STA130 – Winter 2024

Week 7 Problem Set

N. Moon and J. Speagle

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

IMPORTANT - SETUP

```
# Instructions for how students should define tests
STUDENT_NUMBER <- 1010057028;  # replace 130 by your real student number
```

Question 1

In this question, you will explore data about whether countries (or sub regions) have their road conditions set that vehicles drive on the left or right side of the road (link: https://www.worldstandards.eu/cars/list-of-left-driving-countries/).

Here we can see that there are 270 countries (or states/territories) and 86 of them drive on the left side of the road. Note: this is data that covers all regions in the world.

Here is a data frame with the data from the driving study:

```
# Create a data frame
road_side <- c( rep("left", 86), rep("right", 270-86) )
roaddata <- tibble(road_side)
```

(a) Are the observations in `roaddata` the entire population or a sample from a population? Write one sentence summarizing your answer and replace NULL by "SAMPLE" or "POPULATION" in the code chunk below

```
Q1a_sample_vs_pop <- "POPULATION"
```

The observations in `roadata` is the entire population.

(b) Modify the line of code below to use the `sample_n()` function to select a random sample of different 100 countries/regions. Call this new data `Q1b_road_sample`.

```
set.seed(STUDENT_NUMBER) # Do not change this line of code

Q1b_road_sample <- roaddata %>% sample_n(size=100)
```

(c) In this part, you will simulate the bootstrap sampling distribution and create a visualization to examine it.

```
Q1c_i_repetitions <- 2000 ## Replace 1 by your answer
Q1c_i_simulated_values <- rep(NA, 2000) ## Replace NULL by your answer
```

(i) Using the `road_sample` sample you created in (b), we want to simulate 2000 bootstrap samples and calculate the proportion of countries who drive on the left in each of these bootstrap samples. Start by setting up values for the simulation by storing the correct values in the variables below.

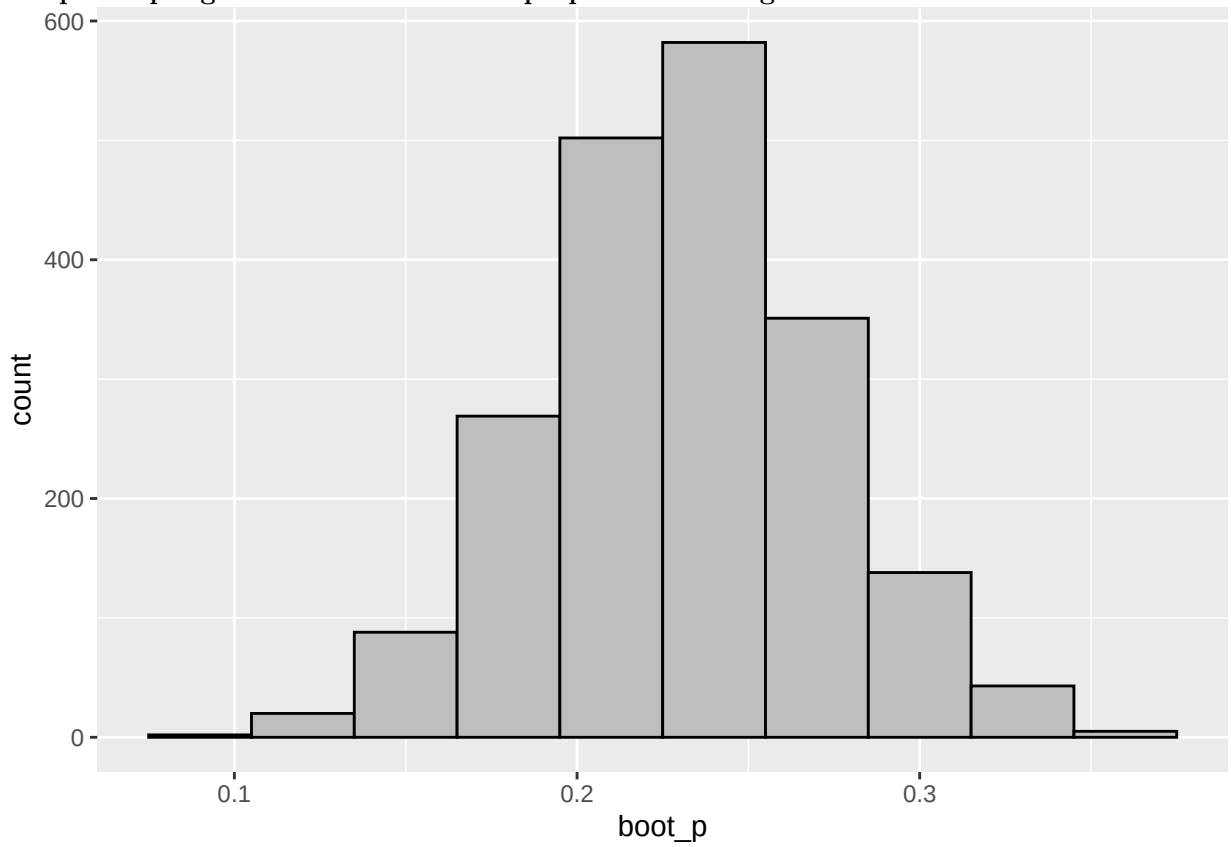
```
road_sample <- Q1b_road_sample   # Do not change this line of code
repetitions <- Q1c_i_repetitions # Do not change this line of code
boot_p <- Q1c_i_simulated_values # Do not change this line of code
set.seed(STUDENT_NUMBER %% 1000) # Do not change this line of code


for (i in 1:repetitions)
{
  boot_samp <- road_sample %>% sample_n(size = 100, replace=TRUE)
  boot_p[i] <- as.numeric(boot_samp %>%
                            filter(road_side == "left") %>%
                            summarize(n()))/100
}
```

```
boot_p <- tibble(boot_p)
```

```
# Produce a histogram of the bootstrap sampling distribution
boot_p %>% ggplot(aes(x=boot_p)) + geom_histogram(color = "black",
                  fill = "gray",
                  bins = 10)
```

