STA130H1S – Fall 2022

Problem Set 6

() and STA130 Professors

Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on October 20 by 5:00 p.m. ET.

```
library(tidyverse)
```

Part 1: Bootstrap Confidence Intervals

Question 1: Driving on the "Right" side of the Road

In this question, you will explore data about cars drive on the right or left side of the road in different countries. World Standards' list of left driving countries shows that 86 of all 270 countries in the world drive on the left side of the road.

```
roaddata <- tibble(road_side = c(rep("left", 86), rep("right", 270-86)))
glimpse(roaddata)
```

```
## Rows: 270
## Columns: 1
## $ road_side <chr> "left", "left", "left", "left", "left", "left", "left", "lef~
```

(a) Are the observations in `roaddata` the entire population or a sample from a population?

Roaddata may possiblity contain the entire population due to the presence of every country being present in the analysis

(b) Pipe the `roaddate tibble` into the `slice_sample(n=100)` function to select a random sample of 100 countries. Call this new data `road_sample`. .

```
set.seed(9770) # Set the seed as the last *three* digits of your student number
# code your answer here
roadsample <- roaddata %>% slice_sample(n=100)
```

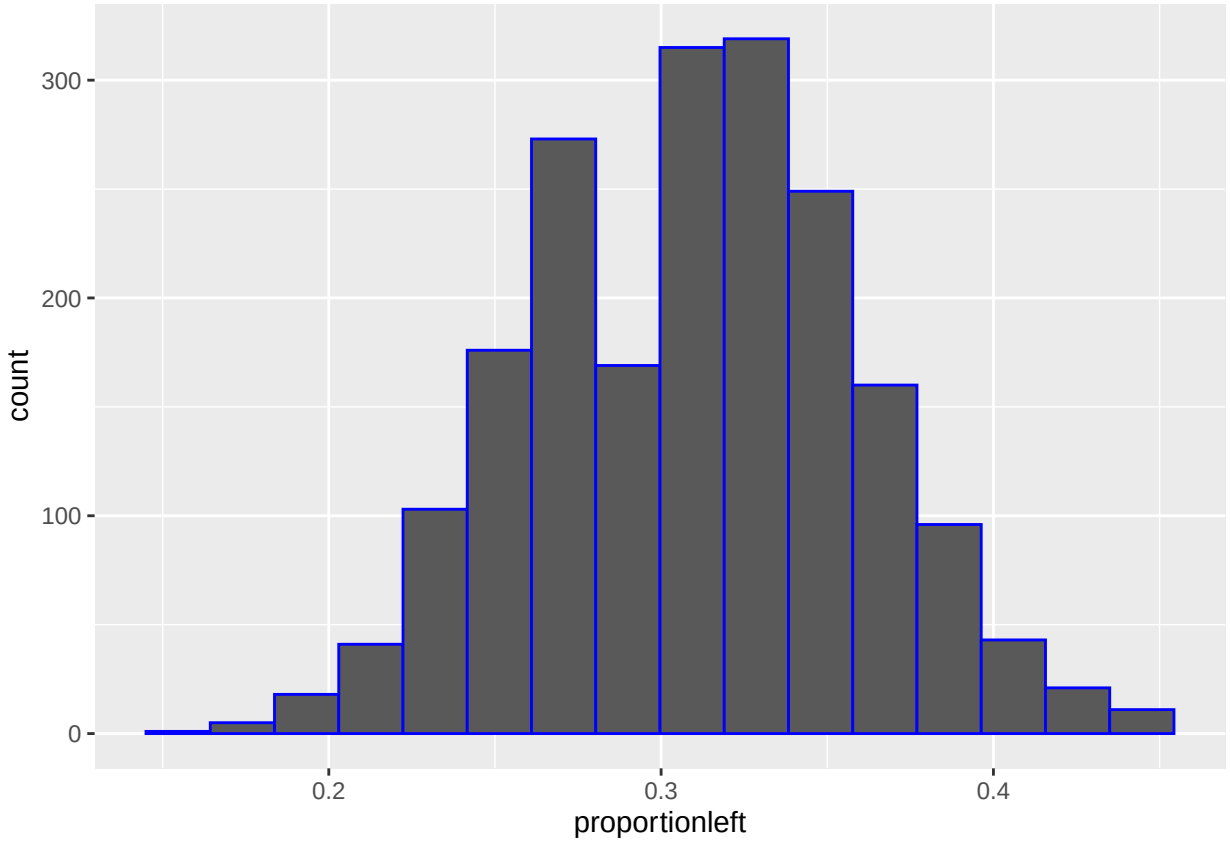
(c) Using the `road_sample` sample you created in (b), simulate 2000 bootstrap samples and calculate the proportion of countries who drive on the left in each of these bootstrap samples. Produce a histogram of the bootstrap sampling distribution of the proportion of regions that drive on the left side.

```
set.seed(123) # Set the seed as the last *three* digits of your student number
# code your answer here
  N <- 2000
```

```
  n <- 100


 set.seed(130)
x <- roadsample
bootstrap_xbars <- rep(NA,N)
for(j in 1:N){
 tmp <- roadsample %>% mutate(binaryresult=case_when(road_side=="left"~1,road_side=="right"~0)) %>%
slice_sample(n=100,replace=TRUE)%>% summarise(x=mean(binaryresult)) %>% as.numeric()
bootstrap_xbars[j] <- tmp

}
table1 <- tibble(proportionleft=bootstrap_xbars)
ggplot(table1,aes(x=proportionleft))+geom_histogram(bins=16,color="blue")
```



