Score: _____ / **0**

PSTAT 5A / **FINAL EXAM** / Spring 2023 Instructor: **Ethan Marzban**

Name:_____ **UCSB NetID:**_____
First, then Last *NOT your Perm Number!*

Circle the section you attend:

Yuan 10 - 10:50am Jason 11 - 11:50am Nickolas 12 - 12:50pm Nickolas 1 - 1:50pm

Your Seat Number:_____

SAMPLE FREE RESPONSE QUESTIONS

Instructions:

- You will have **180 minutes** to complete the entire exam
  - Do not begin working on the exam until instructed to do so.
  - During the final 10 minutes of the exam, we will ask everyone to remain seated until the exam concludes.
- This exam comes in **TWO PARTS**: this is the **FREE RESPONSE** part of the exam.
  - There is a separate booklet containing Multiple Choice questions that should have been distributed to you at the same time as this booklet.
- Write your answers directly in the space provided on this exam booklet.
  - You do not need to write anything on your scantron for this part of the exam.
- Be sure to show all of your work; correct answers with no supporting work will not receive full credit.
- You are allowed the use of two **8.5** $\times$ **11-inch** sheets, front and back, of notes. You are also permitted the use of **calculators**; the use of any and all other electronic devices (laptops, cell phones, etc.) is prohibited.
- **PLEASE DO NOT DETACH ANY PAGES FROM THIS EXAM.**
- Good Luck!!!

---

1. Leonard believes that PSTAT students are just as good at bowling as Math students. To test this, he organizes a bowling match in which a group of 50 PSTAT students and a group of 50 Math students independently and simultaneously each played a game of bowling. The average number of points was used as a metric of assessing how good each team is at bowling: as such, Leonard collected the following information:

	Sample Mean	**Sample Stnd. Dev.**
Math	61	10.0
PSTAT	65	13.5

(a) Classify this as either an observational study or an experiment. Explain your reasoning.

> **Solution:** Since treatment was neither administered nor withheld (i.e. it was not the case that Leonard assigned one team to bowl and another to not), this is an observational study.

(b) Classify this as either a Longitudinal or Cross-Sectional study. Explain your reasoning.

> **Solution:** Since students were not tracked over time, this is an example of a cross-sectional study.

Parts (c) - (h) refer to the following: Suppose that Leonard now wishes to statistically test his claims against a two-sided alternative using a 5% level of significance. Assume all normality and independence assumptions hold. Additionally, let Population 1 refer to Math students and Population 2 refer to PSTAT students.

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

> **Solution:** $\mu_1 =$ the average number of points of Math students; $\mu_2 =$ the average number of points of PSTAT students.

(d) Write down the null and alternative hypotheses.

> **Solution:** The null hypothesis is that the two teams are equally skilled; i.e. that $\mu_1 = \mu_2$. Since we are told to use a two-sided alternative, we adopt the alternative $\mu_1 \neq \mu_2$: i.e.
>
> $$\begin{bmatrix} H_0: & \mu_1 = \mu_2 \\ H_A: & \mu_1 \neq \mu_2 \end{bmatrix}$$
>
> which, if we wanted to, we could phrase in terms of the difference of means:
>
> $$\begin{bmatrix} H_0: & \mu_2 - \mu_1 = 0 \\ H_A: & \mu_2 - \mu_1 \neq 0 \end{bmatrix}$$

(e) Compute the value of the test statistic.

Solution: The test statistic we use is

$$\text{TS} = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{65 - 61}{\sqrt{\frac{10^2}{50} + \frac{13.5^2}{50}}} = 1.684$$

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

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

$$\begin{aligned}
\text{df} &= \text{round}\left\{ \frac{\left[\left(\frac{s_X^2}{n_1}\right) + \left(\frac{s_Y^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2 - 1}} \right\} \\
&= \text{round}\left\{ \frac{\left[\left(\frac{10^2}{50}\right) + \left(\frac{13.5^2}{50}\right)\right]^2}{\frac{\left(\frac{10^2}{50}\right)^2}{50 - 1} + \frac{\left(\frac{13.5^2}{50}\right)^2}{50 - 1}} \right\} \\
&= \text{round}\{90.32934\} = 90
\end{aligned}$$

That is:

$$\text{TS} \overset{H_0}{\sim} t_{90}$$

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

Solution: We use our $t-$table. Specifically, since we have 90 degrees of freedom we look at the row corresponding to 90 degrees of freedom; since we are performing a *two*-sided test at a 5% level of significance we look at the "two-tailed 0.050" column to see the critical value is 1.99.

