STA238 – Winter 2021

Week 6 Bootstrap - Proportion Example

S. Caetano

[Question]

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 in `roaddata` cover all of the 270 countries/regions in the world in the year 2020. Thus, these form the entire population.

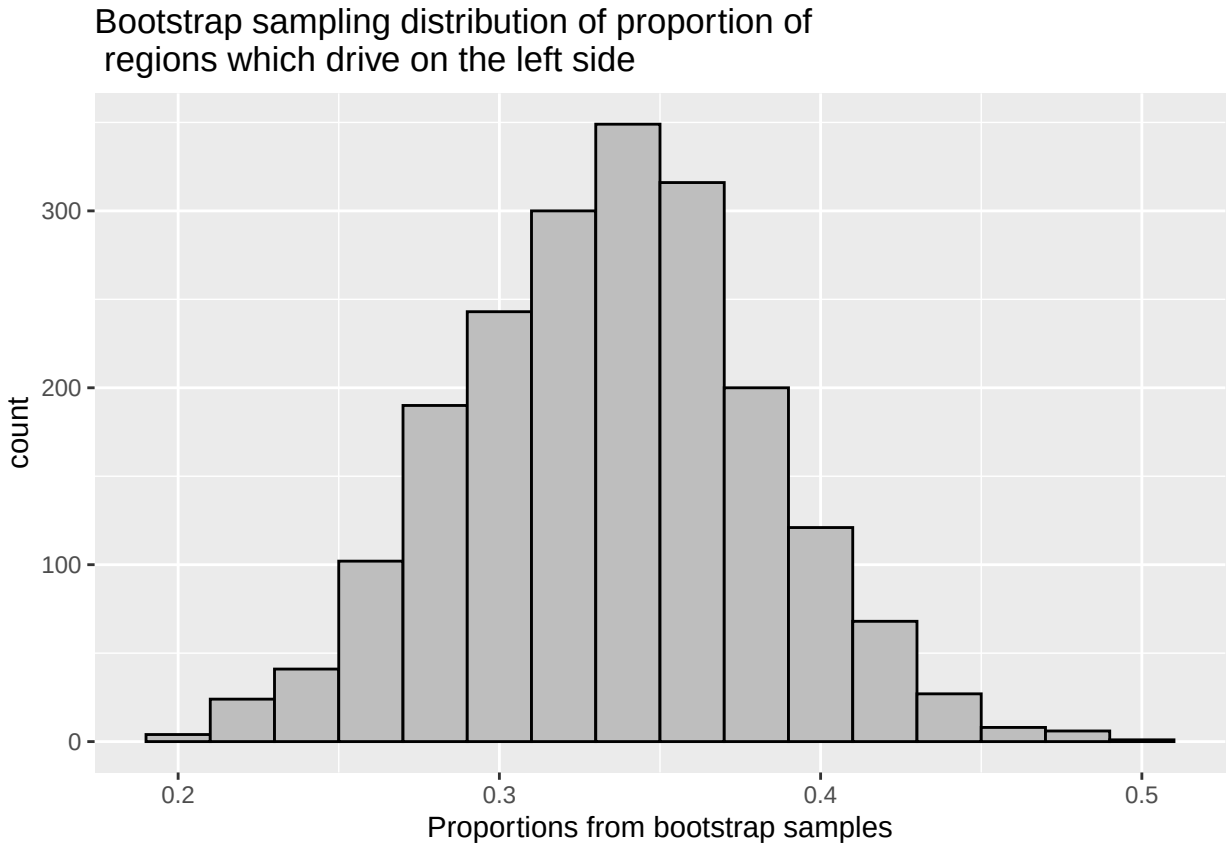
(b) Use the `sample_n()` function to select a random sample of different 100 countries/regions. Call this new data `road_sample`. Set the seed as the last *three* digits of your student number.

```
set.seed(333) # change to the last three digits of your student number

road_sample <- roaddata %>% 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 the seed as the last *three* digits of your student number.

```
set.seed(333) # change to the last three digits of your student number

boot_p <- rep(NA, 2000)  # where we'll store the bootstrap proportions

for (i in 1:2000)
{
```

```
  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)

ggplot(boot_p, aes(x=boot_p)) + geom_histogram(binwidth=0.02, fill="gray", color="black") +
  labs(x="Proportions from bootstrap samples",
    title="Bootstrap sampling distribution of proportion of \n regions which drive on the left side")
```



(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_p$boot_p, c(0.05, 0.95))
```

```
##   5%  95%
## 0.26 0.42
```

Based on the above, we see that our confidence interval for the proportion of countries/regions which drive on the left is (0.26, 0.42).

(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.

Incorrect. We know in this sample that 34% of countries/regions drive on the left ($\hat{p} = 0.34$) so we're 100% sure that $\hat{p}$ is between 0.27 and 0.44.

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

Incorrect. We know that in this population that 32% of countires drive on the left ($p = 86/270$) so we're 100% sure that p is between 0.27 and 0.44. Note that in general, we wouldn't know the true population mean, so we wouldn't know for certain whether a particular confidence interval contains the true population value or not.

(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.

Correct. Although this doesn't report the confidence interval computed, this is a valid description about how confidence intervals behave.

(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.

A wider confidence interval would be more likely to capture the true value of the population parameter. We would get a wider confidence interval if we set a higher confidence level (e.g., instead of 95%, use 98%). We would be extending to more of the bootstrap sampling distribution.