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.

Solution: Let μ_1 denote the average typing speed among gamers, and let μ_2 denote the average typing speed among non-gamers.

(b) State the null and alternative hypotheses.

Solution:

$$\begin{bmatrix} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{bmatrix} = \begin{bmatrix} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 \neq 0 \end{bmatrix}$$

(c) Compute the value of the test statistic.

Solution:

$$\text{ts} = \frac{\overline{y} - \overline{x}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{76 - 80}{\sqrt{\frac{3.51^2}{31} + \frac{10.12^2}{40}}} \approx -2.32$$

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

Solution: Since we are told to assume all necessary independence and normality conditions, we know that the test statistic will, under the null, follow a $t-$distribution with degrees of freedom given by the Satterthwaite approximation:

$$\text{df} = \text{round}\left\{\frac{\left[\left(\frac{s_X^2}{n_1}\right)+\left(\frac{s_Y^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2-1}}\right\}$$

$$= \text{round}\left\{\frac{\left[\left(\frac{3.51^2}{31}\right)+\left(\frac{10.12^2}{40}\right)\right]^2}{\frac{\left(\frac{3.51^2}{31}\right)^2}{31-1}+\frac{\left(\frac{10.51^2}{40}\right)^2}{40-1}}\right\} = \text{round}\{50.46623\} = 50$$

That is,

$$\text{TS} \overset{H_0}{\sim} t_{50}$$

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

Solution: We can consult our $t-$table, since 50 degrees of freedom *does* appear on it. Since we have a two-sided alternative and a 5% level of significance, we have a critical value of 2.01.

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

Solution: After importing the `scipy.stats` module as `sps`, we run

```
2 * sps.t.cdf(-2.32, 50)
```

which gives a $p-$value of around 0.024.

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

Solution: We can proceed either using critical values, or using $p-$values. We reject when either:

- the absolute value of the test statistic exceeds the critical value

- the $p-$value is less than the level of significance

Since both of these are true ($|-2.32| = 2.32 > 2.01$ and $p = 0.024 < 0.05$) we reject the null:

> At a 5% level of significance, there was sufficient evidence to reject the null that gamers and non-gamers have the same average typing speed in favor of the alternative that they do not have the same average typing speed.

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

Solution: Note that our hypotheses now become

$$\begin{bmatrix} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 > \mu_2 \end{bmatrix} = \begin{bmatrix} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 < 0 \end{bmatrix}$$

which is effectively a lower-tailed test. Therefore, the critical value becomes -1.68 and the $p-$value becomes

$$\texttt{sps.t.cdf(-2.32, 50)} = 0.012$$

and we again reject the null:

> At a 5% level of significance, there was sufficient evidence to reject the null that gamers and non-gamers have the same average typing speed in favor of the alternative that they do not have the same average typing speed.

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 $\boldsymbol{y}$.

Solution:

$$\overline{x} = \frac{1}{4}(1+4+5+7) = \frac{17}{4}$$
$$\overline{y} = \frac{1}{4}(3+5+3+5) = 4$$

Name: _____ **Date:** _____

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$= \frac{1}{3}\left[\left(1 - \frac{17}{4}\right)^2 + \left(4 - \frac{17}{4}\right)^2 + \left(5 - \frac{17}{4}\right)^2 + \left(7 - \frac{17}{4}\right)^2\right] = \frac{1}{3}\cdot\frac{75}{4} = \frac{25}{4}$$
$$s_X = \sqrt{\frac{25}{4}} = \frac{5}{2}$$
$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$
$$= \frac{1}{3}\left[(3-4)^2 + (5-4)^2 + (3-4)^2 + (5-4)^2\right] = \frac{1}{3}\cdot(4) = \frac{4}{3}$$
$$s_Y = \sqrt{\frac{4}{3}} = \frac{2}{\sqrt{3}} = \frac{2\sqrt{3}}{3}$$

