STA130 – Winter 2021

Week 6 Problem Set

N. Moon & S. Caetano & Yixing Xu (100734135)

Instructions

How do I hand in these problems for the 11:59 a.m. ET, February 25th deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/555726) by 11:59 a.m. ET, on February 25. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[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?

The observations are samples from a population.

(b) Use the `sample_n()` function to select a random sample of different 100 countries/regions. Call this new data `road_sample`. Set the seed as the last *three* digits of your student number.

```
set.seed(351)

road_sample <- roaddata %>% sample_n(size = 100, replace = FALSE)
```

(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 the seed as the last *three* digits of your student number.

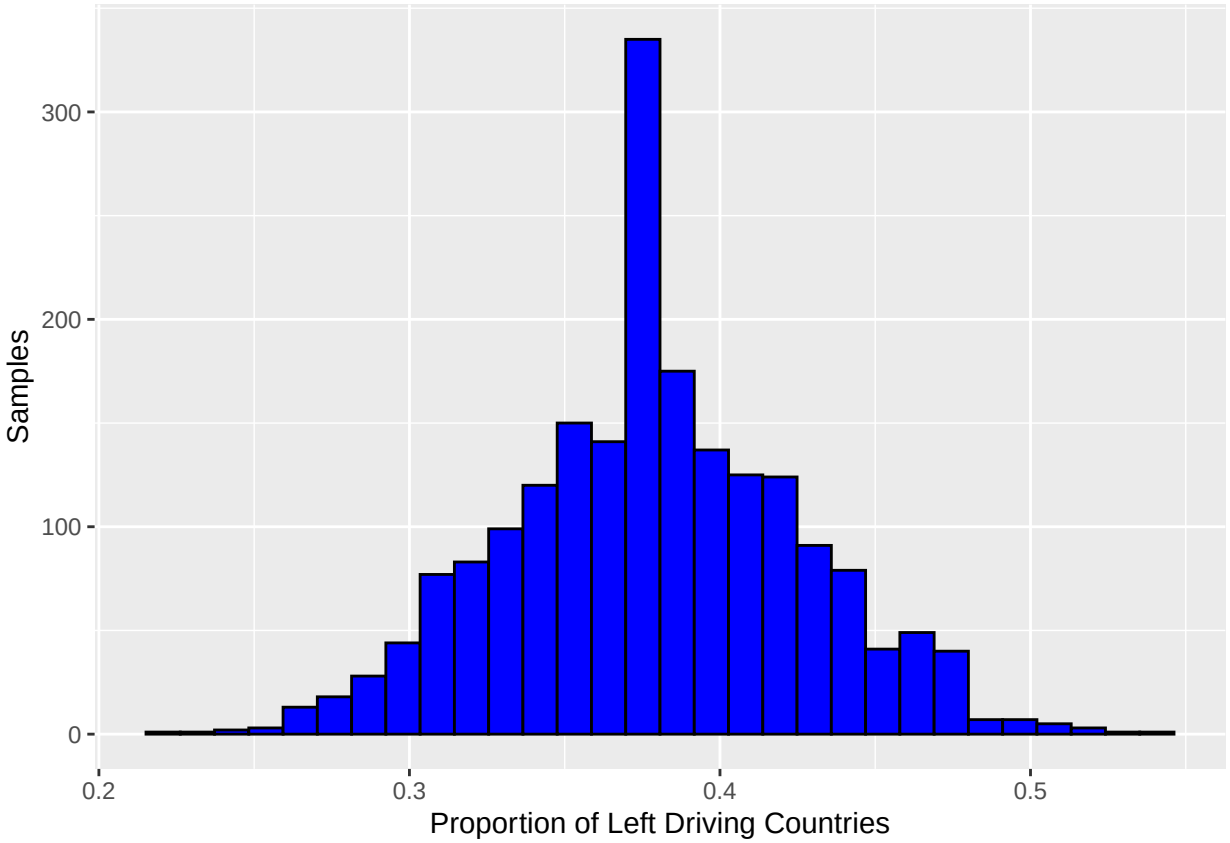
```
set.seed(351)

boot_props <- rep(NA, 2000)
for(i in 1:2000){
  boot_samp <- road_sample %>% sample_n(size = 100, replace = TRUE)
  boot_props[i] <- as.numeric(boot_samp %>%
                              summarise(prop = (sum(road_side == "left")/100)))
  #print(boot_props[i])
}

boot_props <- tibble(props = boot_props)
traceback()
```

```
## No traceback available
```

```
boot_props %>% ggplot(aes(x = props)) + geom_histogram(color = "black", fill = "blue") +
  labs(x = "Proportion of Left Driving Countries", y = "Samples")
```



