STAT 350 Final Exam FALL 2025

# V1

| Name: _____ | PUID _____ |
|---|---|

**Instructor (circle one):** **Heekyung Ahn** **Yu Lin** **Evidence Matangi** **Timothy Reese** **Halin Shin**

**Select Class Meeting Days/Time**

- ○ **T/Th** 9:00AM-10:15AM
- ○ **MW** 1:30PM-2:45PM
- ○ **MWF** 11:30AM-12:20PM
- ○ **MWF** 12:30 PM-1:20PM
- ○ **MWF** 1:30 PM-2:20PM
- ○ **MWF** 2:30 PM-3:20PM
- ○ **MWF** 3:30-4:20PM
- ○ **MWF** 4:30PM-5:20PM
- ○ **Online**

As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue.

**Instructions:**

1. Please write your **name** and **PUID clearly** on every **odd page.**
2. **Write your work in the box. Do not run over into the next question space.**
3. The only materials that you are allowed during the exam are your **scientific calculator**, **writing utensils, erasers, your crib sheet**, and **your picture ID**. Colored scratch paper will be provided if you need more room for your answers. Please write your name at the top of that paper also.
4. The crib sheet can be a handwritten or typed double-sided 8.5in x 11in sheet.
5. If you share your calculator without permission or have a cell phone at your desk, you will get a **zero** on the exam. Do not take out your cell phone until you are next in line to submit your exam.
6. The exam is only 60 minutes long so there will be no breaks during the exam. If you leave the exam room, you must turn in your exam, and you will not be allowed to come back.
7. **For free response questions you must show ALL your work to obtain full credit.** An answer without showing any work may result in **zero** credit. If your work is not readable, it will be marked wrong. Remember that work must be shown for all numbers that are not provided in the problem or no credit will be given for them. All explanations must be in complete English sentences to receive full credit.
8. All numeric answers should have **four decimal places** unless stated otherwise.
9. After you complete the exam, please turn in your exam as well as your table and any scrap paper that you used. Please be prepared to **show your Purdue picture ID**.
10. You are expected to uphold the honor code of Purdue University. It is your responsibility to keep your work covered at all times. Anyone caught cheating on the exam will automatically fail the course and will be reported to the Office of the Dean of Students.
11. It is strictly prohibited to smuggle this exam outside. Your exam will be returned to you on Gradescope after it is graded.

**Your exam is not valid without your signature below. This means that it won't be graded.**

I attest here that I have read and followed the instructions above honestly while taking this exam and that the work submitted is my own, produced without assistance from books, other people (including other students in this class), notes other than my own crib sheet(s), or other aids. In addition, I agree that if I tell any other student in this class anything about the exam BEFORE they take it, I (and the student that I communicate the information to) will fail the course and be reported to the Office of the Dean of Students for Academic Dishonesty.

**Signature of Student:** _____

**You may use this page as scratch paper.**
**The following is for your benefit only.**

| Question Number | Total Possible | Your points |
|---|---|---|
| Problem 1 (True/False)<br>(2 points each) | 20 | |
| Problem 2 (Multiple Choice)<br>(3 points each) | 18 | |
| Problem 3 | 18 | |
| Problem 4 | 31 | |
| Problem 5 | 40 | |
| Problem 6 | 38 | |
| Total | 150+15 = 165 | |

**The rest of this page can be used for scratch work**

1. **(20 points, 2 points each) True/False Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect.

**1.1.** Let $\boldsymbol{X}$ be a Poisson random variable with $\boldsymbol{E[X] = \mu}$ (average per hour rate). Suppose we define two Poisson random variables, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, defined on two three-hour periods, sharing the rate of $\boldsymbol{X}$**.**

Ⓣ or Ⓕ $\boldsymbol{Y}$ and $\boldsymbol{Z}$ must be independent and identically distributed, $\boldsymbol{Poisson(3\mu)}$.

**1.2.** Let $\boldsymbol{A_1, A_2, \dots, A_n}$, and $\boldsymbol{B}$ be events from a sample space $\boldsymbol{\Omega}$ where $\boldsymbol{A_1, A_2, \dots, A_n}$ form a partition of $\boldsymbol{\Omega}$ and $\boldsymbol{P(B) > 0}$.

