Homework 9

PSTAT 5A: Spring 2023, with Ethan P. Marzban

Instructions

- Please submit your work to Gradescope by no later than **11:59pm on Wednesday, June 7**. As a reminder, late homework will not be accepted.
- Recall that you will be asked to upload a **single** PDF containing your work for *both* the programming and non-programming questions to Gradescope.
  - You can merge PDF files using either Adobe Acrobat, or using adobe's online PDF merger at this link.

Caution

Be aware that some parts may be easier (or, in fact, may *need* to be) computed using Python. If you do use Python for any part, please write down the code you used.

Problem 1: Weight Loss

A new weight loss regimen claims to significantly reduce the weights of its participants. To test these claims, a researcher takes a representative sample of 100 volunteers, records their weights before the regimen, and then records their weights after the regimen. (All measurements are in lbs.) The summary statistics are displayed below:

	Sample Mean	**Sample Std. Dev.**
Pre-Regimen	151.2	21.3
Post-Regimen	145.2	19.1

Let Population 1 be the set of all pre-regimen weights, and Population 2 be the set of all post-regimen weights. Additionally, assume (for now) that all independence assumptions are satisfied.

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

b. State the null and alternative hypotheses. (Remember that the null can be thought of as the "status quo".) Think carefully about the alternative: as a hint, this is *not* a two-sided test.

c. Compute the 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?

f. 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.

g. **Critical Thinking:** Do you think our assumption that "all independence assumptions are satisfied" is justified? Why or why not?

Problem 2: ANOVA By Hand

In this problem, we will work through the computations of an ANOVA by hand. You must show all of your work clearly.

The data we will consider is:

$$\begin{aligned} \boldsymbol{x_1} &= \{-1,\ 0,\ 1\} \\ \boldsymbol{x_2} &= \{0,\ 1,\ 1,\ 2\} \\ \boldsymbol{x_3} &= \{1,\ 2,\ 1\} \end{aligned}$$

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).

d. Compute the **sum of squares between groups**:

$$\mathrm{SS_G} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$$

where n_i denotes the size of the i^{th} group.

e. Compute the **sum of squares total**:

$$\mathrm{SS_T} = \sum_{i=1}^{k} (x_i - \overline{x})^2$$

where the sum is taken over *all* observations in the dataset.

f. Compute the **sum of squared errors**:

$$\mathrm{SS_E} = \mathrm{SS_T} - \mathrm{SS}_{\mathrm{SS}_G}$$

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.

Problem 3: Filling In an ANOVA Table

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.

Part (a):

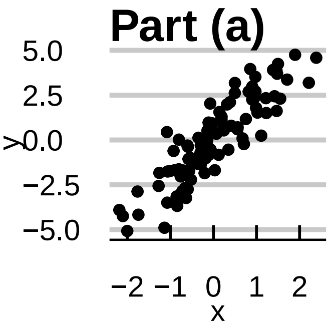
	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Btwn. Groups	4	10	<???>	<???>	<???>
Residuals	<???>	50	0.5		

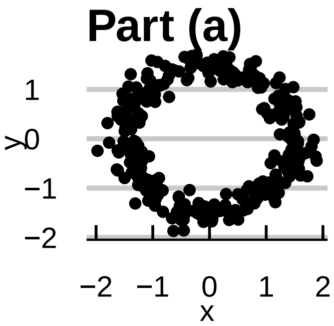
Part (b):

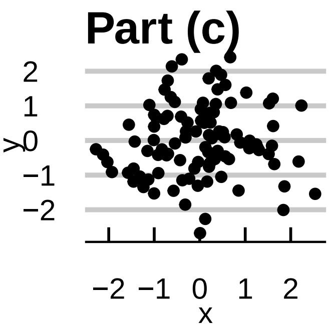
	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Btwn. Groups	10	20	2	<???>	0.8636
Residuals	120	<???>	<???>		

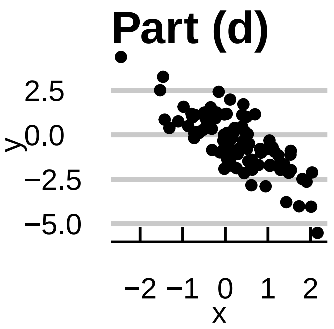
Problem 4: Graphical Correlations

In each of the following parts you will be presented with a scatterplot of two variables x and y. Based on the scatterplot, determine whether you believe the correlation between x and y to be positive, negative, or zero. Justify your answers.









Problem 5: Numerical Computations

Consider the following two sets of numbers:

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

a. 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 **NOT** use any more advanced functions like `numpy.std()`, or `np.mean()`.)

b. Compute the coefficients of the OLS regression line when regressing y onto x (i.e. treating y as the response variable and x as the explanatory variable). Again, only use Python as a calculator for arithmetic computations.

Problem 6: Programming

Part (a)

! Task 1

Write a function called `cor()` that takes in two lists `x` and `y`, and returns the Pearson's Correlation between `x` and `y`. Check that `cor([1, 2, 3], [1, 2, 3])` returns `1`.

Part (b)

! Task 1

Write a function called `reg()` that takes in two lists `x` and `y` and returns the OLS estimates of the intercept and slope of regressing `y` onto `x`. Your function should return a list with two elements: $\widehat{\beta}_0$ and $\widehat{\beta}_1$, in that order.