Name:________________________________ **Date:**________________

PSTAT 5A: Final Exam Practice Problems

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

1 Post-MT2

1. Jeremy believes that the average typing speed among gamers is the same as the average typing speed among non-gamers. To test this claim, he collects the typing speeds (in words per minute) of 31 randomly-selected gamers and 40 randomly-selected non-gamers. The results of his data are displayed below:

	Sample Mean	**Sample Std. Dev.**
Gamers	80	3.51
Non-Gamers	76	10.12

Assume all necessary independence and normality conditions hold, and use a 1% level of significance and a two-sided alternative wherever applicable. Additionally, let “Population 1” refer to “gamers” and “Population 2” refer to non-gamers.

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

(b) State the null and alternative hypotheses.

(c) Compute the value of the test statistic.

(d) Assuming the null is correct, what is the distribution of the test statistic? Be sure to include any/all relevant parameter(s).

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

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

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

(h) Redo the test, this time using the alternative hypothesis that the average typing speed of gamers is higher than that of non-gamers.

2. Consider the following lists of numbers:

$$\boldsymbol{x} = \{1, 4, 5, 7\}$$
$$\boldsymbol{y} = \{3, 5, 3, 5\}$$

(a) Compute $\text{corr}(\boldsymbol{x}, \ \boldsymbol{y})$, the Pearson's correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$.

For parts (b) - (e): Assume we now regress $\boldsymbol{y}$ onto $\boldsymbol{x}$ (i.e. we use $\boldsymbol{y}$ as the response variable and $\boldsymbol{x}$ as the explanatory variable), assuming a linear model.

(b) Compute $\widehat{\beta}_1$, the OLS estimate of the slope of the signal function.

(c) Compute $\widehat{\beta}_0$, the OLS estimate of the intercept of the signal function.

(d) Suppose a new $\boldsymbol{x}-$value of 6 is recorded. What would be the predicted corresponding $\boldsymbol{y}-$value?

(e) Suppose a new $\boldsymbol{x}-$value of 100 is recorded. Would using the OLS regression line to predict the corresponding $\boldsymbol{y}-$value be risky or not? (**Hint:** there is a specific term I'm looking for here.)

3. A particular PSTAT course has been split into two sections, called A and B. It is known that the final exam score of a randomly-selected person from section A is normally distributed with mean 8 and standard deviation 1.2; it is also known that the final exam score of a randomly-selected person from section B is normally distributed with mean 9 and standard deviation 2.3. Additionally, it is known that the scores between two sections are independent of each other.

   A student is selected at random from section A and another student is selected at random from section B, and the final exam scores of these two students is recorded.

   (a) Define the random variables of interest.

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

   (c) What is the probability that the student from section A scored exactly 2 points below the student from section B?

   (d) What is the probability that the student from section A scored more than 2 points below the student from section B?

   (e) What is the probability that the two students scored within 3 points of each other?

4. Leah is interested in determining whether students who listen to music while studying perform better (academically) than those who do not. To do so, she seeks out 50 people who regularly listen to music while studying and 50 that do not. She then collects the average GPA from each group to use as a metric of "performance in school".

   (a) Explain why this is an observational study, and not an experiment.

