Data Analysis 6: Hypothesis Testing

In this data analysis, you will assess the validity of claims using real data and the hypothesis testing procedures we've discussed in class.

If you haven't already done so, work through the tutorial provided on the Data Analysis 6 Canvas page. Once you've worked through the tutorial, write up your responses to the questions listed throughout the tutorial. The same questions are included below to help you format your submissions.

Submit a PDF copy of your responses to Gradescope by the deadline stated on Canvas.

Part 1: Caffeine

Question 1 (0.5 points): What is the parameter of interest in this scenario? Provide context.

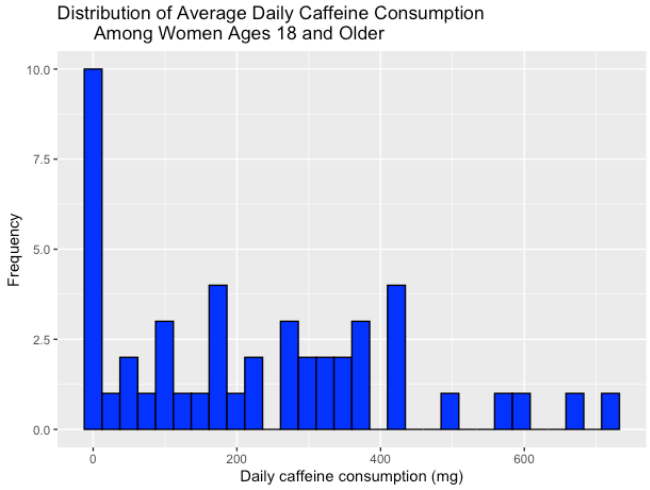
average caffeine consumption among women ages 18 and older greater than 200 mg

Question 2 (2 points): State the null and alternative hypothesis to answer the question of interest. Clearly define any notation you use.

H_A: $\mu > 200$mg H_0: $\mu < 200$mg

Question 3 (1 point): Make either a histogram or boxplot to visualize the variable consumption. Refer back to Data Analysis #1 for guidance on how to use the ggplot() functions to create a histogram or boxplot.

A. (0.5 points) Include your histogram or boxplot in your document. Be sure to include clear axes labels and a title.



B. (0.5 points) Based on your visualization, is there visual evidence the average daily caffeine consumption for adult women is greater than 200 mg?

The average daily caffeine consumption for adult women is greater than 200, according to my visualization.

Question 4 (1 point): Check the conditions required to perform the appropriate hypothesis test for this scenario. State the conditions and whether or not they are met.

There are the common conditions required: first, the data should be random sampled, which satisfy. Second,The sample size is sufficiently large, which sample size n = 47 which is greater than 30. Thus, the conditions are meet.

Question 5 (2 points): Use the sample data to calculate the appropriate test statistic. From the sampled data, you'll need the sample mean and sample standard deviation. You can use the mean () and sd () functions in R to find these values.

A. (1 point) Calculate the test statistic. Report the value of the test statistic and show your work (i.e., demonstrate how you arrived at that value).

First, I used R code to find mean and sd:

> mean(caffeine$consumption) [1] 229.5745

> sd(caffeine$consumption) [1] 196.0696

Then we can get n = 47, sd = 196.0696, and mean = 229.5745

$$\frac{229.5745 - 200}{\frac{196.0696}{\sqrt{47}}} = 1.03408$$

Thus, the test statistic is 1.03408.

B. (1 point) Determine the null distribution of the test statistic. State the name of the distribution and include any parameter values needed to define the distribution.

The name of the distribution is t- distribution and test statistics is 1.034 and with 46 degrees of freedom. To define the distribution of parameter values are sample mean($\bar{x}$), population mean(mu_0), sample standard deviation(s), and sample size(n) and t-distribution relay on degrees of freedom(n-1)

Question 6 (1 point): Calculate and report the p-value. Include any code used to do this calculation.

> 1-pt (1.03408,46) [1] 0.1532524

P-value is: 0.1532524

Question 7 (1 point): Use the t.test() function to verify the hypothesis test calculations you performed in questions 5 and 6. Set the significance level to α=0.05. Include the t.test() output in your assignment here.

The R code: t.test(caffeine$consumption, mu=196.0696, alternative="greater", conf.level= 1-0.05)

Output: One Sample t-test

data: caffeine$consumption

t = 1.1715, df = 46, p-value = 0.1237

alternative hypothesis: true mean is greater than 196.0696.

95 percent confidence interval:

181.5653 Inf

sample estimates:

mean of x

229.5745

Question 8 (2 points): From the R output, write a four-part conclusion describing the results. Use a significance level of α=0.05. Provide a statement of evidence in terms of the alternative hypothesis. State whether (or not) to reject the null. Give an interpretation of the point and interval estimate. Be sure to include the context of the problem in your conclusion.

There is no evidence to suggest the average caffeine consumption among women ages 18 and older greater than 200 mg. we fail to reject the null hypothesis snice our p-value is greater than 0.05. we are 95% confident that the average caffeine consumption among women ages 18 and older is between 173.520 and 285.629 with a point estimation of 229.5745.