(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).

```
quantile(boot_props$props, c(0.05, 0.95))
```

```
##   5%  95%
## 0.30 0.46
```

(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.

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

This is not correct, because we are concerned about the true proportion of all countries/regions that drive on the left side, not the proportion of the countries from our sample in (b). We were not trying to make a conclusion about our original sample; we were trying to make a conclusion about the population.

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

This is incorrect, because we are not looking for the chance or probability that between 27% and 44% of all countries in the population drive on the left side. It is about how confident we are that the interval of (27%, 44%) captures the true proportion of countries that 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.

This is correct, because our 90% confidence signifies that 90% of the intervals calculated and performed in this way will capture the true proportion of countries/regions that drive on the left side.

(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.

We would use a wider confidence level, because larger intervals would encompass a larger range of (potential) values, which would mean that it is more likely that this larger range of values would contain the true proportion of all countries who drive on the left side.

(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:

$H_0 : p = p_0$ $H_A : p \neq p_0$

where p is the proportion of countries who drive on the right side and p_0 is the proportion of countries who drive on the left side.

Question 2

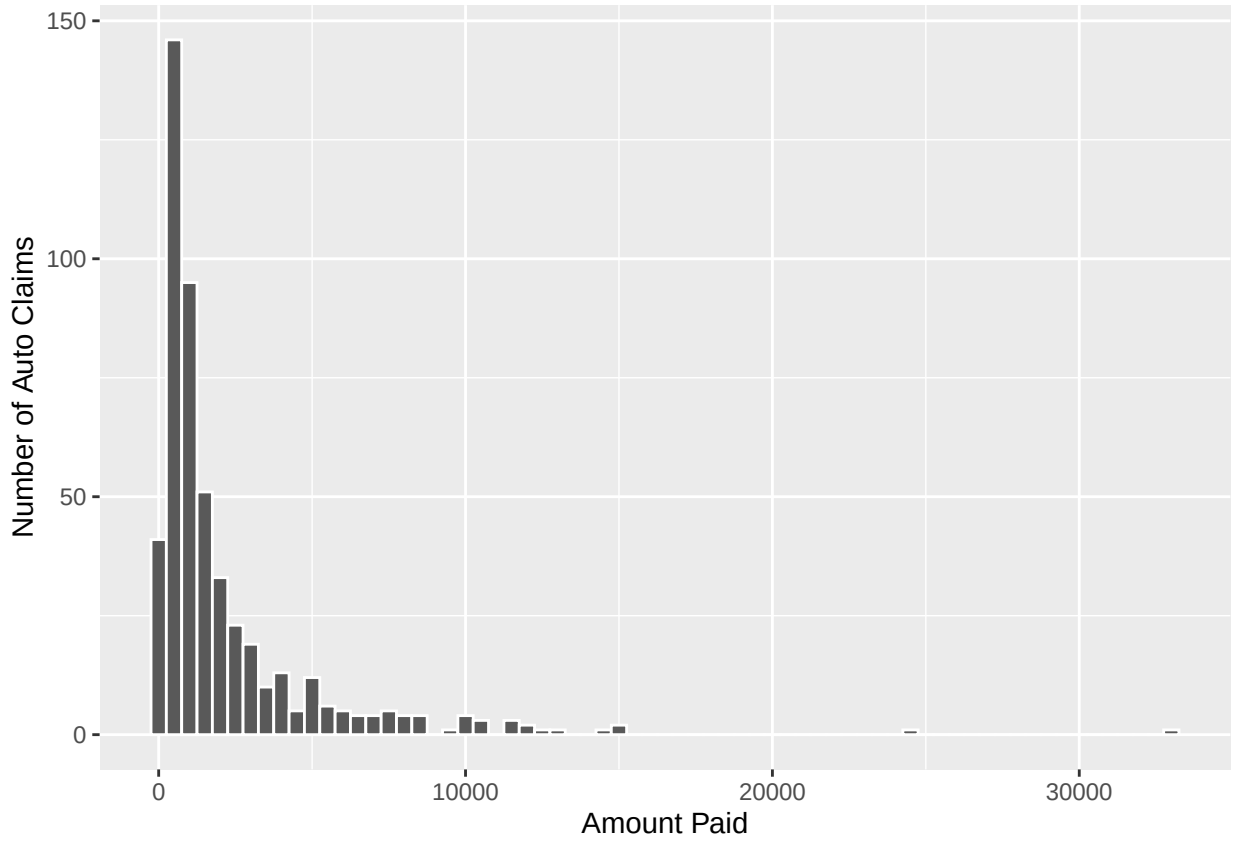
The data set `auto_claims.csv` includes claims paid (in USD) to a sample of auto insurance claimants 50 years of age and older in a specific year. In other words, it represents a 'sample' (the 'original sample') of car insurance claims in that year.

