# Assignment2

Sean Anselmo

2024-01-24

## Question 1

Billy 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)= choose(6,x)*choose(43,6-x)/choose(49,6) x=0,1,2,3,4,5,6. E(X) = X = 36/49 = 0.7347 SD(X) = X = 0.75998 0.76**

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.

```
#Calculate the odds he draws P(x=0) and then subtract that from 1.
PxIs0 = (choose(6,0)*choose(43,6))/choose(49,6)
AtLeastOne = 1 - PxIs0^52 #Do it 52 times for a year
```

**Answer to Q1** Billy's chances of getting 0 numbers on his card for 52 weeks straight is nearing 0, meaning he was correct in assuming that he will on average get at least one number on his card in a years time. This is compounded by his expected value being 0.7347, and his standard deviation being 0.76 This allows us to use the Central Limit Theorem to assume that x is very likely to fall at or above 1 for at minimum one week in the 52 week span. With that being said, E(X) is less than one meaning he is unlikely to get a ticket each week.

## Question 2

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 $\bar{X}$Boot,LC50. Use 2000 "bootstraps" in your work, and display the distribution.

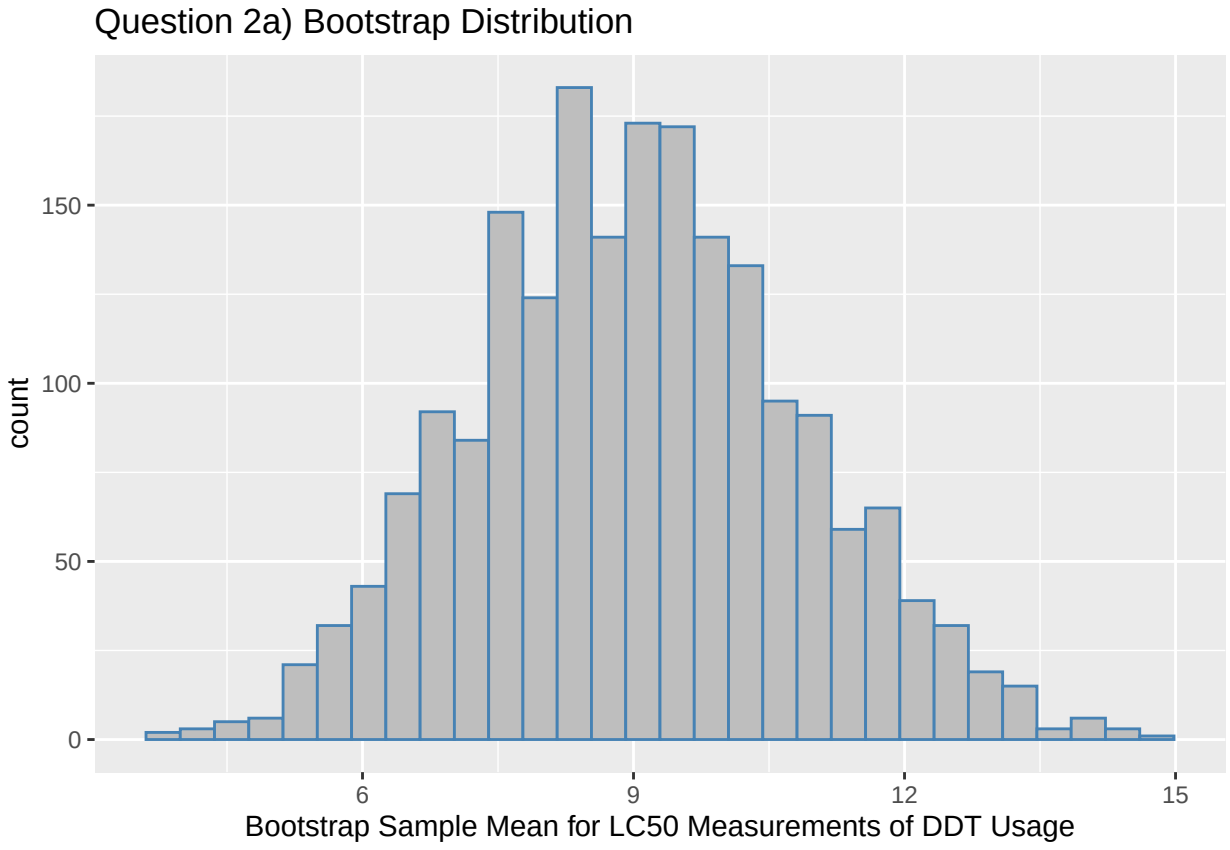
b)From your result in (a), compute the 95% bootstrap (percentile) confidence interval for LC50, the mean LC50 measurement for DDT.

