PSTAT 5A: Homework 5

Please note, you are responsible for completing your homework and submitting it to Gradescope. Though the Syllabus allows for the option to not submit homework, it is highly recommended.

1. According to a market research firm, the average duration of ownership for a laptop in the United States this year (2023) is estimated to be 4.5 years. To test this claim, Maria takes a representative sample of 50 people and notes that the average laptop ownership duration of these 50 people is 4.8 years, with a standard deviation of 0.75 years. Suppose Maria wishes to use a 5% level of significance to test that the average duration of ownership for a laptop in the United States is longer than 4.5 years.
   a. State the null and alternative hypotheses. Be sure to use the proper notation and define your parameters.

      $H_0: \mu = 4.5$
      $H_A: \mu > 4.5$
      μ the average duration of laptop ownership

   b. Check the necessary conditions for this test.

      Is the population normally distributed? No
      Do we have a large enough sample? N = 50 $\geq$ 30 ✓
      Do we have population SD or sample SD? Sample SD

      Therefore $\bar{x} \sim t_{49}$

   c. Calculate the test statistic.

      $TS = \frac{4.8-4.5}{\frac{0.75}{\sqrt{50}}} = \frac{0.3}{0.1061} = 2.83$

d. Write the Python code to calculate the p-value for this test.

```
(1 - stats.t.cdf(abs(t_statistic), df))
```

0.003

e. Suppose the p-value is calculated to be 0.006. What would you tell Maria about their finding?

At the 5% significance level, since p-value < 0.05, we have enough evidence to reject the claim that the average duration of laptop ownership is 4.5 years. Our data suggests that the average duration of laptop ownership is not 4.5 years.

f. Calculate the 90% confidence interval. Remember to include interpretation.

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$4.8 \pm 1.677 \frac{0.75}{\sqrt{50}}$$

$$4.8 \pm 0.1778$$

$$[4.6222,\ 4.9778]$$

We are 90% confident that the true average duration of laptop ownership is between 4.622 and 4.978 years.

2. Michael wondered whether apartments built in Austin in the 1990s are larger, on average, than apartments built in Austin in the 2020s. He pulled data from the Austin property search website, recording the living area (in square feet) and the decade the apartment was built (1990s or 2020s) for a random sample of 60 apartments built in Austin in the 1990s and another random sample of 40 apartments built in Austin in the 2020s. Michael's data set has two variables: living area in square feet and decade (which has two values: 1990s and 2020s. Suppose the mean living area in square feet of apartments from the 1990s is 1917.9 sqft with standard deviation of 622.5 sqft, and the mean living area in square feet of apartments from 2020s is 1737 sqft with standard deviation of 640.2 sqft.
   a. State the null and alternative hypotheses. Be sure to use the proper notation and define your parameters.

$H_0: \mu_{1990} - \mu_{2020} = 0$

$H_A: \mu_{1990} - \mu_{2020} > 0$

μ_{1990} the average size of apartments in 1990s

μ_{2020} the average size of apartments in 2020s

   b. Check the necessary conditions for this test.

1. Independence within each sample is assumed by the random samples taken from Austin
2. Independence between the sample is assumed since these homes are built decades apart
3. Since the sample size of the apartments from the 1990s is 60 and from 2020s is 40, both are greater than 30 so we can relax the normality assumption.

   c. Calculate the test statistic.

$$TS = \frac{\bar{x}_{1990} - \bar{x}_{2020} - 0}{\sqrt{\frac{s^2_{1990}}{n_{1990}} + \frac{s^2_{2020}}{n_{2020}}}} = \frac{1917.9 - 1737 - 0}{\sqrt{\frac{622.5^2}{60} + \frac{640.2^2}{40}}} = 1.40$$

d. Write the Python code to calculate the p-value for this test.

1-stats.t.cdf(t_statistic, df)

e. Suppose the p-value is calculated to be 0.06. What would you tell Michael about their finding at the 5% significance level?

At the 5% significance level, since p-value > 0.05, we do not have enough evidence to reject the claim that the average size of apartments in the 1990s is the same as the average size of apartments in the 2020s. Our data suggests that the average size of apartments in the 1990s are larger than those from the 2020s.

f. Would your conclusion change if Michael tested whether the average living area of the apartments from the 1990s is different from the average living area of those from the 2020s?

No, if Michael was interested in a two-sided test, we would calculate the p-value from two side of center of the distribution:
2 * (1 - stats.t.cdf(abs(t_statistic), 39))
This would be 0.170, which would not be enough evidence to reject the claim that the average size of apartments in the 1990s is the same as the average size of apartments in the 2020s and not different than the one-sided test.

g. Calculate the 85% confidence interval. Remember to include interpretation.

$$(\bar{x}_{1990} - \bar{x}_{2020}) \pm t^* \sqrt{\frac{s^2_{1990}}{n_{1990}} + \frac{s^2_{2020}}{n_{2020}}}$$

$$(1917.9 - 1737) \pm 1.47 \sqrt{\frac{622.5^2}{60} + \frac{640.2^2}{40}}$$

$$180.9 \pm 1.47(129.2)$$

[-8.89, 370.69]
We are 85% confidence that the true difference in average living areas of apartments built in 1990s and the 2020s is about -9.89 and 370.9 sqft.