(ii) Run the code below to run the simulation, then produce a histogram of the bootstrap sampling distribution of the proportion of regions that drive on the left side.



(d) Use the quantile function to 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). Replace the two NULLs below by your calculated values (they should be the exact values from the quantile object - in other words you shouldn't type them in yourself but instead do like in the example below). Write a sentence interpreting the interval you obtained.

```
# Example: suppose you wanted to calculate the median
median <- quantile(boot_p$boot_p, p=0.5)

Q1d_90_ci_lower_bound <- quantile(boot_p$boot_p, p=0.05)
Q1d_90_ci_upper_bound <- quantile(boot_p$boot_p, p=0.95)
```

(e) 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. (Let's assume the CI was (27%, 44%).) Note: your confidence may well be different from this since we are all using different random seeds in earlier parts of this question. In the code chunk below, please indicate TRUE or FALSE (no quotation marks) for each statement by replacing the NULL with your answer

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

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

(iii) If we considered many random samples of 100 countries/regions, and we calculated 90% confidence intervals for each sample, 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.

```
Q1e_i <- FALSE
Q1e_ii <- FALSE
Q1e_iii <- TRUE
```

(f) If we want to be *more* confident about capturing the true proportion of all countries who drive on the left side, we can increase the confidence level (for example, instead of 95%, we could use 98%). When we increase the confidence level, does this make the confidence interval *wider* or *narrower*? Replace NULL by either "WIDER" or "NARROWER" below, and explain your answer in 1-2 sentences.

```
Q1f <- "WIDER"
```

(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. Our hypotheses would be:

$$H_0 : p = 0.5$$

versus

$$H_A : p \neq 0.5$$

where p is the proportion of countries/regions who drive to the left. Using the sampled data, we would get a P-value of 0.0007. Do this hypothesis test and the confidence interval you produced in (d) tell a similar story? Why or why not?

Yes. The 95% confidence interval i got was from 0.16-0.3, so 0.5 is not within this range. Since there's a P-value of 0.0007, then we have strong evidence to reject the null hypothesis.

Question 2

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.