c)Repeat your estimation of LC50, using the "other" confidence interval covered in Data 602. In the context of these data, interpret the meaning of the confidence interval. State any conditions/assumptions that are required in the computation of this confidence interval.

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.

```
data <- c(16,5,21,19,10,5,8,2,7,2,4,9)

set.seed(45)
Q2bootstrap <- do(2000)*mean(resample(data,replace = TRUE))

Q2bootstrap_data_frame <- data.frame(means = unlist(Q2bootstrap))
ggplot(Q2bootstrap_data_frame, aes(x=means))+
  geom_histogram(color = 'steelblue', fill = 'grey')+
  xlab("Bootstrap Sample Mean for LC50 Measurements of DDT Usage")+
  ggtitle("Question 2a) Bootstrap Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
Q2b = quantile(Q2bootstrap$mean,c(0.025,0.975))

ParaMean = mean(data)
ParaSD = sd(data)
```

```
length = length(data)

ParaSem = ParaSD/sqrt(length)
t = qt(0.975,length-1)

moe = t*ParaSem

ParaUpper = ParaMean + moe
ParaLower = ParaMean - moe
```

**Answer for Question 2b** The upper and lower bounds for out 95% confidence interval for the LC50 Measurement of DDT is **5.6666667, 12.5854167**. This means we are 95% sure the mean LC50 values lie between 5.50 and 12.75.

**Answer for Question 2c** The parametric approach yielded an upper limit of **13.0818597** and a lower limit of **4.9181403**. We are 95% certain the mean of LC50 use was between 4.08 and 4.92. For the parametric approach we assumed the data to be normally distributed, and that the sampling was random and independant.

**Answer for Question 2d** I would use the answer from 2b, the bootstrap approach. This is because since it is a small population, and we do not know how the samples are distributed or the standard deviation. The bootstrap approach makes less assumptions about the data as a whole, which is why it is the more robust apporach to determining the confidence interval.

## Question 3

Does one's educational level influence their opinion about vaccinations? A recent Angus Reid 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** respondents 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 education. Of these, 274 disagreed.