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

Solution: Since we are performing a two-sided test, we reject the null whenever the absolute value of the test statistic exceeds the critical value. Since $|\text{TS}| = |1.684| = 1.684 < 1.99$, we fail to reject the null:

> At a 5% level of significance, there was insufficient evidence to reject the null hypothesis that PSTAT students and

Math students are equally good at bowling, in favor of the alternative that they are not.

2. Consider a random variable X with the following probability mass function:

k	-3	2	0	3
$\mathbb{P}(X=k)$	0.25	0.25	a	0.25

(a) What must be the value of a?

Solution: The probability values in a probability mass function (p.m.f.) must sum to 1; as such, we must have

$$0.25 + 0.25 + a + 0.25 = 1 \implies a = 0.25$$

(b) What is $\mathbb{P}(X \leq 1)$?

Solution:

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = -3) + \mathbb{P}(X = 0) = 0.5$$

Or,

$$\mathbb{P}(X \leq 1) = 1 - \mathbb{P}(X > 1) = 1 - \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = 1 - 0.5 = 0.5$$

(c) Compute $\mathbb{E}[X]$, the expected value of X.

Solution:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\text{all } k} k \cdot \mathbb{P}(X = k) \\ &= (-3) \cdot \mathbb{P}(X = -3) + (2) \cdot \mathbb{P}(X = 2) + (0) \cdot \mathbb{P}(X = 0) + (3) \cdot \mathbb{P}(X = 3) \\ &= (0.25) \cdot (-3 + 0 + 2 + 3) = 0.5\end{aligned}$$

(d) Compute $\text{SD}(X)$, the standard deviation of X.

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

$$\begin{aligned}\sum_{\text{all } k} k^2 \cdot \mathbb{P}(X = k) &= (-3)^2 \cdot \mathbb{P}(X = -3) + (2)^2 \cdot \mathbb{P}(X = 2) \\ &\quad + (0)^2 \cdot \mathbb{P}(X = 0) + (3)^2 \cdot \mathbb{P}(X = 3) \\ &= (0.25) \cdot \left[(-3)^2 + (0)^2 + (2)^2 + (3)^2\right] = 5.5\end{aligned}$$

meaning

$$\text{Var}(X) = \left(\sum_{\text{all } k} k^2 \cdot \mathbb{P}(X = k)\right) - (\mathbb{E}[X])^2 = 5.5 - (0.5)^2 = 5.25$$

Or, using the first formula for variance, we would compute

$$\begin{aligned}
\text{Var}(X) &= \sum_{\text{all } k} (k - \mathbb{E}[X])^2 \cdot \mathbb{P}(X = k) \\
&= (-3 - 0.5)^2 \cdot \mathbb{P}(X = -3) + (2 - 0.5)^2 \cdot \mathbb{P}(X = 2) \\
&\qquad + (0 - 0.5)^2 \cdot \mathbb{P}(X = 0) + (3 - 0.5)^2 \cdot \mathbb{P}(X = 3) \\
&= (0.25) \cdot \left[(-3 - 0.5)^2 + (2 - 0.5)^2 + (0 - 0.5)^2 + (3 - 0.5)^2\right] \\
&= 5.25
\end{aligned}$$

Either way, we find $\text{Var}(X) = 5.25$ meaning

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{5.25} \approx 2.2913$$

(e) If $F_X(x)$ denotes the cumulative distribution function of X, what is $F_X(0)$?

Solution: Recall that $F_X(x) := \mathbb{P}(X \leq x)$. Hence,

$$\begin{aligned}
F_X(0) &= \mathbb{P}(X \leq 0) \\
&= \mathbb{P}(X = -3) + \mathbb{P}(X = 0) = (0.25) + (0.25) = 0.5
\end{aligned}$$

3. Recent scientific studies have revealed that there are aliens living among us (sus). Experts believe the true proportion of aliens on Earth is around 26%. To test this claim, a representative sample of 100 people is taken; it is found that 24% of these people are aliens.