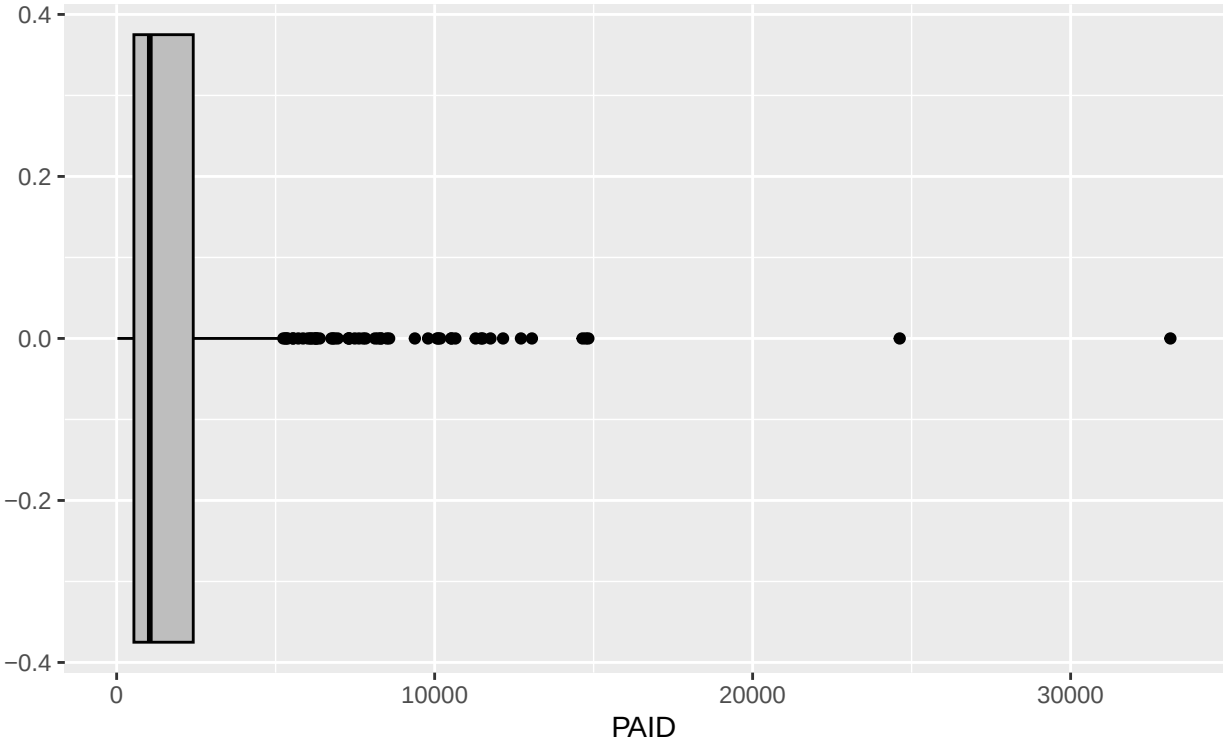
```
AutoClaims <- read_csv("auto_claims.csv")
glimpse(AutoClaims)
```

```
## 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~
```

```
AutoClaims %>% select(PAID) %>% summarize(n = n(), median=median(PAID), iqr = IQR(PAID), range=max(PAID)
```

```
## # A tibble: 1 x 4
##       n median   iqr  range
##   <int>  <dbl> <dbl>  <dbl>
## 1   500  1042. 1866. 33112.
```

```
AutoClaims %>% ggplot(aes(x=PAID)) + geom_boxplot(color = "black",
                fill = "gray",
                bins = 20)
```



(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.

The median amount paid is 1042.345 with a range of $33112.06. The shape is unimodal with a strong right skew.

(b) In this part, you will 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.

```
Q2b_n <- 500 # Replace 1 by your answer
Q2b_repetitions <- 1000 # Replace 1 by your answer
Q2b_vector_for_simulated_values <- rep(NA, 1000) # Replace 1 by your answer
```