Thus, we have

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right) = \frac{1}{3}\left[\left(\frac{1-\frac{17}{4}}{\frac{5}{2}}\right)\left(\frac{3-4}{\frac{2\sqrt{3}}{3}}\right) + \left(\frac{4-\frac{17}{4}}{\frac{5}{2}}\right)\left(\frac{5-4}{\frac{2\sqrt{3}}{3}}\right)\right.$$
$$\left.\left(\frac{5-\frac{17}{4}}{\frac{5}{2}}\right)\left(\frac{3-4}{\frac{2\sqrt{3}}{3}}\right) + \left(\frac{7-\frac{17}{4}}{\frac{5}{2}}\right)\left(\frac{5-4}{\frac{2\sqrt{3}}{3}}\right)\right]$$
$$= \frac{1}{3}\left[\left(-\frac{13}{4}\cdot\frac{2}{5}\right)\left(-\frac{3}{2\sqrt{3}}\right) + \left(-\frac{1}{4}\cdot\frac{2}{5}\right)\left(\frac{3}{2\sqrt{3}}\right)\right.$$
$$\left. + \left(\frac{3}{4}\cdot\frac{2}{5}\right)\left(-\frac{3}{2\sqrt{3}}\right) + \left(\frac{11}{4}\cdot\frac{2}{5}\right)\left(\frac{3}{2\sqrt{3}}\right)\right]$$
$$= \frac{1}{3}\left(\frac{39}{20\sqrt{3}} - \frac{3}{20\sqrt{3}} - \frac{9}{20\sqrt{3}} + \frac{33}{20\sqrt{3}}\right)$$
$$= \frac{1}{3}\cdot\frac{60}{20\sqrt{3}} = \sqrt{\frac{1}{3}}$$

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

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

Solution: It will be easiest to use the second formula for $\widehat{\beta}_1$, which enables us to compute

$$\widehat{\beta}_1 = \frac{s_Y}{s_X}\cdot r = \frac{\frac{2\sqrt{3}}{3}}{\frac{5}{2}}\cdot\frac{1}{\sqrt{3}} = \frac{2\sqrt{3}}{3}\cdot\frac{2}{5}\cdot\frac{1}{\sqrt{3}} = \frac{4}{15}$$

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

> **Solution:**
> $$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \cdot \overline{x} = 4 - \frac{4}{15} \cdot \frac{17}{4} = 4 - \frac{17}{15} = \frac{43}{15}$$

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

> **Solution:** Combining our answers from parts (b) and (c) above, we find the equation of the OLS regression line to be
> $$\widehat{y} = \frac{43}{15} + \frac{4}{15}x = \frac{43 + 4x}{15}$$
> Hence, the predicted y−value of an x−value of 6 is
> $$\widehat{y}^{(6)} = \frac{43 + 4 \cdot 6}{15} = \frac{67}{15}$$

(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.)

> **Solution:** Note that the x−values included in the dataset range from 1 to 7. As such, an x−value of 100 is very far away from the x−values observed and as such trying to use the OLS line to predict the corresponding y−value runs the risk of **extrapolation**.

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.

   > **Solution:** Let X denote the score of the randomly-selected student from section A and let Y denote the score of the randomly-selected student from section B.

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

Solution:

$$X \sim \mathcal{N}(8,\ 1.2); \quad Y \sim \mathcal{N}(9,\ 2.3)$$

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

Solution: To say that the student from section A scored exactly 2 points below the student from section B is to say that $X = Y - 2$ which, in terms of the difference $D = Y - X$, is equivalent to $D = 2$. We know that D will be normally distributed, and, hence, will be continuous; therefore, $\mathbb{P}(D = 2) = 0$ since the probability that a continuous random variable equals any specific value is 0.

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

Solution: We now wish to find the probability $\mathbb{P}(X < Y - 2)$, which can be written as $\mathbb{P}(D > 2)$ where $D = Y - X$. From our result on the linear combination of normally-distribtued random variables, we have

$$D \sim \mathcal{N}\left(8 - 9,\ \sqrt{1.2^2 + 2.3^2}\right) \sim \mathcal{N}\left(-1,\ \sqrt{1.2^2 + 2.3^2}\right)$$

and therefore

$$\begin{aligned}
\mathbb{P}(D > 2) &= \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} > \frac{2+1}{\sqrt{1.2^2+2.3^2}}\right) \\
&= \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} > 1.16\right) \\
&= 1 - \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} \leq 1.16\right) = 1 - 0.8770 = 0.123 = 12.3\%
\end{aligned}$$

