ST 314 Data Analysis 04

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.

ST 314 Data Analysis 04

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.

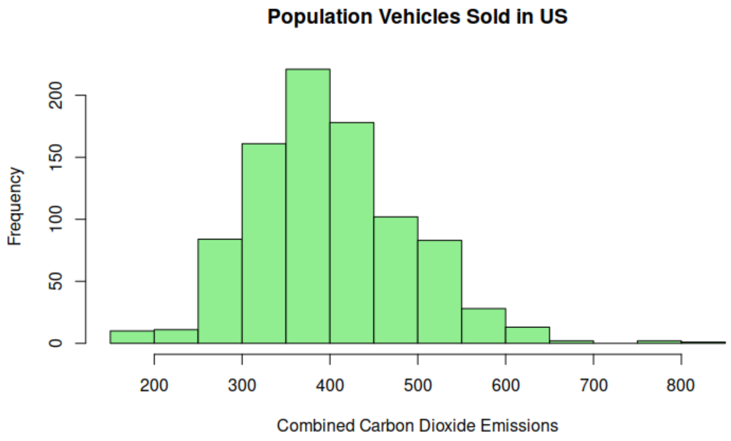
Use the R script titled DA4_Simulation_CLT_and_HypothesisTesting.R to upload the EpaFE2019Data.csv dataset and complete parts 1 through 3.

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.
    - o This means, unlike typical hypothesis test or estimation procedures, we know our population parameters.
- Why should we do this?
    - o To provide an opportunity to evaluate the validity of estimation and hypothesis testing procedures. Does it work like we say it should?

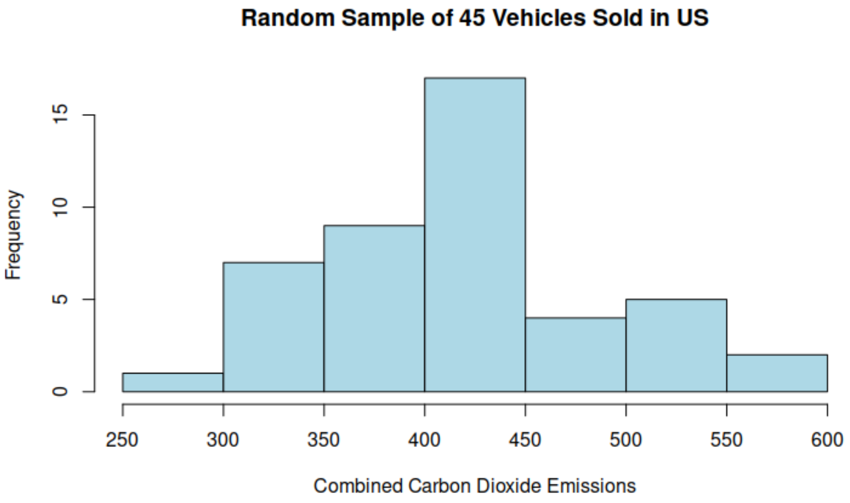
Follow the comments in the R script to complete the following:
The variable combined carbon dioxide emissions, or CombCO2, represents the combined city and highway carbon dioxide emissions for vehicles sold in the US.

a (2 points) Make a histogram of this variable. What are the values of mu and sigma? 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.



This population is described by **mu=399.872**, and **sigma=89.825**. This means that the average carbon emissions of the population is mu, and that the standard deviation is sigma. The population numbers **896** members, and this represents the number of cars that the EPA surveyed as part of this dataset.

b (2 points) Take a random sample of size 45 from the population. From your sample, calculate the sample statistics, $\bar{x}$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?



The sample of **n=45** randomly chosen members from the population in part 1(a) resembles the distribution of the population, however does not resemble it closely. This random sample has parameter values **xbar=417.867** and **sigma_xbar=70.302**. This means that the mean of this new data set lies at 417.867, and the standard deviation is sigma. The shape appears to be left skewed, slightly, however not obviously, and it appears to be centered around 420 or so.

c (2 points) Use $\bar{x}$, your sampled mean, from (Part 1b) and your population standard deviation σ (Part 1a), to calculate the 90% confidence interval (CI) for $\mu_{CO2\,Emissions}$. Show work! Does the interval include the true population mean for fuel efficiency?

$$\bar{x} \mp (Z_{a/2})(\sigma_{\bar{x}}) = \bar{x} \mp (Z_{a/2})\left(\frac{\sigma_x}{\sqrt{n}}\right) = 417.867 \mp (1.645)\left(\frac{70.302}{\sqrt{45}}\right) = (400.627, 435.107)$$

The true population mean, 399.872, is not contained in this interval. This is likely because the random sample of 45 that were selected had emissions that were on average above the mean, shifting the sample mean above the population mean.

d (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?

Given that we are calculating a 90% confidence interval, then 90% of the confidence intervals of random samples of the population will capture the parameter, i.e. include the mean, whereas 10% will not. Therefore, if 220 students each calculate 1 90% confidence interval each, we would expect that about **10% would fail to capture the parameter, which is equivalent to 22 students out of the 220. The remaining 198 students would be expected to get confidence intervals that include the parameter.** I'm one of those 22 students, as my interval did not capture the parameter / the population mean :).