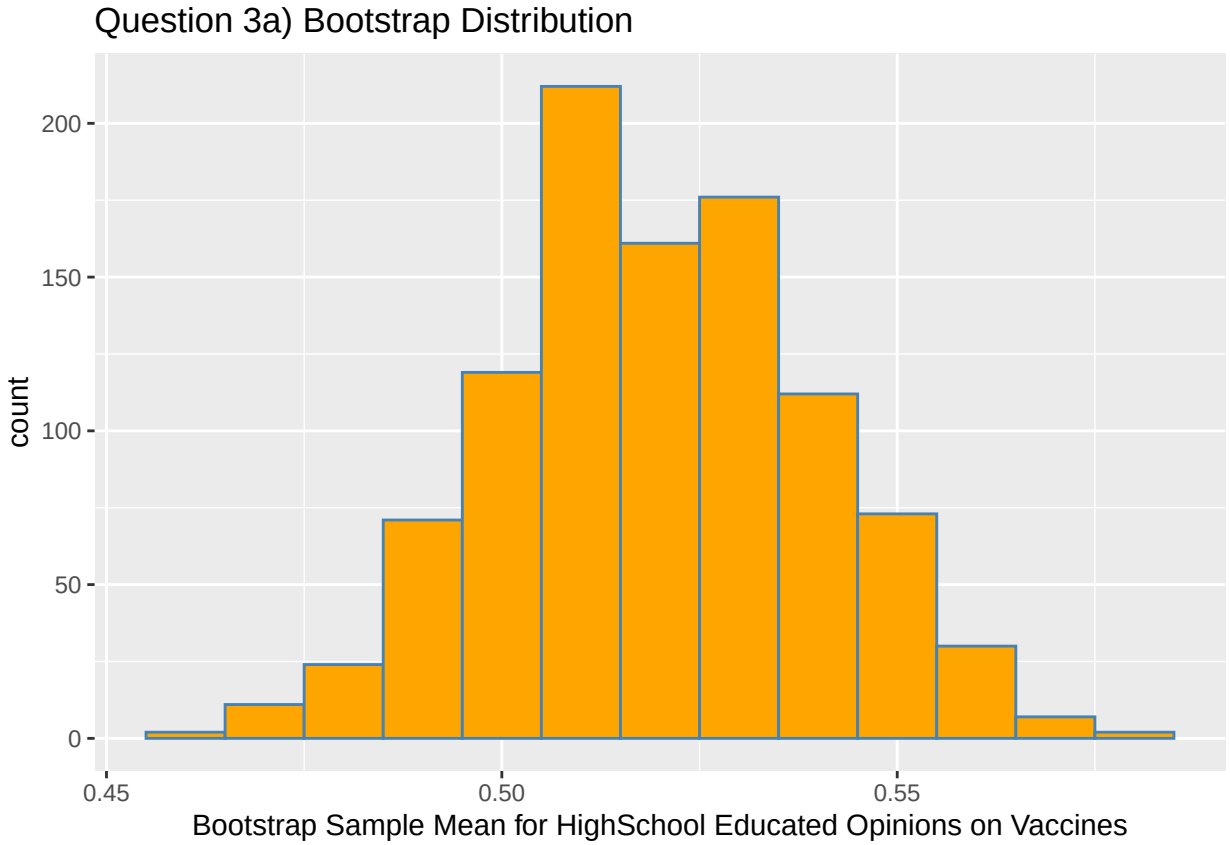
a)Consider the population consisting of all persons, who's highest level of education was high school or less and the bootstrap statistic pˆBoot,HS . Using 1000 iterations/replications, create a bootstrap distribution of pˆHS. Display your distribution.

b)Now consider a different population that consists of all persons who's highest level of education was at least an undergraduate degree. Repeat part (a), creating a bootstrap distribution for pˆBoot,Uni. (Again, display your distribution).