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

Solution: Now we seek $\mathbb{P}(|D| < 3)$, which we compute using the same methodology used in Lecture:

$$\begin{aligned}
\mathbb{P}(|D| < 3) &= \mathbb{P}(-3 < D < 3) \\
&= \mathbb{P}(D < 3) - \mathbb{P}(D < -3) \\
&= \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} < \frac{3+1}{\sqrt{1.2^2+2.3^2}}\right) - \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} < \frac{-3+1}{\sqrt{1.2^2+2.3^2}}\right)
\end{aligned}$$

Name: _____ **Date:** _____

$$= \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} < 1.54\right) - \mathbb{P}\left(\frac{D+1}{\sqrt{1.2^2+2.3^2}} < -0.77\right)$$
$$= 0.9382 - 0.2206 = 0.7176 = 71.76\%$$

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.

   Solution: Since Leah has not explicitly told one group to listen to music while studying, nor has she asked one group to *not* listen to music while studying (i.e. treatment has been neither administered nor withheld), this is an observational study as opposed to an experiment.

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

   Solution: If Leah wanted to conduct an experiment instead of an observational study, she could do the following:

   > Split her group of volunteers into two groups. To one group, she can ask participants to listen to music while studying and to the other she can ask participants to not listen to music while studying. She can then record the average GPA of each group.

   In this way, she is able to control the administration of treatment, which is the hallmark of an experiment.

   (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.

   Solution: There are two potentially correct answers. One argument could be made that Simple Random Sampling is sufficient, as everyone will have an equal chance of being included in the study. However, by chance alone, one major could be either over- or under-represented in the study. As such, a potentially "better" sampling technique to use would

be stratified sampling, with each major assigned to a unique stratum. Cluster sampling is not a good idea, as we want all majors to be included in the study (whereas in cluster sampling not all majors will be included.)

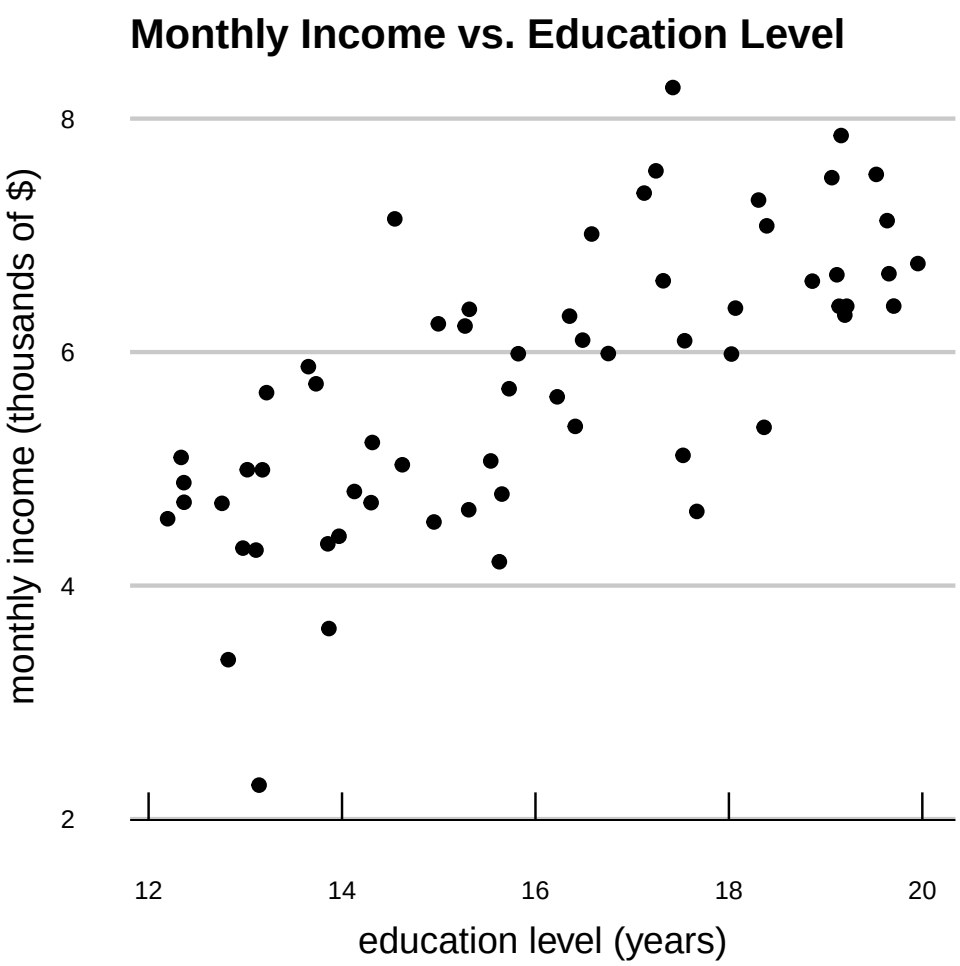
(d) Suppose Leah's study reveals that students who listened to music tended while studying tended to have lower GPAs than those who did not listen to music while studying. Can Leah then conclude that there exists a causal relationship between "listening to music while studying" and "academic performance?"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