   (a) What is the population in this problem?

   Solution: The population is the set of all people on Earth.

   (b) Define the parameter of interest, p.

   Solution: The parameter of interest is $p =$ the true proportion of people on Earth that are aliens.

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

   Solution: The random variable of interest is $\widehat{P} =$ the proportion of people in a representative sample of 100 that are aliens.

   For parts (d) - (h): Assume, wherever relevant, that we are conducting a two-sided test at a 5% level of significance.

   (d) What are the null and alternative hypotheses?

   Solution:

$$\begin{bmatrix} H_0: & p = 0.26 \\ H_A: & p \neq 0.26 \end{bmatrix}$$

   (e) Assuming the null is correct, what is the distribution of the test statistic? Be sure to check any/all relevant conditions.

   Solution: We need to check the success-failure conditions.

   1) $np_0 = (100) \cdot (0.26) = 26 \geq 10$ ✓

   2) $n(1 - p_0) = 100 \cdot (1 - 0.26) = 74 \geq 10$ ✓

   Therefore, the Central Limit Theorem for Proportions tells us that, under the null,

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

   (f) Compute the value of the test statistic.

   Solution:

$$\text{TS} = \frac{0.24 - 0.26}{\sqrt{\frac{0.26 \cdot (1-0.26)}{100}}} = -0.46$$

(g) Compute the $p-$value.

Solution: Since we are using a two-sided alternative, our $p-$value can be computed by

$$2 \cdot \mathbb{P}(Z \leq -0.46)$$

where $Z \sim \mathcal{N}(0,1)$. From our normal table, we see that $\mathbb{P}(Z \leq -0.46) = 0.3228$ meaning our $p-$value is 0.6456.

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

Solution: We only reject when the $p-$value is less than the level of significance. Since, in this case, our $p-$value is 0.6456 which is greater than 5%, we fail to reject the null:

> At a 5% level of significance, there was insufficient evidence to reject the null that 26% of the population are aliens in favor of the alternative that the true proportion of aliens is *not* 26%.

4. The time it takes Juan to commute to school from his apartment is normally distributed with a mean of 20 minutes and a standard deviation of 4 minutes.

(a) What is the probability that Juan will commute to school in under 15 minutes on a randomly selected day?

Solution: Let X denote the time (in minutes) it takes Juan to commute to school on a randomly-selected day. Then, from the problem statement, we know that $X \sim \mathcal{N}(20,\ 4)$. We seek $\mathbb{P}(X < 15)$, which can be computed using Standardization and then looking up the relevant probability in the normal table:

$$\mathbb{P}(X < 15) = \mathbb{P}\left(\frac{X-20}{4} \leq \frac{15-20}{4}\right) = \mathbb{P}\left(\frac{X-20}{4} \leq -1.25\right) = 0.1056$$

(b) What is the probability that Juan will commute to school in over 30 minutes on a randomly selected day?

Solution: Letting X be as defined in the previous part, we seek $\mathbb{P}(X > 30)$ which we again compute using Standardization and our normal

table:

$$\begin{aligned}\mathbb{P}(X > 30) &= \mathbb{P}\left(\frac{X-20}{4} > \frac{30-20}{4}\right)\\ &= 1 - \mathbb{P}\left(\frac{X-20}{4} \leq \frac{30-20}{4}\right)\\ &= 1 - \mathbb{P}\left(\frac{X-20}{4} \leq 2.5\right) = 1 - 0.9938 = \boxed{0.0062}\end{aligned}$$

(c) What is the probability that Juan will commute to school in between 10 and 25 minutes on a randomly selected day?