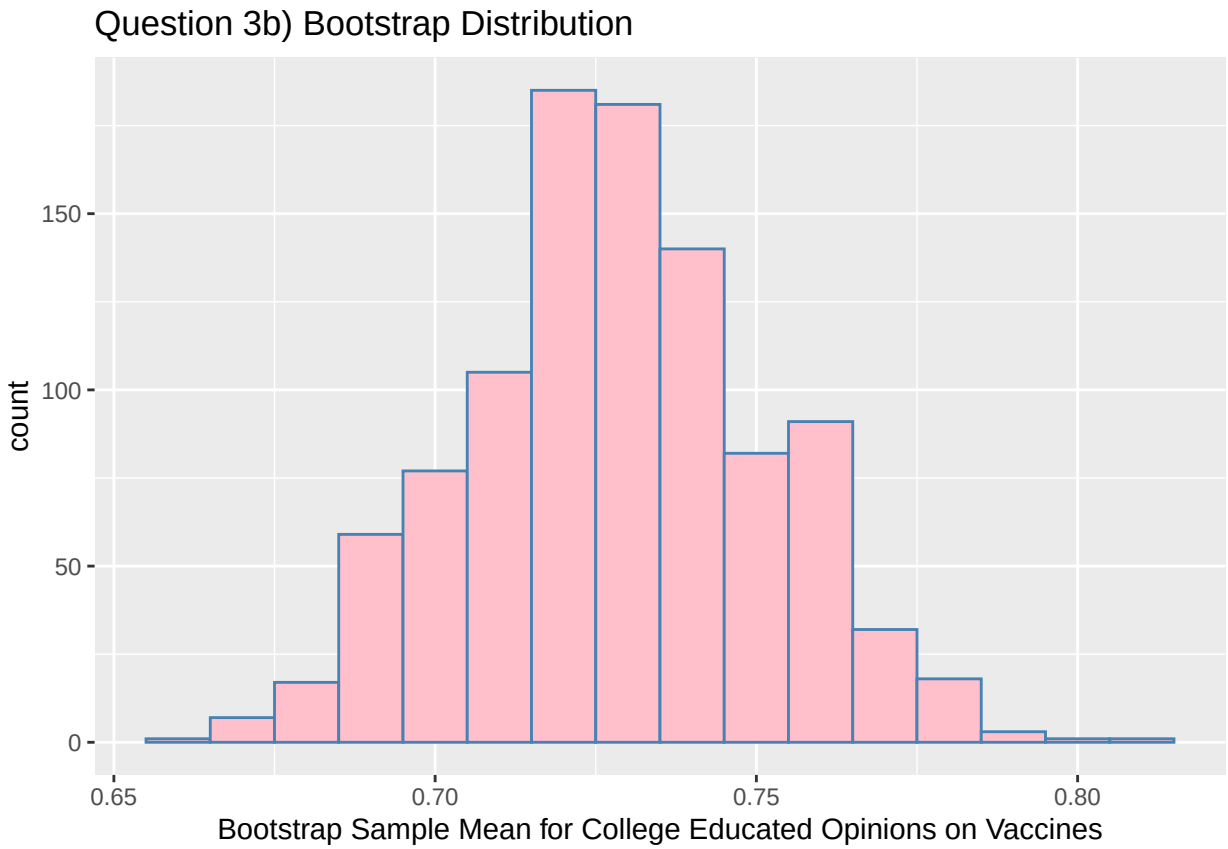
c)You wish to estimate pUni−pHS, the difference between the proportion of all university-educated Canadians who disagree that the science of vaccinations isn't clear and the proportion of all Canadians who's highest level of completed education is high school who believe the same. You wish to have 95% confidence in your result. Think about the code you created to generate the bootstrap distributions on parts (a) and (b). Modify the code to you created in parts (a) and (b) to create a distribution of the bootstrap statistic pˆUni−pˆHS.

d)Consider your finding in part (c). Compute the 95% bootstrap percentile confidence interval for pUni−pHs. From your result, does the proportion of persons 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? Write a paragraph that supports your answer.