(d) Calculate a 90% confidence interval for the proportion of countries/regions which drive on the left based on the bootstrap sampling distribution you generated in (c).

```
table1 %>% summarize(confinterval=quantile(proportionleft,0.95) - quantile(proportionleft,0.05)
)%>% as.numeric()
```

```
## [1] 0.16
```

(e) Assume for the moment that your confidence interval was (27%, 44%). [You own confidence interval is likely different from this based on your random seed]. Indicate whether or not each of the following statements is a correct interpretation of the confidence interval constructed in part (d) and justify your answers.

(A) We are 90% confident that between 27% and 44% of countries/regions in our sample from (b) drive on the left side.

In bootstrap sampling, We are 90% confident that between 27% and 44% of countries/regions in our population from (b) drive on the left side. This is incorrect because it says sample and not population

(B) There is a 90% chance that between 27% and 44% of all countries in the population drive on the left side.

False, bootstrap sampling does not mean the above, it is not the chance that it is 90 % it is just how confident we are that it lies in the interval

(C) If we consider many random samples of 100 countries/regions, and we calculate 90% bootstrap confidence intervals for each sample, approximately 90% of these confidence intervals will include the true proportion of countries/regions in the population who drive on the left side of the road.

True, bootstrap sampling does indicate that how confident we are that it is in the 90% confidence level

(f) If we want to be *more* confident about capturing the proportion of all countries who drive on the left side, should we use a *wider* confidence level or a *narrower* confidence level? Explain your answer.

If we were to make it wider we would have a higher confidence level like from 90 to 95 ,so due to this we will be more confident.

(g) We could carry out an hypothesis test to investigate whether or not countries are equally likely to drive on the right or to the left side of the road. State the NULL hypotheses of such a test.

p1=p2 in which p1 and p2 are the proportion of left and right accordingly

Question 2: Auto Claims

The data set `auto_claims.csv` includes claims paid (in USD) to a sample of auto insurance claimants 50 years of age and older in a specific year. In other words, it represents a 'sample' (the 'original sample') of car insurance claims in that year.

(a) Produce appropriate data summaries (i.e. a summary table and relevant visualization) of paid claims (`PAID`) and comment the shape, centre and spread of this distribution.