Solution: Recall that observational studies cannot be used to establish causal relationships. As Leah has conducted an observational study (and not an experiment), she cannot use her results to make causal claims.

Furthermore, association does not necessarily imply causation, even in an experiment! That is to say, even if Leah had conducted an experiment and still observed an association between "listening to music while studying" and "GPA", she still wouldn't be able to conclude a causal link between the two variables.

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:

Name:_____ **Date:**_____

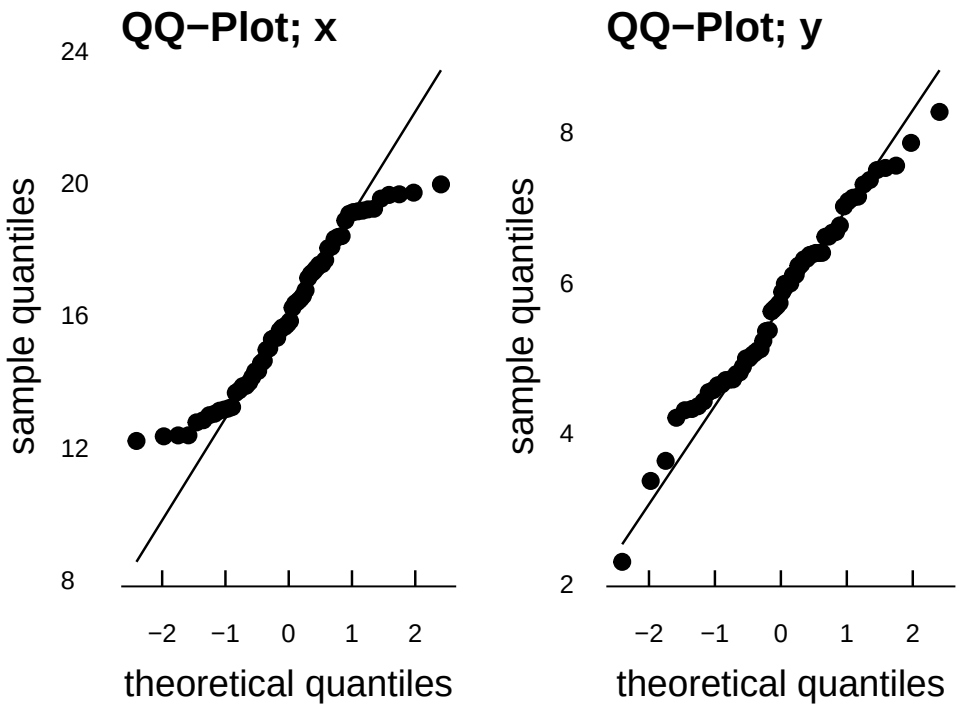


Additionally, the following numerical summaries of his data are provided:

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

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

Name: _____ **Date:** _____



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

Solution:

$$\begin{aligned}
\text{Cor}(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_X}\right)\left(\frac{y_i-\overline{y}}{s_Y}\right) \\
&= \frac{1}{n-1}\cdot\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_X\cdot s_Y} \\
&= \frac{1}{n-1}\cdot\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}} \\
&= \frac{1}{61}\cdot\frac{(122.4954)}{\sqrt{\frac{1}{61}(343.1438)\cdot\frac{1}{61}(87.11993)}} \approx 0.7084724
\end{aligned}$$

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

Solution:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{122.4954}{343.1438} \approx 0.35697$$

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

Solution:

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta}(\overline{x}) = \frac{354.8923}{62} - (0.35697) \cdot \frac{992.7295}{62} \approx \boxed{0.00835}$$

(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?