Ⓣ or Ⓕ Then it must follow that $\boldsymbol{\sum_{i=1}^{n} P(A_i|B) = 1}$.

**1.3.** A special deck of cards contains eight cards: for each number **1, 2, 3,** and **4** there is exactly one **red card** and one **black card** (so the cards are **1R, 1B, 2R, 2B, 3R, 3B, 4R, 4B**).

Two cards are drawn at random **without replacement**, in order.

Let $\boldsymbol{C}$ denote the event that the first card drawn is **red**, and let $\boldsymbol{D}$ denote the event that the second card drawn is either a **1** or a **2**.

Ⓣ or Ⓕ Events $\boldsymbol{C}$ and $\boldsymbol{D}$ are independent.

**1.4.** On each day, a factory may be in **high-stress state** with probability $\boldsymbol{1/4}$ and in a normal operating mode with probability $\boldsymbol{3/4}$. Let $\boldsymbol{Y}$ be a **Bernoulli** random variable that equals **1** if the day is **high-stress** and **0** otherwise, so $\boldsymbol{P(Y = 1) = 1/4}$. Let $\boldsymbol{X}$ denote the random variable representing the number of machine breakdowns that day. Conditionally on the **state** of the factory, $\boldsymbol{X \mid Y = 1}$ has a **Poisson** distribution with mean **10**, and $\boldsymbol{X \mid Y = 0}$ has a Poisson distribution with mean **6**. Define a new random variable $\boldsymbol{V = X \cdot (1 - Y)}$, so that $\boldsymbol{V}$ quantifies the number of breakdowns under normal operating conditions.

Ⓣ or Ⓕ Since $\boldsymbol{V = X \cdot (1 - Y)}$, it follows that $\boldsymbol{E[V] = E[X] \cdot (1 - E[Y]) = 21/4}$ .

**1.5.** On a given day, a warehouse receives online orders in two waves. Let $\boldsymbol{X}$ be the total number of units ordered in the **morning** and $\boldsymbol{Y}$ the total number of units ordered in the **afternoon**. Historical data suggest that the morning and afternoon totals are normally distributed and independent, with

$$\boldsymbol{X \sim N(\mu_X = 120, \sigma_X = 15)} \qquad \boldsymbol{Y \sim N(\mu_Y = 160, \sigma_Y = 20)}$$

The operations manager defines a planning index:

$$\boldsymbol{I = 0.3X + 0.7Y + 50}$$

Ⓣ or Ⓕ The variance of $\boldsymbol{I}$ satisfies $\boldsymbol{V(I) = 0.3 \cdot Var(X) + 0.7 \cdot Var(Y) = 347.5}$.

**1.6.** In a one-way ANOVA, a Tukey's HSD output reports results for 105 unique pairwise comparisons among the group means.

Ⓣ or Ⓕ The single factor variable in the ANOVA model must have exactly 15 levels.

**1.7.** A one-way ANOVA is conducted to compare several treatment means. All model assumptions are checked and found to be reasonable. The resulting F test statistic is approximately 1, and the corresponding p-value is large, so the null hypothesis of equal treatment means is not rejected.

Ⓣ or Ⓕ In this situation, the appropriate next step is to carry out a Tukey multiple comparison procedure to determine which specific treatment means differ.

**1.8.** In a simple linear regression, $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \epsilon_i}$, where the $\boldsymbol{\epsilon_i's}$ are **i.i.d.** normal with variance $\boldsymbol{\sigma^2}$.

Ⓣ or Ⓕ Under the null hypothesis $\boldsymbol{H_0: \beta_1 = 0}$, if the value of $\boldsymbol{\sigma^2}$ is known then the test statistic $\left(\widehat{\boldsymbol{\beta}}_1 \div \sqrt{\frac{\sigma^2}{S_{XX}}}\right)$ follows a standard normal distribution.

**1.9.** In a simple linear regression setting, a residual plot is constructed by plotting the residuals versus the fitted values (or versus $\boldsymbol{x}$). The points appear randomly scattered around the horizontal line at 0 with roughly the same vertical spread for all values of $\boldsymbol{x}$, and no clear curvature or funnel shape is visible.