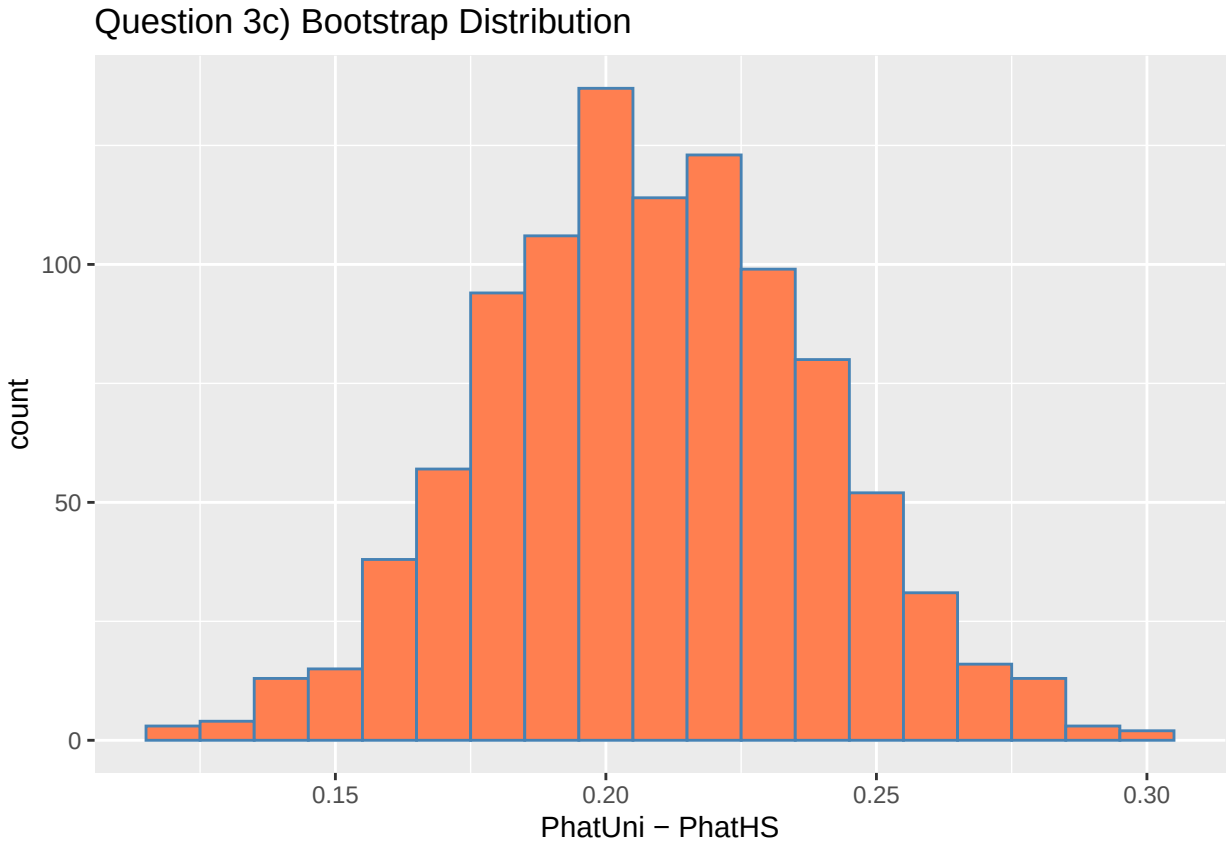
```
HighSchool = 670
HighSchoolDisagree = 348

data <- c(rep(1, HighSchoolDisagree), rep(0, HighSchool-HighSchoolDisagree))
Q3aBootstrap <- do(1000)*mean(resample(data,replace = TRUE))

ggplot(Q3aBootstrap, aes(x=mean))+
  geom_histogram(color = 'steelblue', fill = 'orange', binwidth = 0.01)+
  xlab("Bootstrap Sample Mean for HighSchool Educated Opinions on Vaccines")+
  ggtitle("Question 3a) Bootstrap Distribution")
```




```
dataB <- c(rep(1, 274),rep(0,376-274))
Q3bBootstrap <- do (1000)*mean(resample(dataB, replace = TRUE))

ggplot(Q3bBootstrap, aes(x=mean))+
  geom_histogram(color = 'steelblue', fill = 'pink', binwidth = 0.01)+
  xlab("Bootstrap Sample Mean for College Educated Opinions on Vaccines")+
  ggtitle("Question 3b) Bootstrap Distribution")
```




```
meansHS <- unlist(Q3aBootstrap$mean)
meansCollege <- unlist(Q3bBootstrap$mean)

Q3cBootstrap <- meansCollege - meansHS

Q3cBootstrap_data_frame <- data.frame(mean = Q3cBootstrap)


ggplot(Q3cBootstrap_data_frame, aes(x=mean))+
  geom_histogram(color = 'steelblue', fill = 'coral', binwidth = 0.01)+
  xlab("PhatUni - PhatHS")+
  ggtitle("Question 3c) Bootstrap Distribution")
```




```
Q3d = quantile(Q3cBootstrap,c(0.025,0.975))
```

**Answer for Question 3d:** The 95% confidence interal for pUni - pHS is 0.15, 0.267. Since the interaval is above 0, we can determine that there is a significant difference between the two groups. This difference is represented by the proportion of individuals with at most a highschool education who disagree with the clairty of vaccine science (pHS) is between 15% and 26.7% greater than the proportion of individuals with at least a university degree who disagree. This means we are 95% sure there is an actual significant difference between these two popultions, and the difference observed is not due to random chance.