Solution: Recall that one interpretation of $\widehat{\beta_1}$ in the OLS regression line is that a one-unit increase in x corresponds to a (predicted) $\widehat{\beta_1}$ unit increase in y. As such, the interpretation of the 0.35697 found above is that "a one-year increase in education level corresponds to a predicted 0.35697 thousand-dollar increase in monthly income".

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

Solution: Assuming all independence and normality conditions hold, we know that

$$\frac{\widehat{\beta_1} - \beta_1}{\mathrm{SD}(\widehat{\beta_1})} \sim t_{n-2}$$

Hence, our confidence interval take the form

$$\widehat{\beta_1} \pm c \cdot \mathrm{SD}(\widehat{\beta_1})$$

where c is the 97.5^{th} (i.e. $1 - (1 - 0.95)/2$) percentile of the t_{60} distribution. From our $t-$table (specifically, looking at the row with $\mathrm{df} = 60$ and the one-tailed 0.025 column) we see that this value is 2.00, meaning our confidence interval is

$$(0.35697) \pm (2.00)\sqrt{0.002914} = \boxed{[0.249007\ ,\ 0.464933]}$$

One interpretation of this interval is:

> We are 95% confident that the true slope of the linear relationship between education level (in years) and monthly income (in thousands of dollars) is between 0.249007 and 0.464933.

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

Name:_____ **Date:**_____

Solution: We use the OLS regression line:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x \\ \widehat{y}^{(15.25)} &= \widehat{\beta}_0 + \widehat{\beta}_1 \cdot (15.25) \\ &= 0.00835 + (0.35697)(15.25) = \boxed{5.452142}\end{aligned}$$

(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.)

Solution: Since the x values included in the dataset range between 12 and 20 (as indicated by the scatterplot), trying to use the OLS regression line to predict the income of someone who has 27 years of education *is* risky as we are at risk of performing **extrapolation**.

(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)?

Solution: To answer questions relating to normality, we need to look at the **QQ-Plots**. Recall that deviations from linearity in a QQ-plot, especially near the ends of the plot, indicate non-normality. The QQ-plot for y appears roughly linear, so it is safe to assume y *is* normally distributed. For the QQ-plot of x, however, we see some marked deviations from linearty at both ends of the plot, leading us to believe that x was likely *not* normally distributed.

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.

Solution: Let p denote the proportion of British residents that drink at least one cup of tea per day.

(b) State the null and alternative hypotheses.

Solution:

$$\begin{cases} H_0: & p = 0.75 \\ H_A: & p > 0.75 \end{cases}$$

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

Solution: Let $\widehat{P}$ denote the proportion of a representative sample of 80 British residents that drink at least one cup of tea per day.

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

Solution:

$$\text{ts} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\left(\frac{64}{80}\right) - 0.75}{\sqrt{\frac{0.75 \cdot 0.25}{80}}} = 1.03$$

(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).

Solution: We would like to invoke the Central Limit Theorem for proportions. As such, we need to check:

1) $np_0 = (80) \cdot (0.75) = 60 \geq 10$ ✓

2) $n(1 - p_0) = (80) \cdot (0.25) = 20 \geq 10$ ✓

Since both conditions are met, we conclude that

$$\text{TS} = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} \mathcal{N}(0,\ 1)$$

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

Solution: Since we are using an upper-tailed alternative and a 5% level of significance, the critical value is the 95^{th} percentile of the standard normal distribution (not scaled by anything), which we see to be around 1.645.