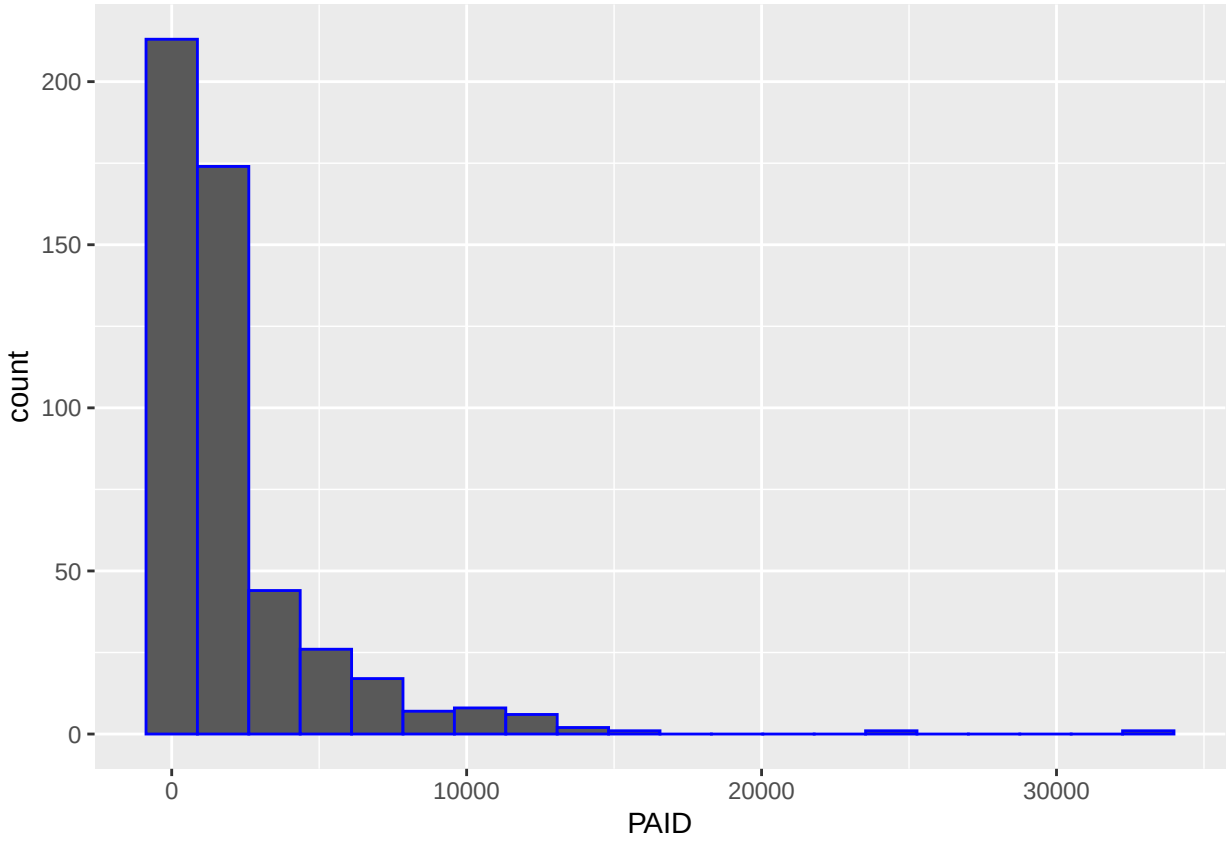
```
PAID_tibble <- read_csv("auto_claims.csv")
glimpse(PAID_tibble)
```

```
## Rows: 500
## Columns: 5
## $ STATE  <chr> "STATE 15", "STATE 15", "STATE 02", "STATE 15", "STATE 04", "ST~
## $ CLASS  <chr> "F6", "F6", "C11", "C11", "C6", "C11", "C6", "C6", "C1", "C11",~
## $ GENDER <chr> "F", "M", "F", "M", "M", "M", "F", "F", "F", "M", "F", "F", "F"~
## $ AGE    <dbl> 95, 95, 92, 91, 91, 90, 90, 90, 90, 88, 88, 88, 88, 88, 88, 88,~
## $ PAID   <dbl> 2384.67, 650.00, 654.00, 3890.07, 295.99, 11756.34, 2402.00, 29~
```

```
PAID_tibble %>% summarise(n=n(), sum_claims=sum(PAID), median_claims=median(PAID),
mean_claims=mean(PAID), var_claims=var(PAID), sd_claims=sd(PAID))
```

```
## # A tibble: 1 x 6
##       n sum_claims median_claims mean_claims var_claims sd_claims
##   <int>      <dbl>         <dbl>       <dbl>      <dbl>     <dbl>
## 1   500   1080067.          1042.       2160.   9608873.     3100.
```

```
ggplot(PAID_tibble,aes(x=PAID))+geom_histogram(bins=20,color="blue")
```



