**Part 1. (12 Points)**

Each year the EPA does an analysis on the current models of vehicles sold the United States. The data provided in the data set EpaFE2019Data.csv is a subset of this analysis, if you are curious you may access the full data set from the EPA website <http://www.fueleconomy.gov/feg/download.shtml>.

Use the R script titled DA5\_Simulations\_and\_Confidence\_Intervals.R to upload the EpaFE2019Data.csv dataset and perform the necessary simulations.

In this exercise, we will use EPA car data as an example of a **population**.

* We will use R to select a simple random sample of vehicles from the population.
* We will then use this sampled data to compute confidence intervals and perform hypothesis tests.
  + This means, unlike typical hypothesis test or estimation procedures, we know our population parameters.
* **Why should we do this?** 
  + To provide an opportunity to evaluate the validity of estimation and hypothesis testing procedures. Does it work like we say it should?

**Follow the comments in the R script to complete the following:**

The variable combined carbon dioxide emissions, or CombCO2, represents the combined city and highway carbon dioxide emissions for vehicles sold in the US.

1. (3 points) Make a histogram of this variable and include your histogram here. What are the values of ? How large is the population? *Note: Consider this data a population. This implies the mean and standard deviation are parameters.* Paste the histogram and give a brief description of the population data.

Chart, histogram

Description automatically generated

The size of the population is 896, meaning 896 vehicles are represented in the dataset. The mean () of the population is 399.8717, and the standard deviation () is 89.825.

1. (3 points) Take a random sample of size 45 from the population. From your sample, calculate the sample statistics, and s. Make a histogram of carbon dioxide emissions for the sample of 45 vehicles. Paste the histogram. Make a brief description of the sampled data. Does it look much like the population?

Chart, histogram

Description automatically generated

From the above sample, = 380.244 and s = 72.431. The histogram for the sample data does not much look like the histogram for the population data because of the greater variations in the mean and standard deviation of the data. Size of 45 is relatively small compared to the population, so the histogram is much more spread out compared to the population histogram.

1. (3 points) Use , your sampled mean, from (Part 1b) and your sample standard deviation (Part 1b), to calculate the 90% confidence interval (CI) for . Show work!

Standard Error =

Confidence Interval =

The 90% CI estimates the to be between 362.483 and 398.005, with a point estimate of 380.244.

1. (1 point) Does your interval from part c include the true population mean for fuel efficiency?

The true population mean is 399.8717, which is outside the confidence interval of the sample.

1. (2 points) There are 400 students this term completing this same assignment. Assuming they calculated the CI correctly, how many students should we expect to have an interval that does not contain the true mean?

For a 90% CI, we should expect 360 (0.90 \* 400) students to have an interval that does contain the true mean, whereas 10% (40 students) would not contain the true mean.

**Part 2. (11 Points)**

A newborn is considered to be premature if it is born before 37 weeks of its gestation period. Assume the proportion of all babies in U.S. that are born prematurely is 0.101. This is the population parameter. We’ll use this parameter to simulate many samples in and make observations about the behavior of confidence interval intervals for a proportion.

To answer the following questions, you’ll need to run the R code given in DA5\_Simulations\_and\_Confidence\_Intervals.R in the section labeled Part 2.

1. (1 point) Randomly generate a single sample proportion , for a sample of size­ from the population described above. Report your value for here. Use lines 54-69 in the R script.

A picture containing text

Description automatically generated

1. (2 points) Use from the previous question to construct a 99% CI for the proportion of newborns born prematurely. Show your work and report your confidence interval here.

Standard Error =

Confidence Interval =

The 99% CI estimates the proportion of newborns born prematurely to be between 0.006 and 0.294, with a point estimate of 0.15.

1. (1 point) Does the interval you constructed in the previous question contain the population proportion ?

The 99% CI does contain the population proportion of 0.101.

1. (1 point) If you were to collect 10,000 random samples, all of size n=40, and construct a 99% confidence interval for the proportion, what percentage of those intervals would you expect to capture the true population proportion ?

We would expect 99% of those intervals to capture the true population proportion.

1. (2 points) Using R, simulate 10,000 random samples and construct the 99% confidence interval for each sample. I’ve written the code for you to do this. I have provided the code and some comments to explain what the code is doing in the DA5\_Simulations\_and\_Confidence\_Intervals.R (lines 71-93); however, if you still have questions about the code or want to learn about R syntax, just ask! Report the percentage of intervals that successfully captured the population proportion. Is this value the same or close to the value you expected (your answer to Part 2d)?



The percentage of intervals that successfully captured the population proportion is 91.78%, which is close to the value we expected in part 2d.

1. (2 points) Before constructing the confidence interval in Part 2b, you were not instructed to check the conditions necessary to construct a confidence interval for the population proportion. In order for the confidence intervals to behave as expected, we need the following two conditions to be met: and . Show that these conditions are *not* met.

Since np < 10 and n(1-p) 10, the first condition is not met.

1. (2 points) Using your answers from parts d and e, comment on how the confidence intervals behave when the sample size condition is not met.

As the sample size decreases, the confidence interval would get wider.

**Gradescope Page Matching (2 points)**

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.