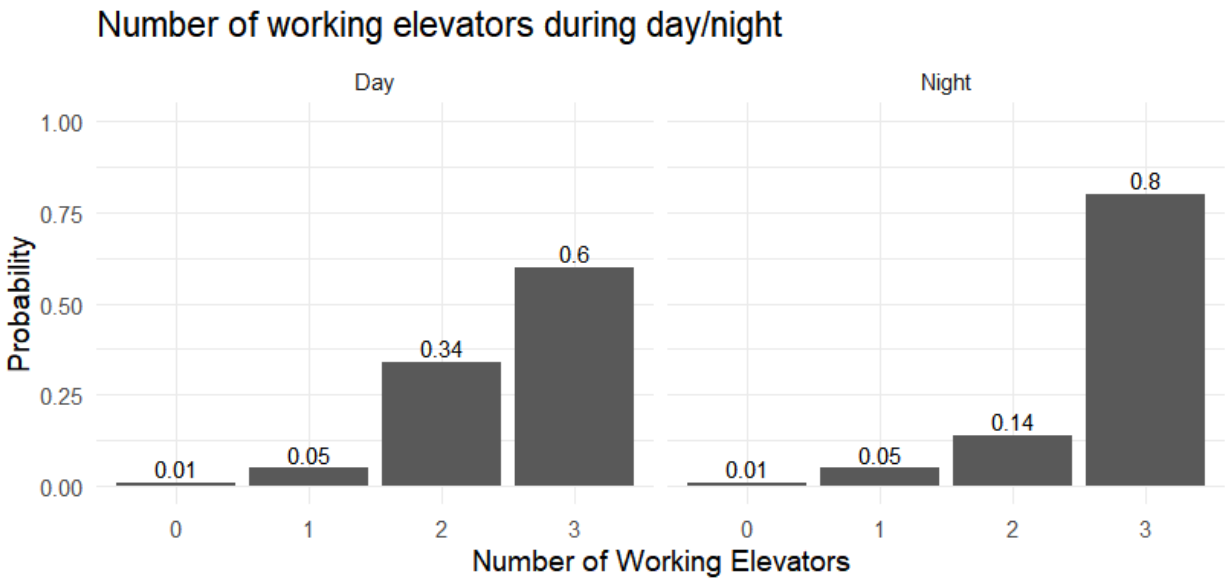
Ⓣ or Ⓕ This residual pattern provides reasonable support for both the **linearity assumption** and the **constant variance** (**homoscedasticity**) assumption.

**1.10.** In a simple linear regression analysis, the sample coefficient of determination $\boldsymbol{R^2}$ is found to be very close to 1.

Ⓣ or Ⓕ From this, we can conclude that a large proportion of the variability in the response variable is explained by its linear relationship with the explanatory variable, and that the relationship between them is in fact linear.

**2.** **(18 points, 3 pts each) Multiple Choice Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect. **For each question, there is only one correct option letter choice unless specified.**

**2.1.** Three elevators in the Math building are known to be old and often malfunction during the daytime hours (8 am – 7 pm). However, people report that the elevators tend to work normally during the nighttime hours (7 pm – 8 am). Let $\boldsymbol{X}$ be a nominal variable indicating the time of day, with sample space $\{\boldsymbol{day}, \boldsymbol{night}\}$, and $\boldsymbol{Y}$ is the number of elevators working properly. Each of the bar graphs below shows the distribution of the number of working elevators during daytime and nighttime hours.




Which of the following probabilities is computed correctly?

Ⓐ $\boldsymbol{P(Y \leq 1) = 0.12}$

Ⓑ $\boldsymbol{P(\{X = day\} \cap \{Y = 3\}) = 0.6}$

Ⓒ $\boldsymbol{P(X = day) = P(X = night) = 1}$

Ⓓ $\boldsymbol{P(Y > 1 | X = night) = 0.94}$

Ⓔ All of the above

**2.2.** Let $\boldsymbol{X}$ be a **discrete random variable** with **support** $\{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$. Find the constant $\boldsymbol{k}$ that makes the **function** below a valid **probability mass function (pmf)**.

$$f_X(x) = \begin{cases} 0.45, & x = 0 \\ \frac{k}{x}, & x = 1, 2, 3 \\ 0, & otherwise \end{cases}$$