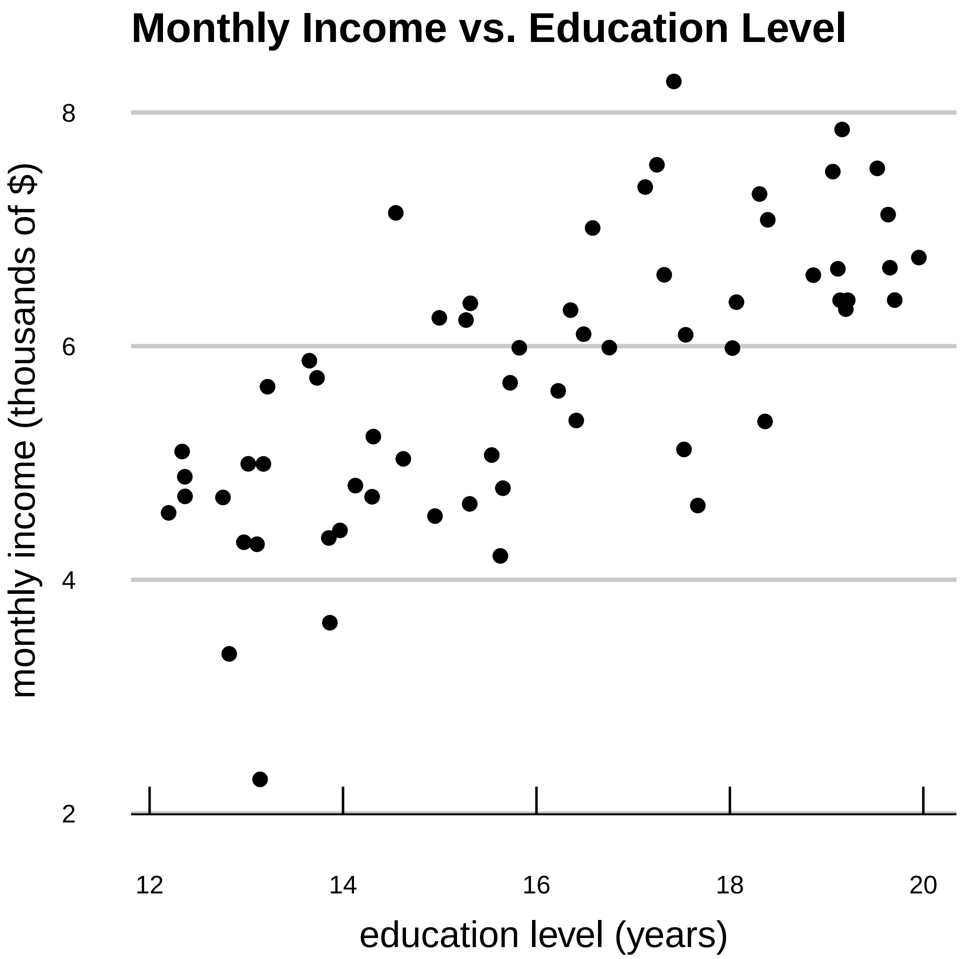
   (b) Briefly explain how Leah might restructure her study to conduct an experiment as opposed to an observational study.

   (c) Suppose that Leah is now interested in seeing whether the results of her study (i.e. whether listening to music while studying affects overall performance) varies between majors. What type of sampling procedure do you think Leah should carry out? Explain your reasoning.

Name:____ **Date:**____________________

(d) Suppose Leah's study reveals that students who listened to music tended while studying tended to have lower GPAs than those who did not listen to music while studying. Can Leah then conclude that there exists a causal relationship between "listening to music while studying" and "academic performance?" Why or why not?