3. Marissa would like to determine whether or not there is a significant difference between the average price of an apartment in Santa Barbara and the average price of an apartment in Los Angeles. Her initial beliefs are that the average prices in these two cities are the same. To test this claim, she takes a representative sample of 20 Santa Barbara apartments and another representative sample of 25 Los Angeles apartments; her data is summarized below (measurements are reported in thousands of dollars per month):

	Sample Mean	Sample Std. Dev.
Santa Barbara	2.69	0.67
Los Angeles	2.71	0.55

Assume that all independence and normality assumptions are met. Additionally, suppose Marissa decides to label "households in Santa Barbara" as Population 1 and "households in Los Angeles" as Population 2. Finally, assume Marissa wants to test her initial assumption against a two-sided alternative.

a. Define the parameters of interest, μ_1 and μ_2.

μ_1: average price of an apartment in Santa Barbara

μ_2: average price of an apartment in Los Angeles

b. State the null and alternative hypotheses.

Ho: $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 - \mu_2 \neq 0$

c. Compute the observed value of the test statistic.

$$TS = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2.69 - 2.71}{\sqrt{\frac{0.67^2}{20} + \frac{0.55^2}{25}}} = -0.1076$$

d. Assuming the null is correct, what is the approximate distribution of the sampling distribution? Be sure to include any / all relevant parameters.

t_{37} *or* t_{19}

e. What is the p-value of the observed test statistic? (You may need to use Python for this part).

2 * (1 - stats.t.cdf(abs(t_statistic), df))

0.915

f. What is the critical value of the test, if we use an $\alpha = 0.05$ level of significance? (You can use Python but you don't need to).

2.03

g. Now, carry out the test at $\alpha = 0.05$ level of significance. Be sure to phrase your conclusions in terms of the context of the problem.

Since p-value > α, we do not have enough evidence to reject the claim that the average price of an apartment in SB is the same as the average price of apartments in LA. Thus, our data suggest that the average price of an apartment in SB is about the same as the average price of apartments in LA.

4. Faculty and students are all concerned about the amount of money that students need to spend on textbooks. In this question, we will examine the differences between the cost of new textbooks available at the UCLA bookstore and on Amazon, and whether a difference exists.
   a. Define your parameter of interest.
   μ_d: mean of difference in textbook prices at UCLA and Amazon

   b. Provided a dataset of 68 books with the following columns: prices of the book from UCLA and prices of the book from Amazon. Describe how you will calculate the sample statistic.

   First, calculate the difference in price between UCLA and Amazon per book.
   Next, take the mean (and standard deviation) of the difference.

   c. State the null and alternative hypotheses. Be sure to use the proper notation and define your parameter.

   Ho: $\mu_d = 0$
   Ha: $\mu_d \neq 0$
   μ_d is the mean of difference in price of textbooks from UCLA and Amazon

d. Check the necessary conditions for this test.

1. Independence within the sample of differences, we assume the sample of 68 books are randomly selected.
2. Normality of the distribution of population of difference, we assume the normality assumption can be relaxed since the sample size of 68 is greater than 30.

e. Suppose the observed mean is 3.5832 with standard deviation of 13.423. Calculate the test statistic.

$$t = \frac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{n}}} = \frac{3.5832 - 0}{\frac{13.42347}{\sqrt{68}}} = 2.201$$

f. Write the Python code to calculate the p-value for this test.

```
2 * (1 - stats.t.cdf(abs(t_statistic), df))
```
0.031

g. The p-value for the test is 0.031. What is the conclusion for our hypothesis test at 5% significance level?

Since the p-value < alpha = 0.05, we have enough evidence to reject the claim that the mean difference in textbook price at UCLA and Amazon is 0. The data suggests that the mean price of textbooks at UCLA is not the same as at Amazon.

h. Would your conclusion change if instead, you want to run a one-sided alternative?

A one-sided test would have a p-value of about half of the two-sided test, approximately 0.0156. We can check using:
```
1 - stats.t.cdf(t_statistic, df)
```
This means the conclusion would not change, since for two-sided, it was already statistically significant.

5. Consider the following two sets of numbers:

$$X = \{1, 2, 4, 4, 6, 5, 3\}$$
$$Y = \{3, 4, 1, 4, 4, 2, 1\}$$

Compute the correlation between x and y by hand. You may use Python to check your work but you must show your work.

$\bar{X} = \frac{1+2+4+4+6+5+3}{7} = 3.57$

$\bar{Y} = \frac{3+4+1+4+4+2+1}{7} = 2.71$

$s_x = 1.72$

$s_y = 1.38$

$r = \frac{1}{n-1}\sum(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$

$\sum(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y}) = (\frac{1-3.57}{1.72})(\frac{3-2.71}{1.38}) + (\frac{2-3.57}{1.72})(\frac{4-2.71}{1.38}) + (\frac{4-3.57}{1.72})(\frac{1-2.71}{1.38}) + (\frac{4-3.57}{1.72})(\frac{4-2.71}{1.38})$

$+ (\frac{6-3.57}{1.72})(\frac{4-2.71}{1.38}) + (\frac{5-3.57}{1.72})(\frac{2-2.71}{1.38}) + (\frac{3-3.57}{1.72})(\frac{1-2.71}{1.38}) = 0.0602$

$r = \frac{1}{6}0.0602 = 0.010$