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

> **Solution:** Since we are using an upper-tailed alternative, our $p-$value is computed as a right-tail area. That is, if $Z \sim \mathcal{N}(0,\ 1)$ then our $p-$value is
>
> $$\mathbb{P}(Z > 1.03) = 1 - \mathbb{P}(Z \leq 1.03) = 1 - 0.8485 = 0.1515$$

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

> **Solution:** We can proceed either using critical values, or using $p-$values. We reject when either:
>
> - the raw value of the test statistic exceeds the critical value
> - the $p-$value is less than the level of significance
>
> Since neither of these are true ($1.03 \not> 1.645$ and $p = 0.1515 \not< 0.05$) we fail to reject the null:
>
> > At a 5% level of significance, there was insufficient evidence to reject the null that the true proportion of British residents that drink tea is 75% in favor of the alternative that the true proportion is greater than 75%.

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.

> **Solution:** Let X denote the amount of tea (in ounces) contained in a randomly-selected packet of *GauchoTea*-brand Matcha.

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

> **Solution:**
>
> $$X \sim \mathcal{N}(8,\ 2.1)$$

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

Name:_____ **Date:**_____

Solution: We seek $\mathbb{P}(X = 12)$. Since X is continuous, we know that $\mathbb{P}(X = k) = 0$ for any value of k: hence, the desired probability is 0.

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

Solution: We seek $\mathbb{P}(11 \leq X \leq 12.5)$, which we compute using our standard procedure:

$$\begin{aligned}
\mathbb{P}(11 \leq X \leq 12.5) &= \mathbb{P}(X \leq 12.5) - \mathbb{P}(X \leq 11) \\
&= \mathbb{P}\left(\frac{X-12}{2.1} \leq \frac{12.5-12}{2.1}\right) - \mathbb{P}\left(\frac{X-12}{2.1} \leq \frac{11-12}{2.1}\right) \\
&= \mathbb{P}\left(\frac{X-12}{2.1} \leq 0.24\right) - \mathbb{P}\left(\frac{X-12}{2.1} \leq -0.48\right) \\
&= 0.5948 - 0.3156 = 0.2792
\end{aligned}$$

(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!

Solution: Let Y denote the number of packets, in the sample of 10, that contain between 11 and 12.5 oz of tea. We suspect Y to be Binomially distributed- to verify this, we check the four Binomial conditions:

1) **Independent trials?** Yes, since sampling is done with replacement.
2) **Fixed number of trials?** Yes; $n = 10$.
3) **Well-defined notion of "success"?** Yes; success on any given trial (i.e. for any given packet) is "packet contains between 11 and 12.5 oz of tea".
4) **Fixed probability of "success"?** Yes; $p = 0.2792$, as found in part (d).

Since all four conditions are met, we conclude

$$Y \sim \text{Bin}(10,\ 0.2792)$$

and so

$$\mathbb{P}(Y = 4) = \binom{10}{4}(0.2792)^4(1-0.2792)^6 \approx 0.1789 = 17.89\%$$

Name:_____ **Date:**_____

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

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

(a) What is the value of a?

Solution: We know that the probability values in a p.m.f. must sum to 1. This yields

$$0.1 + 0.1 + 0.5 + a = 1 \implies a = 1 - (0.1 + 0.1 + 0.5) = 0.3$$

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

Solution: By direct computation,

$$\mathbb{P}(X \geq 0) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 0.5 + 0.3 = 0.8$$

Or, by the complement rule,

$$\begin{aligned}\mathbb{P}(X \geq 0) &= 1 - \mathbb{P}(X < 0) = 1 - \mathbb{P}(X \leq -1) \\ &= 1 - [\mathbb{P}(X = 1) + \mathbb{P}(X = -2)] \\ &= 1 - (0.1 + 0.1) = 0.8\end{aligned}$$

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

Solution: By definition, $F_X(1) := \mathbb{P}(X \leq 1)$. By direct computation, we have

$$F_X(1) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 0.5 + 0.3 = 0.8$$

By the complement rule, we have

$$F_X(1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}(X \leq -1) = 1 - (0.1 + 0.1) = 0.8$$

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

Solution:

$$\mathbb{E}[X] = \sum_{\text{all } k} k \cdot \mathbb{P}(X = k)$$

$$= (-2) \cdot \mathbb{P}(X = -2) + (-1) \cdot \mathbb{P}(X = -1) + (1) \cdot \mathbb{P}(X = 1) + (2) \cdot \mathbb{P}(X = -2)$$
$$= (-2) \cdot (0.1) + (-1) \cdot (0.1) + (1) \cdot (0.5) + (2) \cdot (0.3)$$
$$= 0.8$$

(e) What is $\text{SD}(X)$, the standard deviation of X?