ST 314 Data Analysis 04

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.

$H_0 : \mu = \mu_{CO2Emission}$
$H_a : \mu \neq \mu_{CO2Emission}$

Does the sample data provide evidence the true average carbon dioxide emissions of all vehicles is different than $\mu_{CO2Emission}$?

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

We will most likely fail to reject the null, as our sample average is quite far from the population average, but the sampled population is very small in comparison to the total population, therefore the sample is less likely to properly represent the population. Though our confidence interval does not contain the parameter, demonstrating that our sample is not representative of the population, the size of the sample is taken into account in the calculation of the z score, therefore we can expect that to be reflected in the subsequent p value. Therefore, we will most likely fail to reject the null hypothesis based on the data collected.

b (3 points) Use $\bar{x}$ your sampled mean from (Part 1b) and your population standard deviation σ (Part 1a), to perform a one sample z test for the above hypotheses, where $\mu_{CO2Emission}$ 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.*

$$\alpha = 0.10, \bar{x} = 417.867, \mu_0 = 399.872, \sigma = 89.825, Sided = 2, n = 45$$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{417.867 - 399.872}{89.825 / \sqrt{45}} = 1.34$$

$$Z\ Table\ Value\ at -1.34 = 0.0901$$
$$For\ 2 - sided : \mathbf{p = 2 * 0.0901 = 0.1802}$$

c (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.*

After doing a two tailed hypothesis test, There is not convincing evidence that the sample average of carbon emissions is not representative of the population average of carbon emissions, which is equal to 399.872. The null hypothesis is rejected at a significance level of 0.10, z stat of 1.34, and p-value of 0.18. The sample estimates that the average carbon emissions are about 418, with a 90% confidence interval of 401 to 435.

This sample fails to reject the null hypothesis. This is because p-value $> \alpha$, therefore we fail to reject the null. The evidence is for rejecting the null, however it is somewhat-weak evidence, therefore fails to reject null. When it comes to the point and interval estimate, the confidence interval of our sample does not include the parameter, indicating that the sample mean x_bar is not representative or close to the population mean mu. However, the sample is of a small enough size that this kind if deviation is probable enough that this still fails to reject H_o (however this is also because the difference between x_bar and mu is reasonably small). Given that we calculated a confidence interval of 90%, it is expected that a certain number of these intervals will not include the parameter, such as in my case. These still cannot reject the null, however, as they are close enough to the original population to demonstrate weak evidence against H_0.

d (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?

My interval did not contain the true mean, and yet failed to reject the hypothesis because the certainty of the p value was not high enough – i.e. the p value was not small enough to indicate strong certainty against H_0. My p-value evidence against H_0 is pretty weak, and therefore $p > \alpha$. This is as a result of the p value (and z value) factoring in the sample size, and standard error, in order to generate a value that reflects the conditions of such small samples of a larger population.

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.

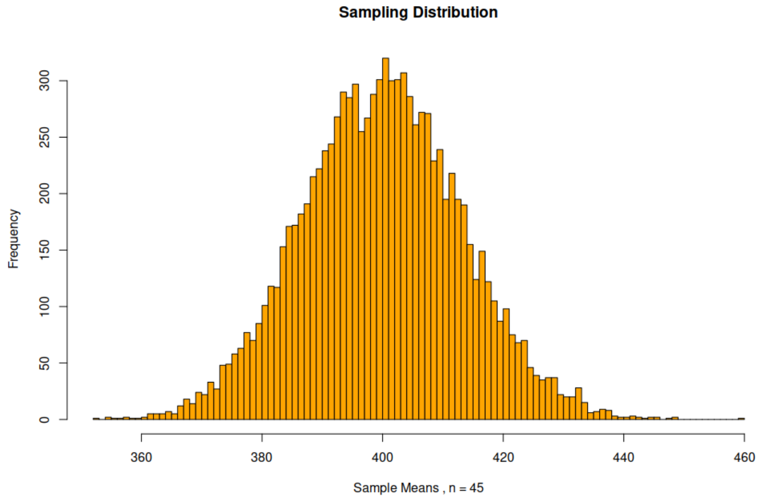
a (2 point) According to the Central Limit Theorem (CLT), what is the distribution of the sample means? Include the theoretical mean and standard deviation values. Show work.

<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.>

The distributions of the sample means in the sampling distribution will be normal, per CLT. The theoretical mean value should be 399.872, our population mean, μ_{CO2} . The theoretical standard deviation value is:

