### Topics:

Sampling Distributions, Central Limit Theorem, Confidence Intervals an Hypothesis Testing

One Variable Visual Displays and Summary Statistics for Quantitative Variables.

Lessons Covered: 20 - 25

Textbook Chapter (Optional) : 5, 6, 7

### Grading:

* Points are listed next to each question and should total 25 points overall.
* Grading will be based on the content of the data analysis as well as the overall appearance of the document.
* Late assignments will not be graded.

### Deadlines:

* Final Submission: **Thursday, October 29th**. All submissions must be PDF files.

### Instructions:

* Clearly label and **type answers** to the questions on the proceeding pages in Word, Google Docs, or other word processing software.
* Insert **diagrams or plots as a picture** in an appropriate location.
* Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. \*, /, sqrt() and ^
* Submit assignment to the Gradescope link as a PDF. **Indicate the pages to the individual questions and also verify the correct document has been uploaded. Failing to follow this direction may result in point deductions.**

### Allowances:

* You may use any resources listed or posted on the Canvas page for the course.
* You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
* Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

### Part 1. (8 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 DA4\_Simulation\_CLT\_and\_HypothesisTesting.R to upload the EpaFE2019Data.csv dataset and complete parts 1 through 3.

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. (2 points) Make a histogram of this variable. 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.

<Delete this text and insert your histogram here>

<Delete this text and write 1-2 sentences which describes what the parameters of the population are and how large the population is. If possible, highlight the numbers so they’re easy for the graders to find>

1. (2 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?

<Delete this text and insert your histogram here>

<Delete this text and write 2-3 sentences which describes what the statistics of the sample are (If possible, highlight the numbers so they’re easy for the graders to find) and also describe the center, shape and spread of the sampled data. Be sure to answer if the distribution of the sample resembles the distribution of the population.>

1. (2 points) Use , your sampled mean, from (Part 1b) and your population standard deviation (Part 1a), to calculate the 90% confidence interval (CI) for . Show work! Does the interval include the true population mean for fuel efficiency?

<Compute the confidence interval by showing your work (Type these equations out).>

<Write a sentence which describes whether or not the true population mean is contained in the confidence interval>

1. (2 point) There are 220 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?

<Write a 1-2 sentences which answers this question>

### Part 2. (10 Points)

Suppose we want to see whether our sampled data from Part 1 will reject the true value of the population mean. Set up a hypothesis test where the claimed average is the actual average carbon dioxide emissions value we found in part 1-a.

Does the sample data provide evidence the true average carbon dioxide emissions of all vehicles is different than

1. (2 point) Before performing the hypothesis test, can we anticipate the outcome? Will we most likely fail to reject or reject the null? Why?

<Write 2-3 sentences which answers this question. Be specific>

1. (3 points) Use your sampled mean from (Part 1b) and your population standard deviation (Part 1a), to perform a one sample z test for the above hypotheses, where is the actual population mean. Use a significance level of 0.10. Show your work for the test statistic and provide a p-value. *Note: You may use R to validate your results but should provide a solution worked by hand.*

<Compute the value of the test statistic “by-hand” and show your work. If possible, highlight the value of the computed test statistic>

<Compute the p-value and describe how you computed this value. If you use R’s “pnorm()” function, include the code you used to compute the p-value. Highlight the p-value if possible.>

1. (3 points) Make a four-part conclusion based on your results. This should include:

* A statement in terms of the evidence in favor of the alternative.
* Whether we should reject the null hypothesis.
* A point and interval estimate.
* Context.
  + *Note: This is just for practice. Given we have all of the population data we know the true average. In reality, we would not know population information.*

<Write a short paragraph for the four-part conclusion>

1. (2 point) If the interval in part 1-c does not contain the true parameter, why will the same sampled data also reject the true null using the hypothesis test?

<This question is a bit of a challenge. Be sure to discuss your solution with your peers or ask for help during office hours if you find this question challenging. Answer this question using a few sentences.>

**Part 3. (7 points)**

Consider your random sample from Part 1b, provided it was obtained randomly, your sample mean and standard deviation values are not static. That is, if we were to take a different sample, these values would change. We discussed this notion when we learned about repeated sampling and sampling distributions. The one sample z test is dependent on these values. Results for the test will vary.

Sample 10000 random samples of size 45 from the population and check out three different things: the sampling distribution for the sample means, the distribution of z test statistics and the distribution of p-values.

1. (2 point) According to the Central Limit Theorem (CLT), what is the distribution of the sample means? Include the theoretical mean and standard deviation values. Show work.

<Use what we know from the CLT to compute the mean and standard deviation for the sampling distribution of the sample mean. Then, describe what the sampling distribution of the sample mean will be using either statistical notation or a few carefully worded sentences.>

1. (1 point) Create a histogram of the sampling distribution for x-bar. Paste your plot. Do the simulated sample means support the Central Limit Theorem? Compare the shape, mean and standard deviation of the simulated sample means to what they should be theoretically.

<Include your plot here>

<Answer the rest of the question using a few complete sentences.>

1. (2 point) Create a histogram of your z test statistics. Paste your plot. What type of distribution will model these test statistics?

<Include your plot here>

<Answer the rest of you question using a complete sentence. Be specific about the type/name of the distribution!>

1. (2 points) Create a histogram of the p-values. **We know the null hypothesis is true,** so there are two things we should expect: the p-values to follow an approximate uniform distribution and just by chance, we will reject the null of the time. Does this seem to be the case? How often do we reject the null? What type of error does this represent?

<Include your plot here>

<Answer the rest of the question using 2-3 sentences>