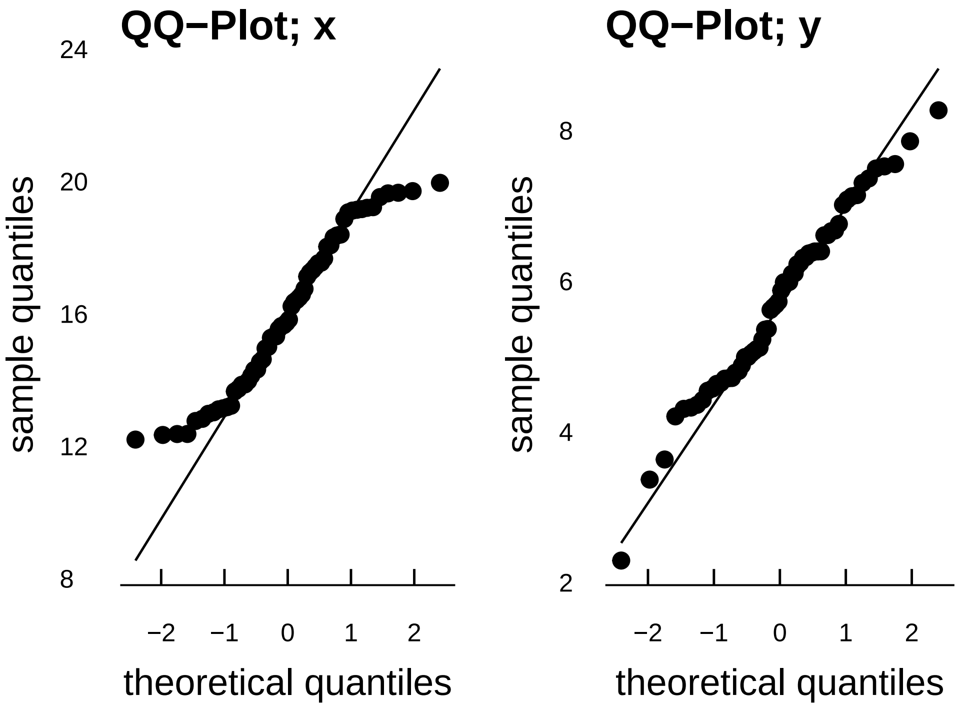
5. Tadhg would like to model the relationship between income and education level (as measured using years of education). He collects a sample of 62 people and records their education level (i.e. years of education) and average monthly income, and produces the following scatterplot from his data:



Additionally, the following numerical summaries of his data are provided:

$$\sum_{i=1}^{62} x_i = 992.7295 \qquad \sum_{i=1}^{62} (x_i - \overline{x})^2 = 343.1438$$
$$\sum_{i=1}^{62} y_i = 354.8923 \qquad \sum_{i=1}^{62} (y_i - \overline{y})^2 = 87.11993$$
$$\sum_{i=1}^{62} (x_i - \overline{x})(y_i - \overline{y}) = 122.4954$$

Finally, below are the QQ-plots of `education level` ($\mathtt{x}$) and `monthly income` ($\mathtt{y}$), respectively:



(a) Compute $\text{Cor}(x,\ y)$, the correlation between $\mathtt{x}$ (education level) and $\mathtt{y}$ (monthly income).

(b) Compute $\widehat{\beta_1}$, the slope of the OLS regression line.

(c) Compute $\widehat{\beta_0}$, the intercept of the OLS regression line.

(d) Provide an interpretation of your value of $\widehat{\beta_1}$. Specifically, what does a one-year change in education level correspond to with regards to a change in monthly income?

(e) It is known that $\text{Var}(\widehat{\beta_1}) = 0.002914$. Construct a 95% confidence interval for β_1, the slope of the true underlying linear relationship between $\mathtt{x}$ and $\mathtt{y}$. Interpret your confidence interval.

(f) What is the predicted monthly income (in thousands of dollars) of someone with 15.25 years of education?

(g) Is it dangerous to try and use the OLS regression line to predict the monthly income (in thousands of dollars) of someone with 27 years of education? (There is a specific word/term I'm looking for here.)

(h) Does x appear to be normally distributed? What about y? Why or why not (i.e. what *specifically* did you look at to answer this question)?

2 Miscellaneous (including additional Post-MT2 problems)

6. *Forbes* has claimed that 75% of British residents drink at least one cup of tea per day. To test this claim, Sean takes a representative sample of 80 British residents and finds that 64 of these people drink at least one cup of tea per day. Suppose Sean wishes to test *Forbes'* claim against an upper-tailed alternative, using a 5% level of significance.

   (a) Define the parameter of interest, and call it p.

   (b) State the null and alternative hypotheses.

   (c) Define the random variable of interest, and call it $\widehat{P}$.

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

   (e) Assuming the null is correct, what distribution does the test statistic follow? Be sure to check any/all relevant conditions, and include any/all relevant parameter(s).

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

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

   (h) Conduct the test, and phrase your conclusions in the context of the problem.

7. A single packet of *GauchoTea*-brand Matcha is advertised to contain 12 oz. of tea; in actually, the amount of tea included in a randomly-selected packet is normally distributed with mean 12 oz. and standard deviation 2.1 oz. A packet of *GauchoTea*-brand Matcha is selected, and the amount of tea it contains is recorded.