(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.

```
autoclaims <- read_csv(file = "auto_claims.csv")
autoclaims %>% summarise(lowest = min(PAID), avg = mean(PAID),
                         med = median(PAID), highest = max(PAID))
```

```
## # A tibble: 1 x 4
##   lowest   avg   med highest
##    <dbl> <dbl> <dbl>   <dbl>
## 1   25.4 2160. 1042.  33138.
```

```
autoclaims %>% ggplot(aes(x = PAID)) +
  geom_histogram(binwidth = 500, color = "white") +
  labs(x = "Amount Paid", y = "Number of Auto Claims")
```



The distribution of the number of auto claims vs. the amount paid appears to be skewed right, and it is unimodal with a peak at around $500-1000 in the Amount Paid. There seems to be somewhat of a variability in how much were paid for each auto claim, as the range seems to be around 0 to over 32000. Furthermore, the values are spread more concentratedly in around the 0 to 15000 range, which is still a wide range.

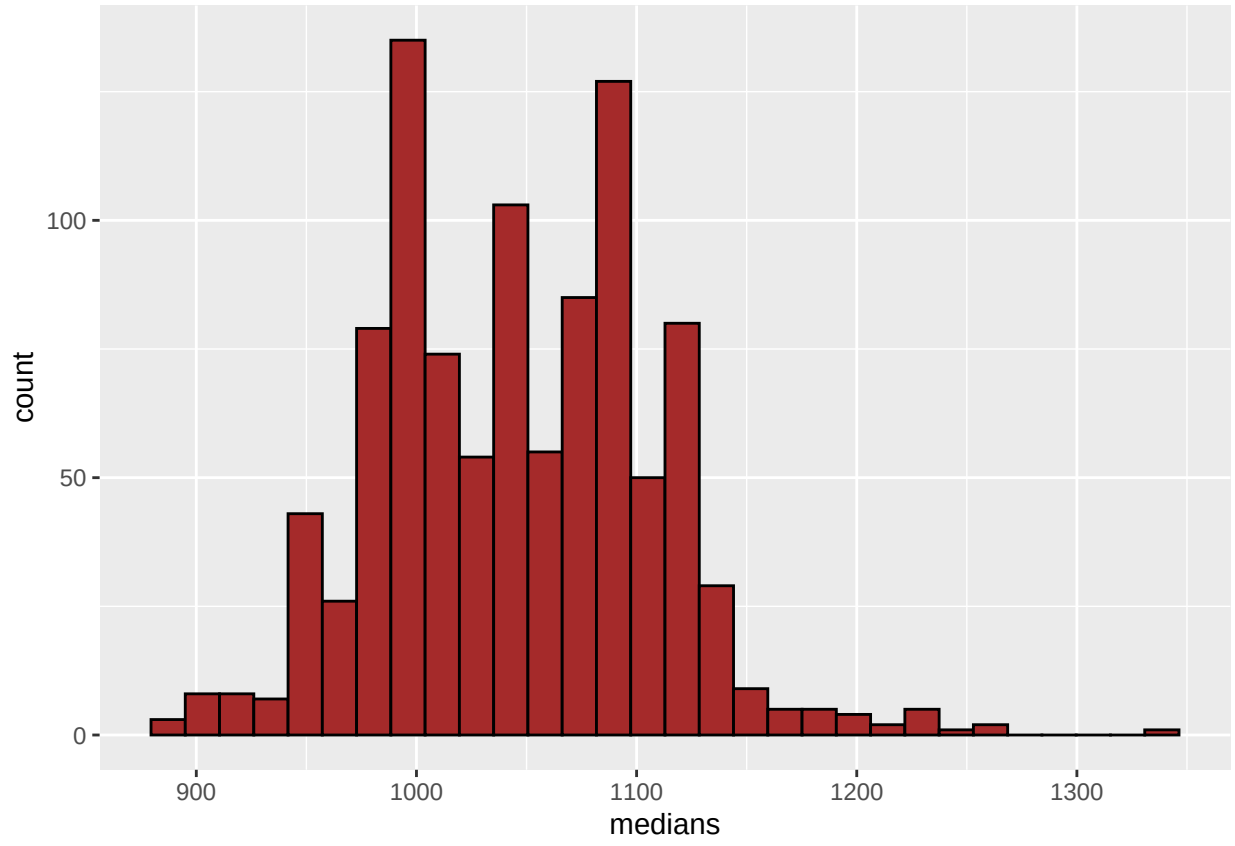
(b) Estimate the sampling distributions of sample *median* of paid claims by taking 1000 samples of size n=500 (to match the sample size in the data) and produce appropriate data summaries. Set the seed as the last *four* digits of your student number for each set of simulations.

```
set.seed(1351)

paid_medians <- rep(NA, 1000)
for (i in 1:1000){
  paid_samp <- autoclaims %>% sample_n(size = 500, replace = TRUE)
  paid_medians[i] <- as.numeric(paid_samp %>% summarise(med = median(PAID)))

}

paid_medians <- tibble(medians = paid_medians)
paid_medians %>% ggplot(aes(x = medians)) +
  geom_histogram(color = "black", fill = "brown")
```



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

```
quantile(paid_medians$medians, c(0.025, 0.975))
```

```
##     2.5%    97.5%
##  940.280 1153.317
```

We can predict with 95% confidence that the true median of paid claims falls between $940.280 and $1153.317.

Part 2