Description of the visualisation– The graph is left skewed and it has a wide spread from 0 to 3000.This spread may contain anonmulous results, however no such conclusions may be made until ideal bin size is investigated. The graph is also centered at 0.

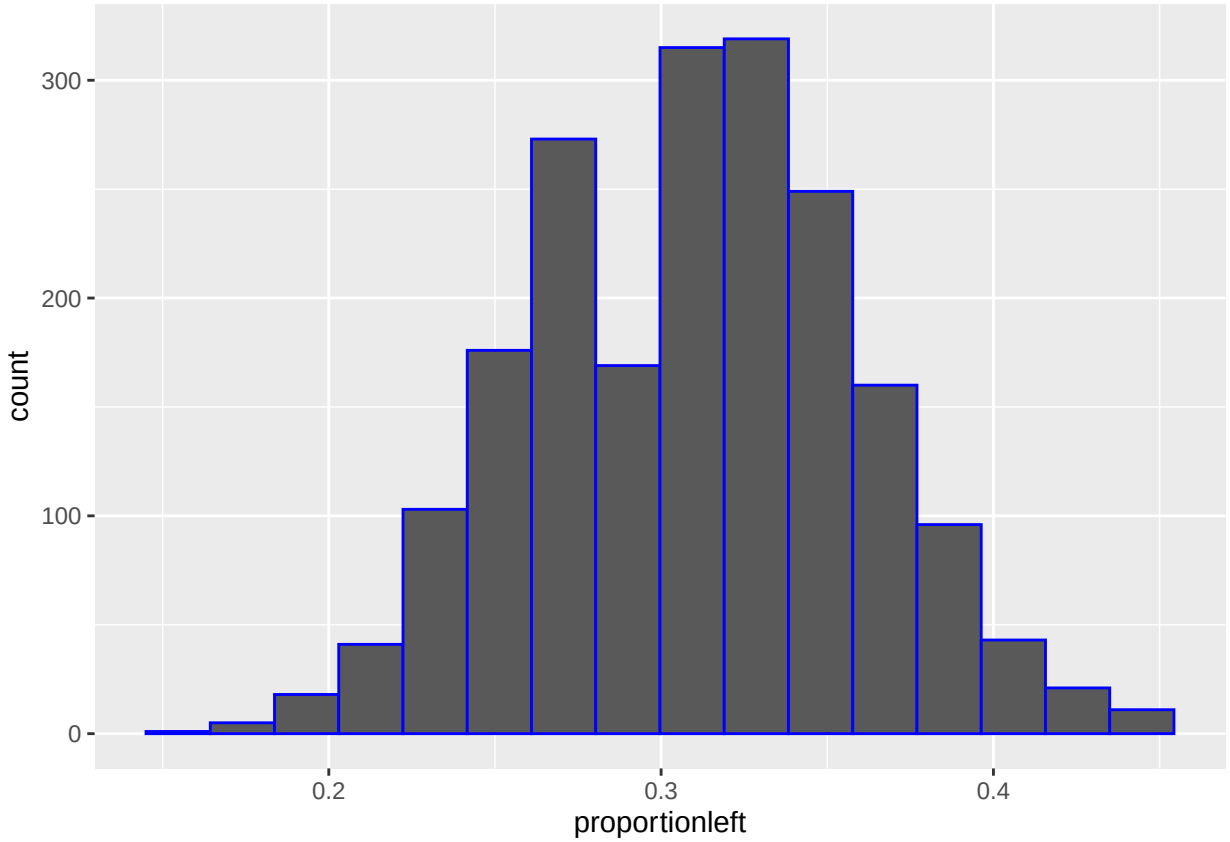
(b) Estimate the sampling distributions of sample *median* of paid claims by taking 1000 samples of size n=500 (to match the sample size in the data) and produce appropriate data summaries.

```
autosample <- PAID_tibble
set.seed(9770)
N<-1000


x <- autosample
bootstrap_xbars <- rep(NA,N)
for(j in 1:N){
tmp <-PAID_tibble %>% slice_sample(n=500,replace=TRUE)%>% summarise(x=median(PAID)) %>% as.numeric()
bootstrap_xbars[j] <- tmp

}
table2 <- tibble(proportionleft=bootstrap_xbars)
```

```
ggplot(table1,aes(x=proportionleft))+geom_histogram(bins=16,color="blue")
```