Ⓐ $\boldsymbol{k = 0.45}$

Ⓑ $\boldsymbol{k = 0.3}$

Ⓒ $\boldsymbol{k = 0.5455}$

Ⓓ $\boldsymbol{k = 0.9124}$

Ⓔ $\boldsymbol{k = 0.5006}$

Ⓕ $\boldsymbol{k = \infty}$

**2.3.** A simple linear regression is fit to relate product price ($\boldsymbol{x}$) to weekly sales ($\boldsymbol{Y}$). All standard simple linear regression assumptions (linearity, constant variance, normality of errors, and independence) are judged to be reasonably satisfied based on diagnostic plots and study design. A **95% confidence interval** for the slope $\boldsymbol{\beta_1}$ is $(\mathbf{-12.4}, \mathbf{-4.1})$. At significance level $\boldsymbol{\alpha = 0.05}$, what is the correct conclusion about the **population** linear association between price and sales?

Ⓐ There is no evidence of a linear association because 0 is not in the interval.

Ⓑ There is evidence of a linear association because 0 is not in the interval.

Ⓒ There is evidence that changing price will *cause* weekly sales to decrease in the population, because 0 is not in the interval.

Ⓓ We can only conclude that there is a linear association in this particular sample; a confidence interval cannot be used to draw conclusions about the population.

Ⓔ We cannot draw any conclusion about linear association from this interval without also knowing the p-value of the slope test.

**2.4.** A researcher fits a simple linear regression model to relate a response variable $\boldsymbol{Y}$ to an explanatory variable $\boldsymbol{x}$. The usual model assumes that the error terms are normally distributed. Which of the following plots is **most appropriate** for checking the normality assumption of the errors?

Ⓐ Scatterplot of $\boldsymbol{y}$ versus $\boldsymbol{x}$.

Ⓑ Plot of residuals versus $\boldsymbol{x}$ values.

Ⓒ Normal probability plot (qq-plot) of the residuals.

Ⓓ Normal probability plot (qq-plot) of the response.

Ⓔ Histogram of the response.

Ⓕ Histogram of the explanatory variable.

**2.5.** The ANOVA **F-statistic** always falls within which range

Ⓐ The positive real numbers $(\mathbf{0}, \infty)$.

Ⓑ The real numbers $(-\infty, \infty)$.

Ⓒ The negative real numbers $(-\infty, \mathbf{0})$.

Ⓓ The real numbers between $\mathbf{0}$ and $\mathbf{1}$.

Ⓔ None of the above.

**2.6.** A researcher compares four insomnia drugs by randomly assigning seniors to one of the four treatments and analyzing the outcomes with a one-way ANOVA. In this fixed study, which statements correctly describe how **between drug variation**, **within drug variation**, and the **F** test statistic are related?

Ⓐ The ANOVA **F** statistic is the ratio of between group variation to within group variation.

Ⓑ A **large F** occurs when the differences among the four drug sample means are large relative to the typical person to person variability within each drug group.

Ⓒ The **within drug variation** sets the noise level for judging whether the observed differences among drug sample means look unusually large.

Ⓓ An **F** statistic near 1 indicates that between drug variation is about the same size as within drug variation, so the drug means do not stand out beyond the background variability.

Ⓔ All of the above.

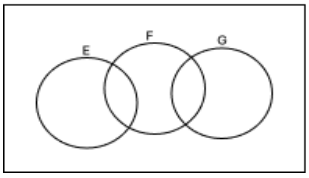
Ⓕ None of the above.

**Free Response Questions.** Show all work, label your answers, use **four decimal places**.

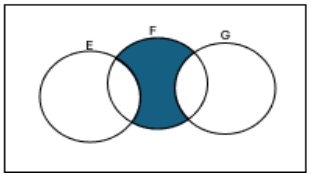
3. **(18 points)** Events E, F, and G belong to the same sample space, S, each with a non-zero probability.