Solution: Letting X be defined as in part (a), we seek $\mathbb{P}(10 \leq X \leq 25)$:

$$\begin{aligned}\mathbb{P}(10 \leq X \leq 25) &= \mathbb{P}(X \leq 25) - \mathbb{P}(X \leq 10)\\ &= \mathbb{P}\left(\frac{X-20}{4} \leq \frac{25-20}{4}\right) - \mathbb{P}\left(\frac{X-20}{4} \leq \frac{10-20}{4}\right)\\ &= \mathbb{P}\left(\frac{X-20}{4} \leq 1.25\right) - \mathbb{P}\left(\frac{X-20}{4} \leq -2.5\right)\\ &= 0.8944 - 0.0062 = \boxed{0.8882}\end{aligned}$$

(d) Let X denote the time in minutes it takes Juan to commute to school on any given day, and let Y denote the time in hours it takes Juan to commute to school on any given day. What is the distribution of Y? be sure to include any/all relevant parameter(s)!

Solution: We know that if X follows a normal distribution, then any linear combination of X also follows a normal distribution. Since $Y = X/60$, we also have

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[(1/60) \cdot X] = (1/60) \cdot \mathbb{E}[X] = (1/60) \cdot (20) = 1/3\\ \text{Var}(Y) &= \text{Var}((1/60) \cdot X) = (1/60)^2 \cdot \text{Var}(X) = (1/60)^2 \cdot (4)^2 = (1/15)^2\end{aligned}$$

Therefore,

$$\boxed{Y \sim \mathcal{N}\left(\frac{1}{3}, \frac{1}{15}\right)}$$

5. In a very old town, it is found that each building has a 15% chance of containing asbestos, independently of all other buildings. Suppose a sample of 15 buildings is taken (with replacement), and the number of buildings in this sample that contain asbestos is recorded.

(a) Define the random variable of interest, and call it X.

Solution: Let X denote the number of buildings, in the sample of 15, that contain asbestos.

(b) What is the distribution of X? Be sure to check any/all relevant conditions.

Solution: We surmise that X is binomially distributed: to verify this, we check the Binomial Criteria:

1) **Independent Trials?** Yes, since sampling is done with replacement.
2) **Fixed number of Trials?** Yes; $n = 15$ trials
3) **Well-defined notion of "success"?** Yes; "success" = "building contains asbestos"
4) **Fixed probability of success?** Yes; $p = 0.15$.

Since all four conditions are met, we conclude that

$$X \sim \text{Bin}(15,\ 0.15)$$

(c) What is the probability that exactly 3 buildings in the sample of 15 contain asbestos?

Solution:

$$\mathbb{P}(X = 3) = \binom{15}{3} \cdot (0.15)^3 \cdot (1 - 0.15)^{15-3} \approx 0.2184$$

(d) What is the standard deviation of the number of buildings (in the sample of 15) that contain asbestos?

Solution:

$$\text{SD}(X) = \sqrt{np(1-p)} = \sqrt{15 \cdot (0.15) \cdot (1 - 0.15)} = 1.9125$$

6. 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 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:** The key point is that Leah has neither administered nor withheld treatment of any kind; participants were simply *observed* based on their current habits (i.e. whether or not they regularly listen to music while studying).

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

   > **Solution:** Once again, the key distinction between an observational study and an experiment is that in an observational study treatment is neither administered nor withheld. If Leah wanted to restructure her study as an experiment, she should seek out 100 volunteers, randomly split these into two groups of size 50 each, assign one group to listen to music and the other to not listen to music, and then record the average GPA of the two groups at the end of the observation period.