(c) Using the simulation in part (b) derive a 95% confidence interval for the median of paid claims.

```
table1 %>% summarize(confinterval=quantile(proportionleft,0.975) - quantile(proportionleft,0.025)
)%>% as.numeric()
```

```
## [1] 0.18
```

```
#Calculation check-sample mean +_ confidence level * (sd/sqrt(n))
```

Part 2: Optional

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

Question 3: More Auto Claims

(a) Select 1000 samples of size 20 from the population of claims stored in the `auto_claims_population.csv` data set. Ensure each sample is taken without replacement, so there are no repeated observations within each sample. Compute the mean `age` of claimants for each sample, calculate the overall mean of sample means, and make a histogram of the simulated sample means.

```
# code your answer here
```

(b) Now create a single random sample of 20 car insurance claims and store these 20 observations in a `tibble` called `ages20`. Create a histogram of the sample, and compute the `min()`, `mean()`, `median()`, `max()`, and `sd()` of the `age` variable using the `summarise()` function and include the `n=n()` summary as well.

```
# code your answer here
```

(c) Use R to take bootstrap 1000 samples from `ages20`. Compute the mean age of claimants for each bootstrap sample, calculate the overall mean of the bootstrap sample means, and make a histogram of the simulated sample means.

```
# code your answer here
```

(d) What distribution do the distributions we simulated in (a) and (c) both estimate? Comment on the similarities and differences for a given random seed initializations, and speculate how this could change under different random seed initializations.

REPLACE THIS TEXT WITH YOUR ANSWER

Question 3: Gestation Data

In this question we will look at data from the Child Health and Development Studies. Our data are adapted from the `Gestation` data set in the `mosaicData` package. Birth weight, date, and gestational period were collected as part of the Child Health and Development Studies in 1961 and 1962 for a sample of 400 mothers who had babies in these two years. Information about the baby's parents—age, education, height, weight, and whether the mother smoked—was also recorded.

We will find confidence intervals for parameters related to the distribution of the mother's age, which for this sample is stored in the variable `age`.