   **a) (8 points)** Match each Venn diagram with the probability statement that correctly represents the colored region.

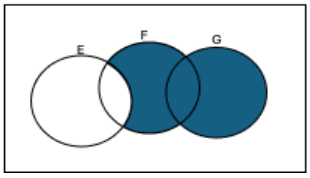
A



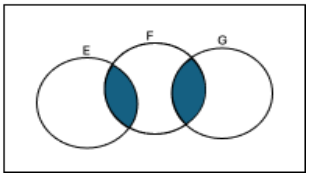

B




C

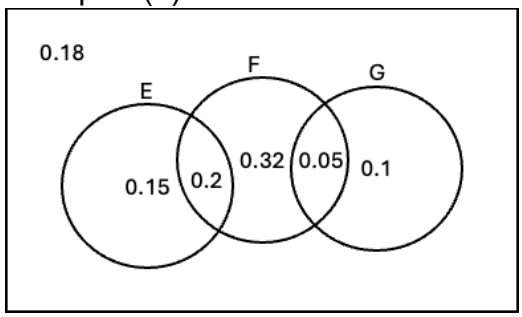



D




| | Notation | Venn diagram (Letter) |
|---|---|---|
| **i** | $\boldsymbol{P(E' \cap F \cap G')}$ | |
| **ii** | $\boldsymbol{P(E \cap G)}$ | |
| **iii** | $\boldsymbol{P\left(F \cap \left((E \cap F) \cup G\right)\right)}$ | |
| **iv** | $\boldsymbol{P(E' \cap (F \cup G))}$ | |

**b) (10 points)** The probability of each region in the Venn diagram is given below. Further, define events $\boldsymbol{A, B, C}$, and $\boldsymbol{D}$ according to the colored regions of the Venn diagrams provided in part (a).




i. **(4 points)** Compute the probability of the intersection of $\boldsymbol{A, B, C}$, and $\boldsymbol{D}$.

ii. **(6 points)** Compute the probability of the union of $\boldsymbol{A, B, C}$, and $\boldsymbol{D}$.

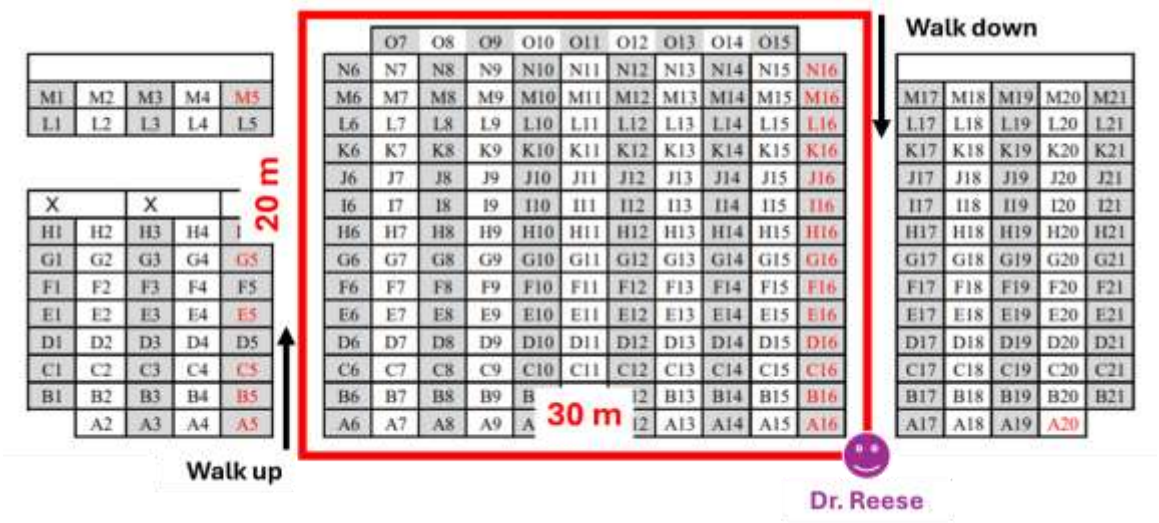
4. **(31 points)** Dr. Reese asked two TAs, Zhenghao and Haoyu, to walk around the exam room following a clockwise direction along the rectangular path indicated in the diagram below. The rectangle has a length of 30 meters and a width of 20 meters, so the total total distance is 100 meters. Each TA walks exactly one full lap of the rectangle, starting and ending at Dr. Reese.

   Zhenghao walks at a constant pace around the path. Haoyu walks at a baseline pace on flat horizontal segments, twice as fast when walking down the path, and twice as slow when walking up the path. Consider a single traversal (one lap) of the rectangular path, starting at Dr. Reese and returning to the starting point.

   Define two independent random variables by observing each TA at a randomly chosen time during their respective lap, with every instant from departure to return equally likely to be chosen. The random observation times for Zhenghao and Haoyu are assumed independent.

