 0.11 and the upper bound
of the 95% confidence interval is 0.15
```

$$0.11 \leq p_{NDP} \leq 0.15$$

d.

Consider your results in parts (a) and (c). What can you infer about the proportion Canadians voting in the upcoming 2019 Federal Election that will vote for their Liberal Party MP-candidate?

Nothing - our results in parts (a) and (c) were looking primarily at the NDP and since there are others parties other than the Liberal Party and the NDP, we cannot use the results effectively. If the purpose of this question was to infer about the proportion of Candians voting for their NDP candidate, then we can infer that between ~11 to 15% of all Canadians will vote for the NDP with 95% confidence.