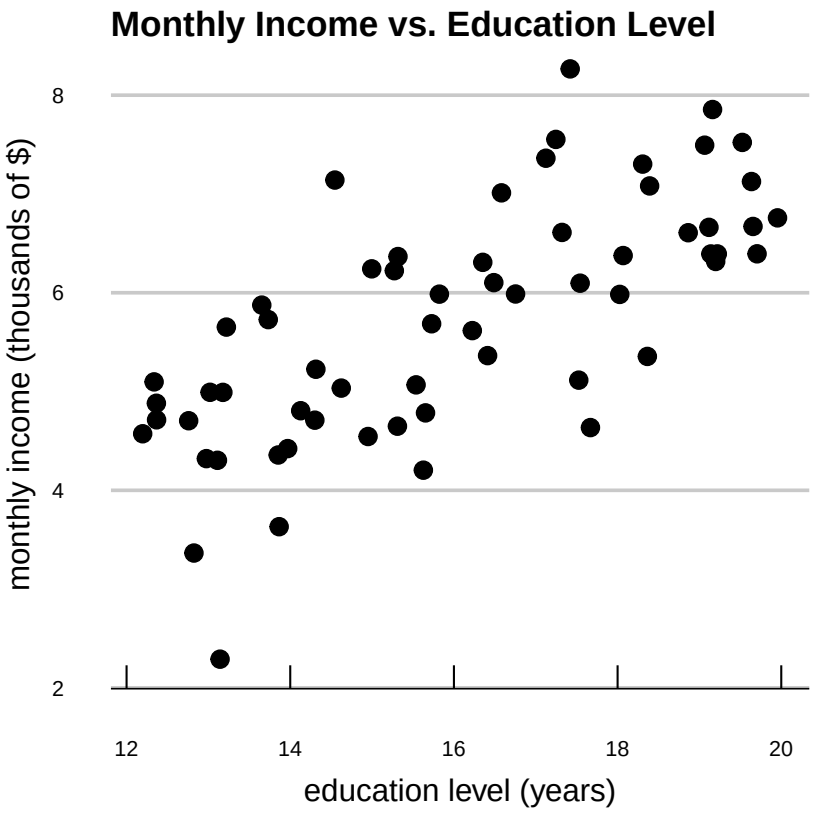
   (c) Is this a longitudinal or cross-sectional study? Explain your reasoning.

   > **Solution:** This is a cross-sectional study.

   (d) 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.)

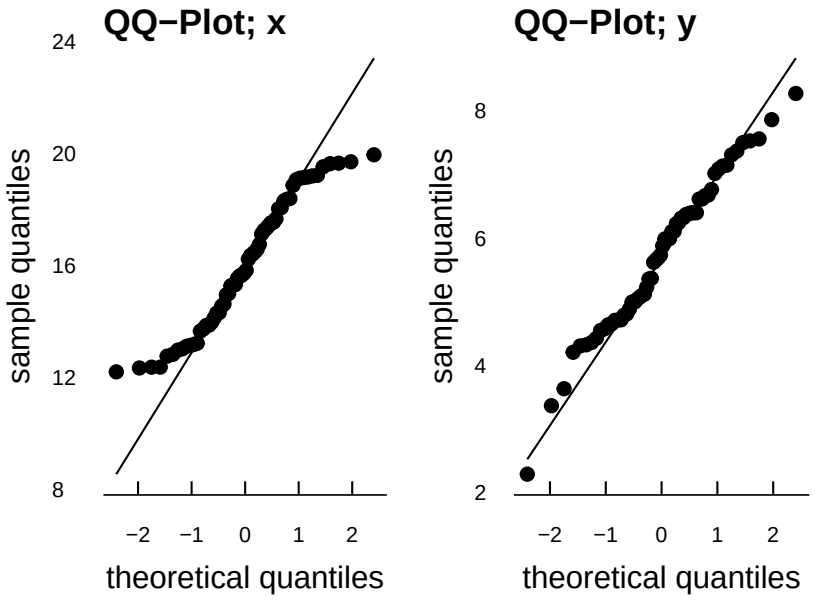
7. Tadhg would like to model the relationship between income and education level (as measured using years of education). He collects a sample of 62 people and records their education level (i.e. years of education) and average monthly income, and produces the following scatterplot from his data:



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

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

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



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

Solution:

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

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

Solution:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{122.4954}{343.1438} \approx 0.35697 \\
\widehat{\beta_0} &= \overline{y} - \widehat{\beta}(\overline{x}) = \frac{354.8923}{62} - (0.35697)\cdot\frac{992.7295}{62} \approx \boxed{0.00835}
\end{aligned}$$

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

Solution: As was computed in the previous part, $\widehat{\beta_1} \approx 0.35697$ (on the actual exam I would likely ask this part before the previous part).

(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 $\text{Var}(\widehat{\beta_1}) = 0.002914$. 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

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

Hence, our confidence interval take the form

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

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 df $= 60$ and the one-tailed 0.025 column) we see that this value is 2.00, meaning our confidence interval is

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

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?

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.