Part 2: Combined CO_2 Emissions

Question 9 (0.5 points): Since we are treating this entire data set as the population, we can calculate the population mean, μ. Using the mean () function and the vector data we're interested in, population$CombCO2, calculate population mean and store it in mu_co2. It's important that you store the mean value so that we can reference it later on in this exercise. Report the calculated population mean.

mu_co2 [1] 399.8717

Question 10 (1 point): Before performing the hypothesis test, can we anticipate the outcome? Will we most likely fail to reject or reject the null? Why?

From my perspective, we cannot anticipate the outcome before performing the hypothesis test, but in this scenario, since we already used the R code: `n <- 45`

```
sample <- population %>%
sample_n(size = n)
sample_mean <- mean(sample$CombCO2)
```

to calculate the sample mean (3888.5333): given the H_0 is average equal to 399.8717, The H_A is not equal to 399.8717, since the sample mean is lower than the population mean, there is a possibility that we will most reject the null hypothesis. However, the anticipated outcome will depend on many things, such as variability within the population, the sample size, therefore, based on sample mean alone we cannot anticipate the outcome.

Question 11 (3 points): Using the information from your sample, (the sample mean, $\overline{x}$, the sample standard deviation, s, and the sample size, n), perform a t-test for the mean.

A. (1 point) Calculate the t test statistic. Since we're assessing the performance of a t-test, you should use your **sample standard deviation**, s, in the calculation of the test statistic, despite the fact that we have access to the population standard deviation.

sample mean = 394.2444, sample standard deviation = 85.9974, population mean = 399.8717, n = 45

```
> t_statistic <- (394.2444 - 399.8717) / (85.9974 / sqrt (45))
 > t_statistic
[1] -0.438956
```

B. (1 point) Determine the p-value.

```
> 2*(1-pt (abs (-0.438956),44))
[1] 0.6628458
```

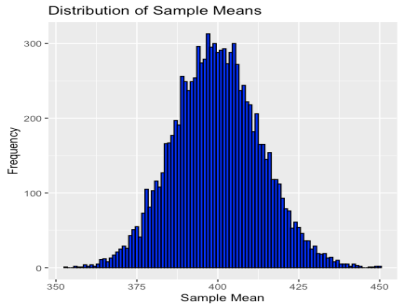
p-value is: 0.6628458.

C. (1 point) Using a significance level of α=0.05, does your p-value lead you to reject or fail to reject the null hypothesis? Does this conclusion align with or contradict what you expected to happen in question 10?

Yes, my p-value lead me to fail to reject the null hypothesis because 0.6628458 > α=0.05. It contradicts with what I expected to happen in question 10.

Question 12 (2.5 points): Construct a histogram of the 10,000 sample means stored in the vector sample_means45$mean. Refer back to Data Analysis #1 for guidance on how to use the ggplot() functions to create a histogram.

A. (0.5 points) Include your histogram in your document. Be sure to include clear axes labels and a title.



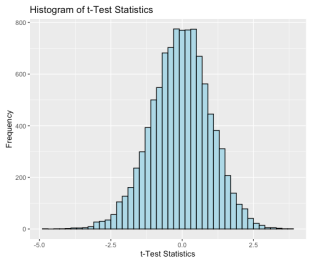
B. (1 point) Describe the distribution. Include a description of the shape, center, and spread.

Shape: nearly symmetric and unimodal.
Center: about 400
Spread: The range of the histogram is between 353-453

C. (1 point) According to the Central Limit Theorem (CLT), what is the distribution of the sample means? According to the CLT theorem, the distribution of the sample means approaches a normal distribution, the sample mean of sample mean is equal to the population mean. Therefore, the distribution of the sample mean is 399.8717.

Question 13 (2.5 points): Construct a histogram of the 10,000 t-test statistics stored in the vector sample_means45$t.

A. (0.5 points) Include your histogram in your document. Be sure to include clear axes labels and a title.



B. (1 point) Describe the distribution. Include a description of the shape, center, and spread.

Shape: nearly symmetric and unimodal.
Center: at 0.0
Spread: The range of the histogram is between -5.0 to 5.3

C. (1 point) In our conversations around estimating and testing the population mean when the sample standard deviation is used in the standard calculation, we discussed the theoretical distribution of $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$. What is the name of this distribution? What parameter value(s) do we expect it have?

The name of distribution is t- distribution, parameters we are expected to have been sample mean(x-bar), population mean(mu_0), standard deviation(s), and sample size(n) and degrees of freedom(n-1)

Question 14 (3 points): Consider the distribution of p-values displayed in the histogram you just created.

A. (1 point) Describe the distribution. Include a description of the shape, center, and spread.

Shape: the shape is approximately uniform, most of bins have the same height.
Center: at 0.5
Spread: The range of the histogram is between 0.00to 1.00, spread is relatively small

B. (1 point) **Remember in this atypical situation, we know that the null hypothesis is true;** therefore, we expect that just by chance, we will falsely reject the null hypothesis α×100% of the time. Use the code mean (sample_means45$p_val <= 0.05) * 100 to calculate the percentage of hypothesis tests that rejected the null hypothesis even though the null hypothesis is true. Does the percentage from your simulation align with the percentage we expected?

The percentage (5%) from my simulation aligns with the percentage we expected.

C. (1 point) The tests that produce a p-value less than the significance level will lead us to falsely reject the null hypothesis. This incorrect conclusion represents one type of error that can occur when performing a hypothesis test. What type of error is this?

The type of error occurs when we falsely reject the null hypothesis. Based on a test that produces a p-value less than significance level, is called a Type I error.

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”.