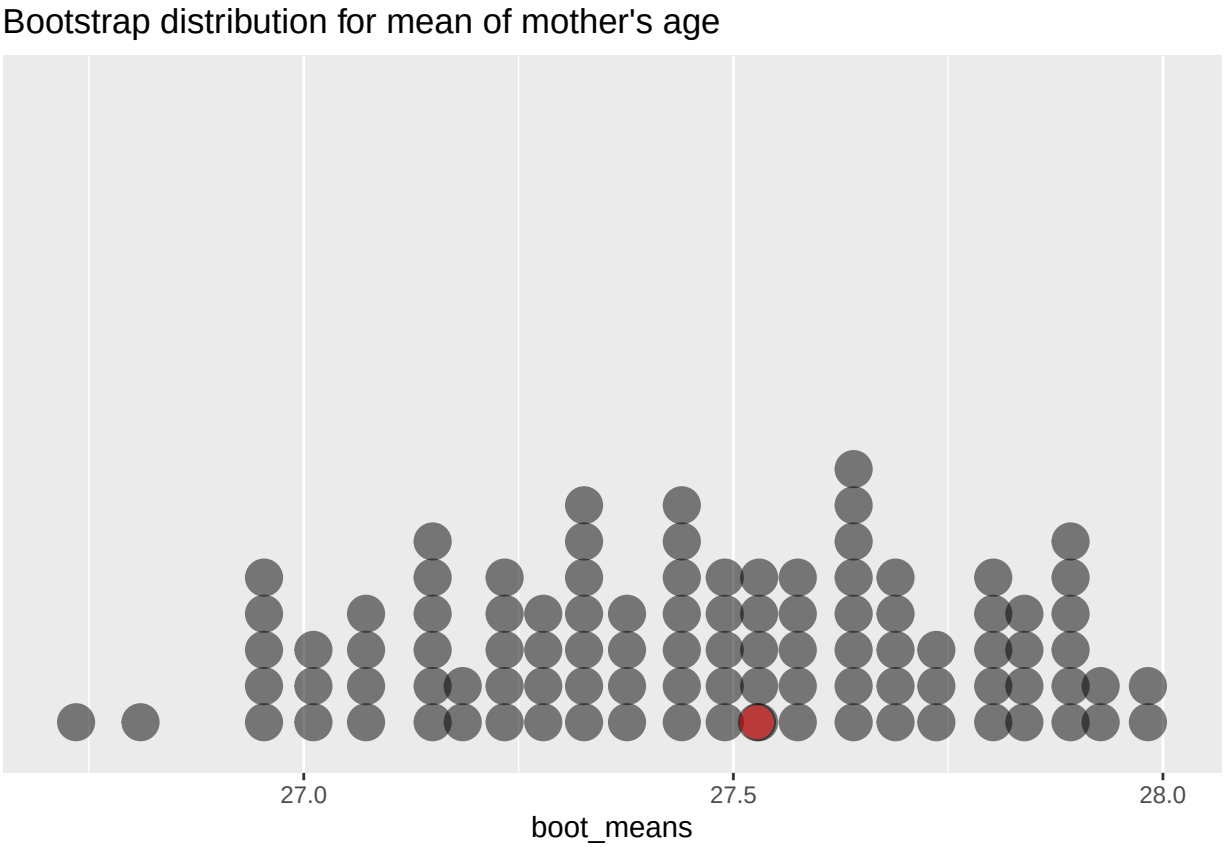
```
Gestation <- read_csv("gestation.csv")
```

(a) Suppose we are interested in how means of random samples of n=400 mothers vary across possible samples of 400 mothers we could take from the population. Explain why it is not possible to use these data (i.e., 'Gestation') to estimate this like we did in Case Study 1, question a).

REPLACE THIS TEXT WITH YOUR ANSWER

(b) The plot below shows the bootstrap distribution for the mean of mother's age for 100 bootstrap samples. The red dot is the estimate of the mean for the first bootstrap sample, and the grey dots are the estimates of the mean for the other 99 bootstrap samples. Explain how the value of the red dot is calculated; then, using the plot, estimate a 90% confidence interval for the mean of mother's age.

REPLACE THIS TEXT WITH YOUR ANSWER




```
## # A tibble: 1 x 6
##     min  mean median   max    sd     n
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <int>
## 1  26.7  27.5   27.5  28.0 0.299   100
```

(c) Use R to find a 99% bootstrap confidence interval for the mean of mother's age. Use 2000 bootstrap samples.

```
# code your answer here
```

(d) Explain why the interpretation *"There is a 99% chance that the true mean of a mother's age at the time this sample was taken is between 26.8 and 28.2 years."* is *INCORRECT*. What is a correct interpretation?

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Use R to find a 95% bootstrap confidence interval for the *median* of mother's age. Use 2000 bootstrap samples.

```
# code your answer here
```

(f) Write a statement interpreting this interval.

REPLACE THIS TEXT WITH YOUR ANSWER