Solution: We first find the variance of X. Using the second formula for variance, we would first compute

$$\sum_{\text{all } k} k^2 \cdot \mathbb{P}(X = k) = (-2)^2 \cdot \mathbb{P}(X = -2) + (-1)^2 \cdot \mathbb{P}(X = -1) + (1)^2 \cdot \mathbb{P}(X = 1) + (2)^2 \cdot \mathbb{P}(X = -2)$$
$$= (-2)^2 \cdot (0.1) + (-1)^2 \cdot (0.1) + (1)^2 \cdot (0.5) + (2)^2 \cdot (0.3) = 2.2$$

and so

$$\text{Var}(X) = \left(\sum_{\text{all } k} k^2 \cdot \mathbb{P}(X = k) \right) - (\mathbb{E}[X])^2 = 2.2 - (0.8)^2 = 1.56$$

Using the first formula for variance:

$$\text{Var}(X) = \sum_{\text{all } k} (k - \mathbb{E}[X])^2 \cdot \mathbb{P}(X = k)$$
$$= (-2 - 0.8)^2 \cdot \mathbb{P}(X = -2) + (-1 - 0.8)^2 \cdot \mathbb{P}(X = -1) + (1 - 0.8)^2 \cdot \mathbb{P}(X = 1) + (2 - 0.8)^2 \cdot \mathbb{P}(X = -2)$$
$$= (-2 - 0.8)^2 \cdot (0.1) + (-1 - 0.8)^2 \cdot (0.10) + (1 - 0.8)^2 \cdot (0.5) + (2 - 0.8)^2 \cdot (0.3) = 1.56$$

Either way we find $\text{Var}(X) = 1.56$ and so

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{1.56} = 1.249$$

9. Consider the following list of numbers:

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

Name: _____ **Date:** _____

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

Solution:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= \frac{1}{4}(-1 + 0 + 2.1 + 3.9) = \frac{5}{4}$$

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

Solution:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$= \frac{1}{3}\left[\left(-1 - \frac{5}{4}\right)^2 + \left(0 - \frac{5}{4}\right)^2 + \left(2.1 - \frac{5}{4}\right)^2 + \left(3.9 - \frac{5}{4}\right)^2\right] = \frac{1}{3} \cdot \frac{1437}{100} = \frac{479}{100}$$
$$s_X = \sqrt{s_X^2} = \frac{\sqrt{479}}{10} \approx 2.1886$$

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

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

Solution: Let's focus on the denominator first. Note that:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n} \cdot \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$= \left(\frac{n-1}{n}\right) \cdot s_X^2$$
$$= \frac{3}{4} \cdot \frac{479}{100} = \frac{1437}{400}$$

Hence, the denominator is simply

$$\left(\frac{1437}{400}\right)^2 = \frac{2064969}{160000}$$

For the numerator, we compute

$$\frac{1}{4}\left[\left(-1-\frac{5}{4}\right)^4+\left(0-\frac{5}{4}\right)^4+\left(2.1-\frac{5}{4}\right)^4+\left(3.9-\frac{5}{4}\right)^4\right]=\frac{3116313}{160000}$$

Therefore,

$$\begin{aligned}\widehat{\alpha}_4 &= \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2}\right)-3\\ &= \left(\frac{3116313}{160000}\cdot\frac{160000}{2064969}\right)-3\\ &= \frac{3116313}{2064969}-3=-\frac{342066}{229441}\approx -1.491\end{aligned}$$

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.)

    Solution: Recall that causal relationships can only be made from data resulting from an experiment, not an experiment. Note that Sam wants to test a *causal* claim, meaning they should perform an experiment.

    (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).

    Solution: To conduct a cross-sectional study, there should be no tracking over time. Hence, Sam should start by dividing volunteers into two groups, which they can call Group 1 and Group 2. Sam should instruct members of one of the groups to drink caffeine regularly (perhaps over some fixed set of time) and then record the heartrates of participants in the group afterwards. Similarly, Sam should instruct members of the other group to refrain from drinking caffeine regularly (over the same set of time as with the first group) and then record the heartrates of participants in the group afterwards.