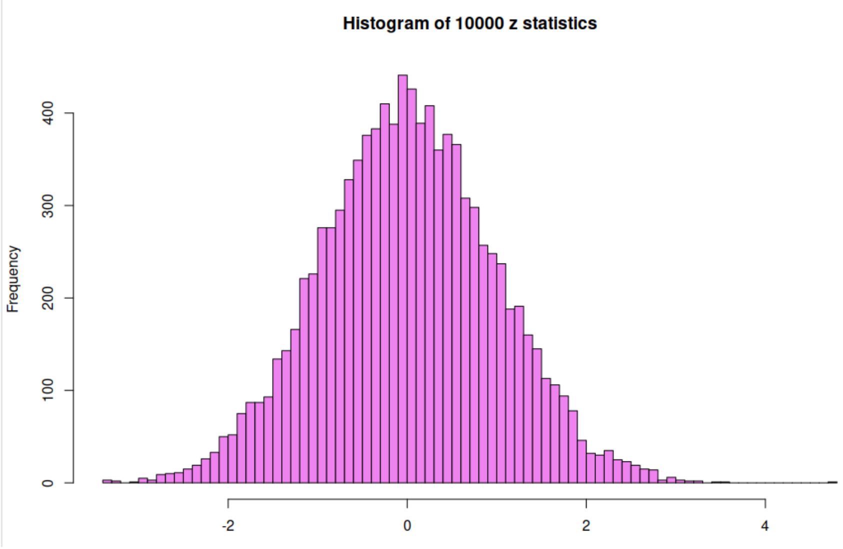
$$\sigma_x = 89.825, n = 45, SD(\bar{x}) = \frac{\sigma_x}{\sqrt{n}} = \frac{89.825}{\sqrt{45}} = 13.39$$

b (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.



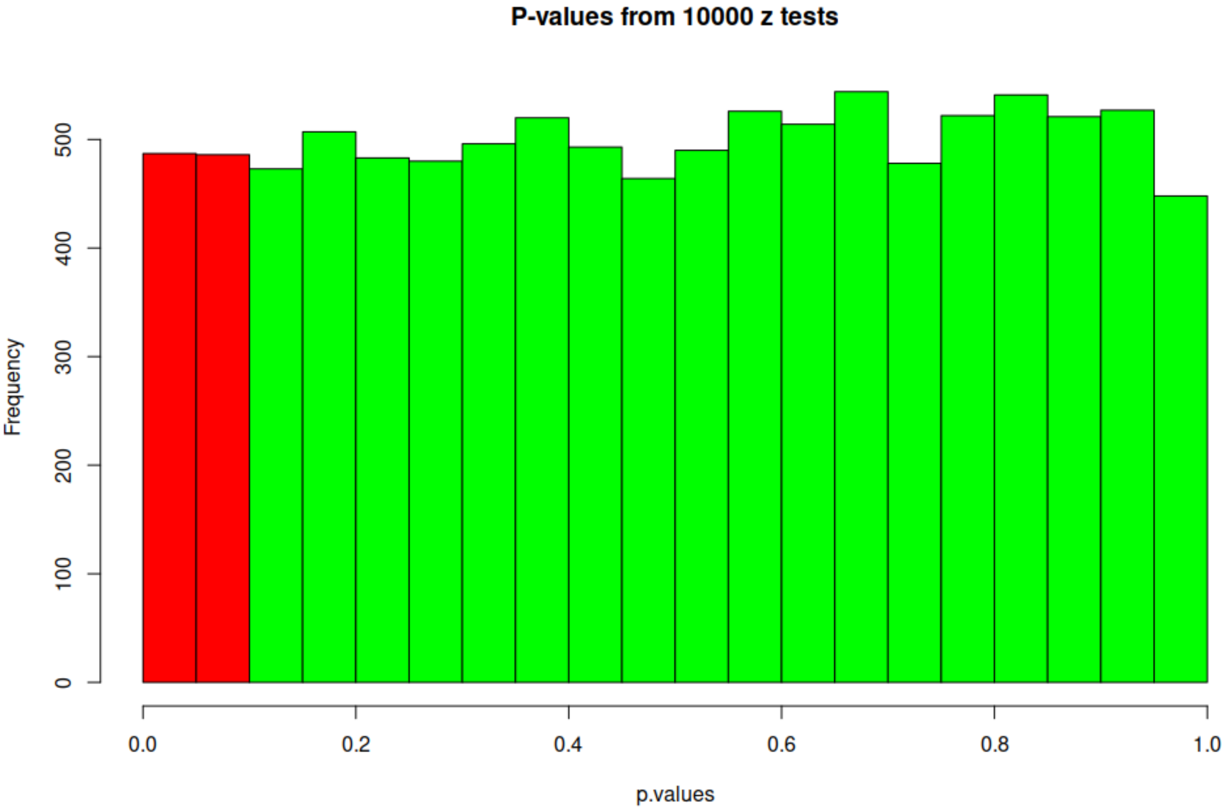
The sampling mean of 399.927 is very close to our estimate sampling mean of 399.872. Our standard deviation is also very close to the estimate, where the sampling SD is 13.22 and our estimated SD is 13.39. The shape of the graph is as expected, a normal curve.

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



This graph is a standard normal distribution, as the mean is about 0 (experimentally: 0.0066), and the standard deviation is about 1 (0.9734).

d (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 $\alpha \times 100\,\%$ of the time. Does this seem to be the case? How often do we reject the null? What type of error does this represent?



The p-values do follow an approximately uniform distribution. We reject the null 10% of the time, which is consistent with the red-labeled data points, which number 2 out of 20, or 10%. This represents Type I error, as H_0 is true, and in these cases we reject H_0.