   $\boldsymbol{Z}$: the distance that Zhenghao has traveled from Dr. Reese along the path

   $\boldsymbol{H}$: the distance that Haoyu has traveled from Dr. Reese along the path




   a) **(4 points)** Determine the distribution of $\boldsymbol{Z}$ and its parameter(s)

**b) (4 points)** What is the probability that Zhenghao is walking up the path?

**c) (10 points)** Based on Haoyu's speed, the probability density function of $\boldsymbol{H}$ is given by

$$f_H(x) = \begin{cases} k, & 0 < x \leq 30, \\ 2k, & 30 < x \leq 50, \\ k, & 50 < x \leq 80, \\ k/2, & 80 < x \leq 100, \\ 0, & otherwise. \end{cases}$$

Determine the value of $\boldsymbol{k}$ that makes $f_H(x)$ a valid probability density function.

**d) (6 points)** What is the probability that Haoyu has traveled between 40 m and 70 m from Dr. Reese?

**e) (4 points)** Compare the average travel distance of Zhenghao and Haoyu. Who travels farther on average?

**f) (3 points)** Suppose Zhenghao and Haoyu now walk counterclockwise instead of clockwise around the same path. Which of the following quantities will change?

Ⓐ The expected value of $\boldsymbol{Z}$

Ⓑ The expected value of $\boldsymbol{H}$

Ⓒ The variance of $\boldsymbol{Z}$

Ⓓ The variance of $\boldsymbol{H}$

Ⓔ All of the above

5. **(40 points)** A data analyst wants to compare four statistical techniques in terms of their mean effectiveness at capturing hidden correlations between depression indices and a range of human health indicators relevant to women's health in Indiana. The techniques under consideration are regression, ANOVA, Taguchi methods, and structural equation modeling (SEM). Each technique is applied to the same dataset with $\boldsymbol{m}$ replications per technique, so that the total sample size is $\boldsymbol{n = 4m}$. The test is conducted at the $\boldsymbol{\alpha = 0.05}$ significance level, and the following ANOVA output is obtained.

| Source | Degrees of Freedom | Sum of Squares | Mean Square | F statistic | $Pr(> F)$ |
|---|---|---|---|---|---|
| **Treatment** | | **931.46** | | | **7.465069e-07** |
| **Error** | **84** | | | | |
| **Total** | | **3011.09** | | | |

   a) **(6 points)** Complete the missing information in the ANOVA Table above. Clearly, show your work in the box provided below.

**b) (6 points)** Determine the value of **m**.

**c) (2 points)** Given that the following summary statistics were realized from the data. Check whether the constant variance assumption was met or not.

| Technique | Regression | ANOVA | Taguchi | SEM |
|---|---|---|---|---|
| $\overline{x}_{i.}$ | 54.4 | 45.1 | 83.4 | 74.3 |
| $s_i$ | 5.3 | 6.3 | 3.8 | 4.1 |

**d) (5 points)** Using the ANOVA output and assuming all four techniques share the same error variance, compute the **pooled estimate of the error standard deviation**. Show your work clearly.

**e) (5 points)** Now ignore which technique each observation came from and treat all replicate effectiveness measurements from the four techniques as **one combined sample**. Compute the **standard deviation of this combined sample.** Show your calculation clearly. (Note this is not the same quantity as in part d))

**f)** **(3 points)** Which statement best describes the difference between the standard deviations computed in parts (d) and (e)?

Ⓐ Both standard deviations measure exactly the same quantity, just computed using different formulas.

Ⓑ The standard deviation in part (d) measures only variability between technique means, while the standard deviation in part (e) measures only variability within each technique.

Ⓒ The standard deviation in part (d) measures how much all observations vary around a single overall mean, while the standard deviation in part (e) measures how much observations vary around their own technique's mean.

Ⓓ The standard deviation in part (d) measures how much observations vary around their own technique's mean (assuming a common error variance), while the standard deviation in part (e) measures how much all observations vary around a single overall mean when technique labels are ignored.

**g)** **(2 points)** Provide the first two steps of the four-step hypothesis testing procedure.

**h)** **(5 points)** Based on the ANOVA results, state your formal decision for the hypothesis test and write a conclusion in the context of this study.