## Question 4

Nanos research recently completed a survey of **n=1000** Canadians aged 18 years of age or older, asking each "what is your most important national issue of concern?" 163 responded "Inflation", 149 responded "Environment", 131 responded "Jobs/Economy". Those were the Top Three.

a)Compute a 95% confidence interval for **pInflation** , the proportion of all Canadians aged 18 years or older for whom "Inflation" is the most important national concern.

b)Similar to your work in Question 4(a), create the distribution of the bootstrap statistic **p^Boot,Inflation** and a 95% bootstrap percentile confidence interval for **pInflation**.

c)A similar survey of Canadians in August 2023 - a little over a month ago - suggested that the proportion of all Canadians who identified "Inflation" as the most important national concern was **pInflation,Aug_23=0.13**

d)From your results in (a) and (b), can you infer that the proportion of all Canadians who believe "Inflation" is the most important national issue has increased since August of this year? Why or why not? Ensure you

invoke a statistical justification.

```
n = 1000
nInflation = 163
nEnviroment = 149
nEconomy = 131

pHatInflation = nInflation/n
seInflation = sqrt(pHatInflation*(1-pHatInflation)/n)

z = qnorm(0.975)

moe = z*seInflation

lower = (pHatInflation-moe)*100
upper = (pHatInflation+moe)*100

set.seed(45)
dataQ4 = c(rep(1, nInflation), rep(0, n-nInflation))
Q4bootstrap = do(1000)*mean(resample(dataQ4, replace = TRUE))

Q4B = quantile(Q4bootstrap$mean, probs = c(0.025,0.975))
```

**Answer for Question 4a)** The lower bound for the 95% confidence interval for Inflation concerns is **14.0106899%**, and the upper bound is **18.5893101%**

**Answer for Question 4b)** 0.141, 0.187

**Answer for Question 4c)** We are 95% sure the number of Canadians who identify Inflation as their primary concern has risen since August. This is because the pInfaltion(Aug 2023) was **13%**, which is lower than our lower bound of **14%**. **13%** also does not fall within our bootstrap method lower bound of **14.1**, further coorberating our reasoning. Since it does lie within our confidence interval, we can make this claim with 95% certainty.

## Question 5

A national survey 4 of **n=399** "Gen Z"-ers - someone who is born in the years 1996 - 2010 (inclusive) was taken. Each was then asked the following question: "If a federal election were held tomorrow, which one of the following parties would you vote for in your constituency?" The results?

128 responded "Conservative" (Conservative Party of Canada) 96 responded "Liberal" (Liberal Party of Canada) 104 responded "NDP" (New Democratic Party of Canada)

Respondents were provided with a few more "closed options", including the Bloc Quebecois, People's Party, and Green Party.

