Name:______________________________ **Date:**_______________

PSTAT 5A: Homework 05

Summer Session A 2023, with *Ethan P. Marzban*

1. According to *Statista*, the average lifespan (i.e. time between consecutive replacements) of a smartphone this year (2023) is estimated to be 2.6 years. To test these claims, Stephen takes a representative sample of 47 people and notes that the average smartphone lifespan of these 47 people is 2.8 years and the standard deviation of lifespans in his sample is 0.61 years. Suppose Stephen wishes to use a 1% level of significance to test *Statistia*'s claim against a two-sided alternative.

   (a) Define the parameter of interest.

   (b) State the null and alternative hypotheses. Remember to use proper notation.

   (c) Define the random variable of interest.

   (d) What distribution should Stephen use to find the critical value of the test? Be sure to check any/all relevant conditions.

   (e) What is the observed value of the test statistic?

   (f) What is the critical value of the test?

   (g) Now, conduct the test and phrase your conclusions in the context of the problem.

2. A game developer claims the average completion time of their new game *GauchoAdventures* is 12 hours. However, among a representative sample of 61 gamers the average completion time was 12.75 hours. Additionally, previous market research has revealed the standard deviation of all completion times (among all individuals) to be 3.1 hours. Suppose we wish to use this data to test the developer's claims against an upper-tailed hypothesis, using a 5% level of significance.

   (a) Define the parameter of interest.

   (b) Define the random variable of interest.

   (c) What distribution should we use to find the critical value of the test? Be sure to check any/all relevant conditions.

   (d) What is the observed value of the test statistic?

   (e) What is the critical value of the test?

   (f) What is the $p-$value of the observed test statistic?

   (g) Now, conduct the test and phrase your conclusions in the context of the problem.

Name:____ **Date:**____________________

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.

(b) State the null and alternative hypotheses.

(c) Compute the observed value of the test statistic.

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

(e) What is the $p-$value of the observed test statistic? (As a reminder, you may need to use Python for this part.)

(f) What is the critical value of the test, if we use an $\alpha = 0.05$ level of significance? (Though you *can* use Python, you do not *need* to.)

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

4. The temperature at a randomly-selected weather station in San Francisco is normally distributed with mean 74°F and standard deviation 8.2°F. The temperature at a randomly-selected weather station in Los Angeles is normally distributed with mean 93°F and standard deviation 9.4°F. Assume that the temperature in San Francisco is independent of the temperature in Los Angeles. A weather station from San Francisco is selected at random and the temperature is recorded; another weather station from Los Angeles is selected at random and the temperature is recorded.

(a) Define the random variables of interest.

(b) Using proper notation, state the distributions of the random variables of interest.

(c) What is the probability that the temperature recorded in Los Angeles is exactly 4°F higher than that in San Francisco?

(d) What is the probability that the temperature recorded in Los Angeles is more than 4°F higher than that in San Francisco?

(e) What is the probability that the temperature recorded in Los Angeles lies within 3 degrees of the temperature in San Francisco?

5. In this problem, we will work through the computations of an ANOVA by hand. The data we will work with is:

$$\boldsymbol{x_1} = \{-1,\ 0,\ 1\}$$
$$\boldsymbol{x_2} = \{0,\ 1,\ 1,\ 2\}$$
$$\boldsymbol{x_3} = \{1,\ 2,\ 1\}$$

(a) Compute the numerator and denominator degrees of freedom.

(b) Compute the **group means,** $\overline{x}_1$, $\overline{x}_2$, and $\overline{x}_3$.

(c) Compute the **grand mean**, $\overline{x}$ (i.e. the mean across all observations, ignoring groups).

(d) Compute the **sum of squares between groups**:

$$\mathrm{SS_G} = \sum_{j=1}^{k} n_j(\overline{x}_j - \overline{x})^2$$

where n_j denotes the size of the j^{th} group.

(e) Compute the **sum of squares total**:

$$\mathrm{SS_T} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2$$

In other words, take each point in the dataset, subtract off the grand mean, square the difference, add that to the square of the difference between the second point and the grand mean, etc.

(f) Compute the **sum of squared errors**:

$$\mathrm{SS_E} = \mathrm{SS_T} - \mathrm{SS_G}$$

(g) Compute the **mean-square between groups** and **mean-square error**:

$$\mathrm{MS_G} = \frac{\mathrm{SS_G}}{\mathrm{df_G}}; \qquad \mathrm{MS_E} = \frac{\mathrm{SS_E}}{\mathrm{df_E}}$$

(h) Compute the value of the F−statistic.

(i) Compute the $p-$value of the statistic. (You will need to use Python.)

(j) Finally, combine your answers to produce an ANOVA table.

6. In the following parts, you will be presented with an ANOVA table that has some entries missing. Fill in the missing entries, and provide justification as to how you found those missing values. You may need Python for certain entries.

(a)

	DF	**Sum Sq.**	**Mean Sq.**	$F-$**value**	$\mathbb{P}(> F)$
Btw. Groups	4	10	<???>	<???>	<???>
Residuals	<???>	50	0.5		

(b)

	DF	**Sum Sq.**	**Mean Sq.**	$F-$**value**	$\mathbb{P}(> F)$
Btw. Groups	10	20	2	<???>	0.8636
Residuals	120	<???>	<???>		

7. Consider the following two sets of numbers:

$$x = \{1, 2, 3, 1, 2, 5, 4\}$$
$$y = \{3, 4, 1, 4, 4, 2, 1\}$$

Compute the correlation between `x` and `y`. Do **not** use Python, except for arithmetic computations (i.e. you may use Python as a calculator, but do **use** any more advanced functions like `numpy.std()`, or `np.mean()`.)