You are once again chatting on the phone to your friend. Your friend enjoyed your previous conversation about data visualization so much that your friend asked you if you had learned anything new in your STA130 course. You decided to tell them about the fancy new technique you just learned: bootstrapping! Be sure to include at least 2 vocabulary words from this week and explain them in simple terms for a lay audience.

Other things to consider: - Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 but less than 400 words.
- Remember to include a conclusion that reiterates the key points your friend should understand about bootstrapping.
- Use full sentences.
- Grammar is not the main focus of this assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).

Vocabulary

- Parameter
- Statistic
- Population
- Sample
- Sampling distribution
- Random sampling
- Resampling
- Bootstrap
- Percentile (quantile)
- Confidence interval
- Confidence level
- Testing
- Estimation
- Representative

Hey, friend! I'm glad you were interested in our last conversation about what I've been learning in my statistics STA130 course. What we've been learning lately, since then, is about bootstrapping. It's just like the name sounds; we're making the most of what we've got. Sometimes we don't have enough data to take several samples from. In this case, we use bootstrapping when we only have enough data to take one sample, so we take what we call "bootstrap samples" from that one sample. This is done with replacement and with the same sample size as the original sample. We can then use all the statistics from the bootstrap sample and use them to estimate a confidence interval at a certain confidence level. Confidence levels determine how confident (in %) we are that the confidence interval generated based off of the bootstrap sampling data and the confidence level will capture the true parameter of the population. It's so cool that we can just make estimations on a population parameter based on resampling from one sample from the population!