   (a) Define the random variable of interest.

   (b) Using proper notation, state the distribution of the random variable of interest.

   (c) What is the probability that this packet of tea contains exactly 12 oz. of tea?

   (d) What is the probability that this packet of tea contains between 11 oz. and 12.5 oz. of tea?

(e) Suppose now that a sample of 10 *GauchoTea*-brand Matcha packets is taken with replacement, and the number of these packets containing between 11 oz. and 12.5 oz. of tea is recorded. What is the probability that exactly 4 of these packets contain between 11 oz. and 12.5 oz. of tea? Check any/all conditions!

8. The random variable Y has the following probability mass function (p.m.f.):

$\boldsymbol{k}$	-2	-1	1	2
$\mathbb{P}(\boldsymbol{X} = \boldsymbol{k})$	0.1	0.1	0.5	$\boldsymbol{a}$

(a) What is the value of $\boldsymbol{a}$?

(b) What is $\mathbb{P}(X \geq 0)$?

(c) If $F_X(x)$ denotes the cumulative distribution function (c.d.f.) of X at x, what is the value of $F_X(1)$?

(d) What is $\mathbb{E}[X]$, the expected value of X?

(e) What is SD(X), the standard deviation of X?

9. Consider the following list of numbers:

$$X = \{-1, \ \ 0, \ \ 2.1, \ \ 3.9\}$$

(a) Compute $\overline{x}$, the mean of X.

(b) Compute s_X, the standard devaition of X.

(c) Compute the **sample kurtosis** of X, defined as

$$\widehat{\alpha}_4 = \left(\frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right]^2} \right) - 3$$

10. Sam believes that drinking excess caffeine can cause a raise in resting heartrate, and would like to statistically test this belief.

(a) Should Sam conduct an observational study or an experiment? Why or why not? (**Hint:** examine the statement of Sam's claims carefully.)

(b) Describe how Sam should structure their study/experiment if they want to conduct a cross-sectional study. Specifically, discuss how they should divide their participants (if at all), what "treatment" entails, and how treatment should be administered/withheld (if at all).

For parts (c) - (g): Assume Sam divides their participants into two groups, both of size 20. They ask members of group 1 to drink exactly three cups of caffeinated drinks per day for a week, and ask members of group 2 to refrain from drinking caffeinated drinks for a week. Sam then records the average resting heartrate (in beats per minute) of each group, and produces the following summaries of their data:

	Sample Mean	**Sample Std. Dev.**
Group 1	78.3	4.51
Group 2	75.7	3.23

Let μ_1 denote the average resting heartrate of participants in Group 1, and let μ_2 denote the average resting heartrate of participants in Group 2. Suppose Sam adopts the null that $\mu_1 = \mu_2$, and an alternative that $\mu_1 > \mu_2$. Assume all independence and normality conditions are met.

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

(d) the distribution that the test statistic follows under the null. Be sure to include any/all relevant parameter(s).

(e) Conduct the test at a 5% level of significance, and state your conclusions in the context of the problem.

(f) Conduct the test at a 1% level of significance, and state your conclusions in the context of the problem.

(g) Conduct the test at a 10% level of significance, and state your conclusions in the context of the problem.

11. Recall that the `palmerpenguins` dataset contains observations on 344 different penguins, collected at a weather station in Antarctica. Juno believes that the data included in the `palmerpenguins` dataset supports the claim that there is a relationship between a penguin's flipper length (in mm) and its species. Specifically, she believes that you can predict a penguin's species from its flipper length (in mm).

(a) What is the response variable?

(b) What is the explanatory variable?

(c) Assuming a linear signal function, what is Juno's claimed model? Use `y` for the response variable and `x` for the explanatory variable.

(d) Is the above model a regression model or a classification model? Explain (and there is a very specific explanation I'm looking for here- remember the definitions of regression and classification problems!)

(e) Would we be able to use the method of Ordinary Least Squares (OLS) to estimate the parameter(s) of the model from part (c)? Why or why not?