**i) (6 points)** The following multiple comparison results were obtained using the Tukey HSD method. Create a graphical display to communicate the pattern of differences using the underline notation and briefly comment on which statistical technique appears to have the highest mean effectiveness at capturing hidden correlations.

| i-j | Significant? |
|---|---|
| Regression - ANOVA | No |
| Regression - SEM | No |
| Regression - Taguchi | Yes |
| ANOVA- SEM | Yes |
| ANOVA- Taguchi | Yes |
| SEM-Taguchi | Yes |

**6. (38 points)** A marketing analyst is studying the relationship between a store's weekly online advertising spending and its weekly sales.

- Let $\boldsymbol{x}$ denote the weekly online advertising spending (in hundreds of dollars)
- Let $\boldsymbol{Y}$ denote the weekly sales (in thousands of dollars)

Data were collected for 8 weeks and are shown below.

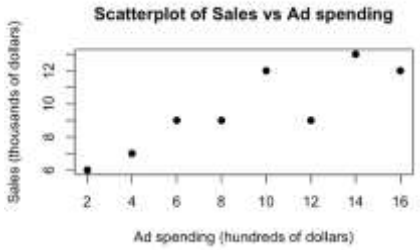
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ad spending (hundreds of dollars) | **2** | **4** | **6** | **8** | **10** | **12** | **14** | **16** |
| Sales (thousands of dollars) | **6** | **7** | **9** | **9** | **12** | **9** | **13** | **12** |

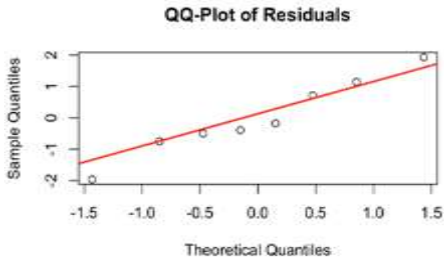
From these data, the following summary statistics have been computed:

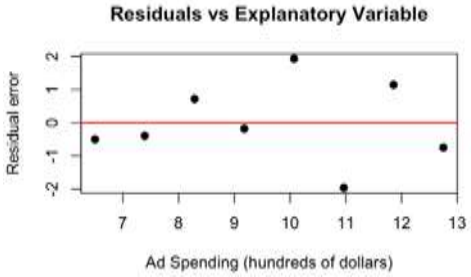
$$\sum \boldsymbol{x_i = 72} \qquad \sum \boldsymbol{y_i^2 = 785}$$

$$\sum \boldsymbol{x_i^2 = 816} \qquad \sum \boldsymbol{x_i y_i = 768}$$

$$\sum \boldsymbol{y_i = 77}$$




**a) (10 points)** Compute the least squares regression line for predicting weekly sales from weekly ad spending, clearly reporting the estimated slope $\boldsymbol{b_1}$ and intercept $\boldsymbol{b_0}$, and then writing the fitted regression equation.

**b) (7 points)** A simple linear regression of $\boldsymbol{Y}$ on $\boldsymbol{x}$ was fit, and the following ANOVA table was obtained. Complete all missing entries. You do not need to show your work.

| Source | Degrees of Freedom | Sum of Squares | Mean Square | F statistic | $Pr(> F)$ |
|---|---|---|---|---|---|
| **Model** | | **33.48** | | | **0.0046** |
| **Error** | | **10.39** | | | |
| **Total** | | | | | |

**c) (4 points)** Compute the coefficient of determination $\boldsymbol{R^2}$ and interpret it in the context of this study.

**d) (4 points)** Use your results from parts **a)**, **b),** and **c)** to compute the Pearson correlation coefficient $\boldsymbol{r}$. Interpret $\boldsymbol{r}$ in the context of this study.

**e) (4 points)** Compute $\boldsymbol{s}$, the **estimate** of $\boldsymbol{\sigma}$ (the standard deviation of the error terms**)**, for this study.

**f) (4 points)** To test $\boldsymbol{H_0: \beta_1 = 0}$ versus $\boldsymbol{H_A: \beta_1 \neq 0}$ in this simple linear regression, what is the value of the **t test statistic**?

**g) (5 points)** Is there a **significant linear association** between online advertising spending and sales at a $\boldsymbol{\alpha = 0.01}$ level of significance? State the hypotheses and provide a formal conclusion in context.