Name: _____ **Date:** _____

For illustrative purposes, here is how Sam would structure their experiment if they were conducting a longitudinal study. There would be no division of groups: rather, Sam would record the heartrates of all participants at the start of the study, then instruct all participants to drink caffeine regularly, and finally record the post-treatment heartrates of all participants.

The real key is to think about the data. In both cases, Sam would end up with pairs of observations $(x_i,\ y_i)$ where x_i is a pre-treatment heartrate and y_i is a post-treatment heartrate. In the cross-sectional study, x_i and y_i would be taken from *different* individuals, whereas in the longitudinal study x_i and y_i would be taken from the *same* individual (and would therefore be correlated).

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.

Solution:

$$\text{ts} = \frac{\overline{y} - \overline{x}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{75.7 - 78.3}{\sqrt{\frac{4.51^2}{20} + \frac{3.23^2}{20}}} \approx \boxed{-2.25}$$

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

Solution: Since we are told to assume all necessary independence and normality conditions, we know that the test statistic will, under the null, follow a $t-$distribution with

Name:_____ **Date:**_____

degrees of freedom given by the Satterthwaite approximation:

$$\text{df} = \text{round}\left\{\frac{\left[\left(\frac{s_X^2}{n_1}\right)+\left(\frac{s_Y^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2-1}}\right\}$$

$$= \text{round}\left\{\frac{\left[\left(\frac{4.51^2}{20}\right)+\left(\frac{3.23^2}{20}\right)\right]^2}{\frac{\left(\frac{4.51^2}{20}\right)^2}{20-1}+\frac{\left(\frac{3.23^2}{20}\right)^2}{20-1}}\right\} = \text{round}\{34.43126\} = 34$$

That is,

$$\text{TS} \overset{H_0}{\sim} t_{34}$$

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

Solution: On an exam, we would need to proceed using critical values (since we wouldn't be able to compute $p-$values under the $t-$distribution without the use of Python, except in certain special circumstances). So, let's use critical values to answer parts (e) - (g).

Since we are assuming a two-sided alternative, our critical value becomes 2.03. Since $|\text{ts}| = |-2.25| = 2.25 > 2.03$, we reject the null:

> At a 5% level of significance, there was sufficient evidence to reject the claim that the average resting heartrate of people who regularly drink caffeine is different than the average resting heartrate of people who do not regularly drink caffeine, in favor of the alternative that the two groups have different average heartrates.

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

Solution: Now we are using a 1% level of significance (still with a two-sided alternative) meaning our critical value becomes 2.73. Since $|\text{ts}| = |-2.25| = 2.25 \not> 2.73$, we fail to reject the null:

> At a 1% level of significance, there was insufficient evidence to reject the claim that the average resting heartrate of people who regularly drink caffeine is

different than the average resting heartrate of people who do not regularly drink caffeine, in favor of the alternative that the two groups have different average heartrates.

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

Solution: Now we are using a 10% level of significance (still with a two-sided alternative) meaning our critical value becomes 1.69. Since $|ts| = |-2.25| = 2.25 > 1.69$, we reject the null:

At a 10% level of significance, there was sufficient evidence to reject the claim that the average resting heartrate of people who regularly drink caffeine is different than the average resting heartrate of people who do not regularly drink caffeine, in favor of the alternative that the two groups have different average heartrates.

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?

Solution: The response variable is `species`.

(b) What is the explanatory variable?

Solution: The response variable is `flipper length (mm)`.

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

Solution: In general, a univariate model states

$$\texttt{y} = f(\texttt{x}) + \texttt{noise}$$

Name:_____ **Date:**_____

If we assume a linear form for the signal function $f()$, we then have

$$\texttt{y} = \beta_0 + \beta_1 \cdot \texttt{x} + \texttt{noise}$$

(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!)

Solution: Since the response variable (`species`) is categorical, this is a classification problem.

(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?

Solution: No, since OLS only works in a regression setting (i.e. where both the explanatory and response variables are numerical).