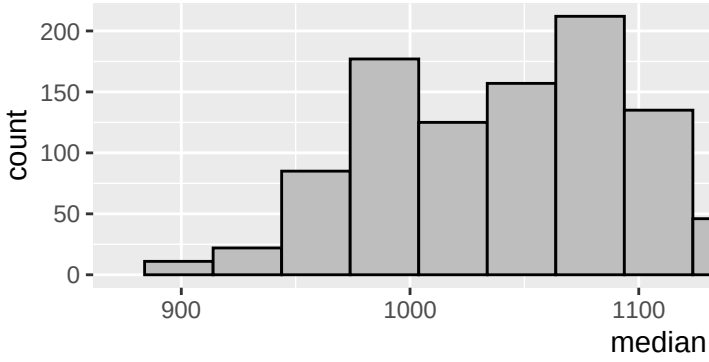
(i) Fill in the values below to set up the simulation, based on what was specified in the question. The values you specify will be used to run the simulation in the next part of this question.

```
n <- Q2b_n                              # Run this line but do not change it
repetitions <- Q2b_repetitions          # Run this line but do not change it
sim <- Q2b_vector_for_simulated_values # Run this line but do not change it
set.seed(STUDENT_NUMBER %% 250)         # Run this line but do not change it

for (i in 1:repetitions)
{
  new_sim <- sample(AutoClaims$PAID ,size = n, replace=TRUE)
  sim_median <- median(new_sim)
  sim[i] <- sim_median
}
sim <- tibble(median = sim)
```

```
# Create your visualization here
sim %>%  ggplot(aes(median)) + geom_histogram(color = "black",
                 fill = "gray",
                 bins = 15)
```

(ii) Run the code chunk below to generate the bootstrap samples, then create a histogram to visualize the estimated bootstrap sampling distribution.



(c) Based on the simulation in part (b), use the quantile function to calculate a 95% confidence interval for the median of paid claims based on samples of size 500. Replace the two NULLs below by your calculated values (they should be the exact values from the quantile object - in other words you shouldn't type them in yourself but instead do like in the example below).

```
# Example: suppose you wanted to calculate the median
median <- quantile(sim$median, p=0.5)

Q2c_95_ci_lower_bound <- quantile(sim$median, p=0.05)
Q2c_95_ci_upper_bound <- quantile(sim$median, p=0.95)
```

Question 3 (Optional, for extra practice)

In this question, you will explore data about whether couples observed kissing in an airport tilt their heads to the left or the right. The data is adapted from a real study, Gunturkun (2003) (link: https://www.nature.com/articles/421711a).

This is the opening of the article:

> "I observed kissing couples in public places (international airports, large railway stations, beaches and parks) in the United States, Germany and Turkey. The head-turning behaviour of each couple was recorded for a single kiss, with only the first being counted in instances of multiple kissing. The following criteria had to be met to qualify: lip contact, face-to-face positioning, no hand-held objects (as these might induce a side preference), and an obvious head-turning direction during kissing. Subjects' ages ranged from about 13–70 years.
>
> Of 124 kissing pairs, 80 (64.5%) turned their heads to the right and 44 (35.5%) turned to the left."

That's kinda creepy

Here is a data frame with the data from the kissing study:

```
# Create a data frame
direction <- c( rep("right", 80), rep("left", 124-80) )
kissdata <- tibble(direction)
```

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

(b) Simulate 1000 bootstrap samples and calculate the proportion of couples who kiss to the left in each of these bootstrap samples. Produce a histogram of the bootstrap sampling distribution of the proportion of people who kiss to the left.

```
set.seed(STUDENT_NUMBER)

# Your code here
```

(c) Calculate a 95% confidence interval for the proportion of people who kiss to the left based on the Bootstrap distribution you generated in (b).

(d) Indicate whether or not each of the following statements is a correct interpretation of the confidence interval constructed in part (c) and justify your answers.

(i) We are 95% confident that between 27% and 44% of kissing couples in this sample tilt their head to the left when they kiss.

(ii) There is a 95% chance that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.

(iii) We are 95% confident that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.

(iv) If we considered many random samples of 124 couples, and we calculated 95% confidence intervals for each sample, 95% of these confidence intervals will include the true proportion of kissing couples in the population who tilt their heads to the left when they kiss.

(e) If we want to be *more* confident about capturing the proportion of all couples who tilt their heads to the left when kissing, should we use a *wider* confidence level or a *narrower* confidence level? Explain your answer.

(f) We could carry out an hypothesis test to investigate whether or not couples are equally likely to tilt their heads to the right or to the left when they kiss. Our hypotheses would be:

$$H_0 : p = 0.5$$

versus

$$H_A : p \neq 0.5$$

where p is the proportion of couples who tilt their heads to the left when they kiss. Using Gunturkun's data, we would get a P-value of 0.003. Do this hypothesis test and the confidence interval you produced in (c) tell a similar story? Why or why not?