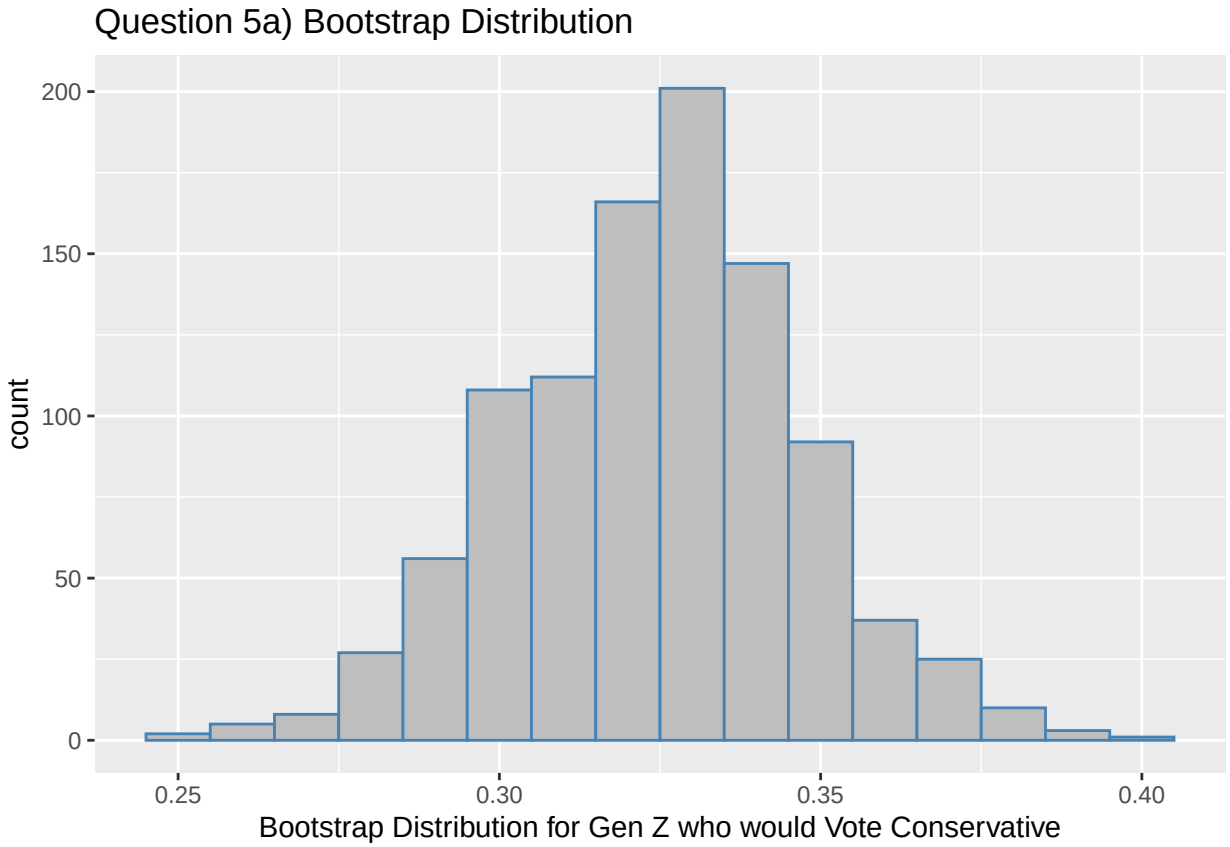
a)Compute the 95% confidence interval for **pCon**, the proportion of all Gen Z-ers in Canada that will vote for their respective Conservative Member of Parliament candidate/constituency, in an election were "held tomorrow".

b)Consider the bootstrap statistic **p˜Con=XCon+2/(399+4)**. Write the R code that will generate a bootstrap distribution for **p˜Con**. Use 1000 as the number of replications/iterations.

c)From your result in part (b), compute a 95% bootstrap confidence interval for **pCon**.

d)Consider your results in parts (a) and (c). Compare the two results. If you had to pick one as the "best" estimate for the unknown value of **pCon**, which one would you select? Provide a justification for your choice.

```r
n = 399
nLiberal = 96
nConservative = 128
nNDP = 104

pHatCon = nConservative/n
seCon = sqrt(pHatCon*(1-pHatCon)/n)

z = qnorm(0.975)

moe = z*seCon

lowerCon = (pHatCon-moe)*100
upperCon = (pHatCon+moe)*100

#Code Answer for Q5B
set.seed(45)
dataQ5 <- c(rep(1, nConservative), rep(0, n - nConservative))
Q5bootstrap <- replicate(1000, {
  sample_mean <- mean(resample(dataQ5, replace = TRUE))

  (sample_mean * n + 2) / (n + 4)
})

Q5bootstrap_data_frame = data.frame(means = unlist(Q5bootstrap))

Q5C = quantile(Q5bootstrap_data_frame$means, probs = c(0.025, 0.975))


ggplot(Q5bootstrap_data_frame, aes(x=means))+
  geom_histogram(color = 'steelblue', fill = 'grey', binwidth = 0.01)+
  xlab("Bootstrap Distribution for Gen Z who would Vote Conservative")+
  ggtitle("Question 5a) Bootstrap Distribution")
```




**Answer for Question 5a)** The lower bound for the 95% confidence interval for Gen Z'ers who would vote conservative is **27.5000645%**, and the upper bound is **36.6603365%**

**Answer for Question 5c)** 0.2779156, 0.369727

**Answer for Question 5d)** We should use the bootstrap approach from part 5c. There is a sufficient population number, and we are not sure of the distribution to use the parametric approach. Unlike in Question 2, we have a sufficient n value and therefore we should use the bootstrap approach, but the decision is close.