STAT 350 Final Exam SPRING 2026

V1

Name: _____ PUID _____

Instructor (circle one):

Heekyung Ahn **Evidence Matangi** **Halin Shin**
Yu Lin **Timothy Reese** **Chenzhong Wu**

Select Class Meeting Days/Time

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

Exam Instructions

1. **Identification:** Write your last name clearly on every odd page and on any provided scratch paper. Have your Purdue picture ID ready when submitting the exam.
2. **Allowed Materials:** Scientific calculator, writing utensils, erasers, and two double-sided 8.5" x 11" crib sheets (handwritten or typed).
3. **Formatting & Work:** Keep all work strictly within the designated boxes. Round all numeric answers to **four decimal places** unless stated otherwise.
4. **Grading Requirements:** You must show ALL work for free-response questions. Unsupported numbers, unreadable work, or missing explanations (which must be in complete English sentences) will receive zero credit.
5. **Logistics:** The exam is 120 minutes. Leaving the room finalizes your submission. After you complete the exam, please turn in your exam as well as any scrap paper that you used. Please be prepared to **show your Purdue picture ID**. Exams will be returned via Gradescope.

Strict Prohibitions & Honor Code

- **Zero Tolerance:** Absolutely no cell phones at your desk. No sharing calculators. Violating either rule results in an immediate zero on the exam.
- **Academic Integrity:** You are responsible for keeping your work covered at all times. The unauthorized removal of exam materials from the room, any form of academic dishonesty, or discussing exam contents with students who have not yet tested will result in an **automatic course failure** and a formal report to the Office of the Dean of Students.

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

Your exam is not valid without your signature below. This means that it won't be graded.
I attest that I have read and followed the instructions above honestly. The work submitted is my own, produced without unauthorized assistance. I agree that if I share information about this exam with any student before they take it, both parties will fail the course and be reported 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) (2 points each)	20	
Problem 2 (Multiple Choice) (3 points each)	18	
Problem 3	24	
Problem 4	26	
Problem 5	40	
Problem 6	37	
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. A sensor records temperatures in Celsius. A data analyst converts every observation to Fahrenheit using F = 1.8C + 32.

Ⓣ or Ⓕ The sample standard deviation of the Fahrenheit data is exactly 1.8 times the sample standard deviation of the Celsius data.

1.2. A mechanical engineer classifies each manufactured part as exactly one of three grades: $\boldsymbol{A}$ (premium), $\boldsymbol{B}$ (standard), or $\boldsymbol{C}$ (substandard). The historical probability for the classification to each grade are $\boldsymbol{P(A) = 0.3}$, $\boldsymbol{P(B) = 0.6}$, and $\boldsymbol{P(C) = 0.1}$.

Ⓣ or Ⓕ Events $\boldsymbol{A}$ and $\boldsymbol{C}$ are **dependent**.

1.3. During an NFL season, a sports analytics team tracks all reported injuries sustained by a single team per game quarter, including those not immediately apparent to viewers such as minor strains and aggravations recorded on the official injury report. It has been historically observed that the team averages approximately 1.4 reported injuries per quarter across all games. However, detailed records reveal that injury rates in the 4th quarter are consistently higher than in the 1st quarter, as cumulative fatigue increases injury risk throughout a game.

Ⓣ or Ⓕ A single $\boldsymbol{\text{Poisson}(\lambda = 1.4)}$ model applied uniformly across all four quarters would **violate** an assumption of the Poisson process.

1.4. Two CNC machines produce bolts whose diameters (in mm) follow continuous uniform distributions. Machine A produces bolts with diameters following $\boldsymbol{\text{Uniform}(9.5, 10.5)}$ and Machine B produces bolts with diameters following $\boldsymbol{\text{Uniform}(9.8, 10.8)}$.

Ⓣ or Ⓕ The probability that Machine A produces a bolt with a diameter between 9.8 and 10.0 is the same as the probability that Machine B does.

1.5. A quality engineer models the lifespan of a sensor component using a continuous distribution. She computes $\boldsymbol{f_X(500) = 0.003}$, where $\boldsymbol{f_X}$ is the PDF of the lifespan in hours.

Ⓣ or Ⓕ Therefore the engineer is certain that a lifespan of 500 hours is very rare.

1.6. A **one-way ANOVA** with **5 groups** produces a **significant $\boldsymbol{F}$-test**.

Ⓣ or Ⓕ If the researcher wants to compare all 10 possible pairs of means, Dunnett's method is more appropriate than Tukey's method.

1.7. In a **one-way ANOVA**, if the **between-group variability** is large relative to the **within-group variability**,

Ⓣ or Ⓕ then the **$\boldsymbol{F}$-test** statistic will tend to be large, giving more evidence against the null hypothesis that all population means are equal.

1.8. A biostatistician plans to fit a simple linear regression line to predict male adults' height using their Femur bone length. Both variables are measured in inches in the original data. Before fitting a regression line, the unit has changed to millimeters for universal applications in medical fields.

Ⓣ or Ⓕ The p**-value** of a regression slope remains constant even after the unit change.

1.9. Suppose that $(-3, 4)$ is a **95% confidence interval** for β_0 in a simple linear regression. For some constant c, we perform hypothesis testing H_a: $\beta_0 \neq c$ at $\alpha = 0.05$.

Ⓣ or Ⓕ We reject the null hypothesis if c is within the confidence interval.

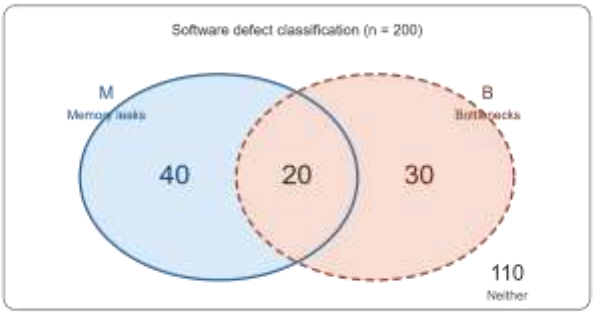
1.10. A materials engineer fits a simple linear regression to predict tensile strength (Y) from carbon content (X) in steel alloys. Before checking residuals, the engineer examines the distribution of Y values alone and finds them to be strongly right-skewed.

Ⓣ or Ⓕ The normality assumption of the simple linear regression model is violated.

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. A software quality team reviews 200 applications for two categories of defects before deployment. Let $M = \{\textbf{has memory leak issues}\}$ and $B = \{\textbf{has performance bottleneck issues}\}$.

The Venn diagram shows the number of applications in each region.



Which of the following counts is computed correctly? (|·| denotes size of set)

Ⓐ $|M' \cap B'| = 110$

Ⓑ $|M' \cap B| = 180$

Ⓒ $|M \cap B'| = 30$

Ⓓ $|M \cup B'| = 150$

2.2. A semiconductor fabrication line experiences random equipment faults. The time (in hours) between faults follows an Exponential distribution with rate λ = 0.5 per hour. The line has been running fault-free for at least 6 hours.

What is the probability it continues running fault-free for at least two more hours?

Ⓐ 0.0183

Ⓑ 0.0498

Ⓒ 0.1353

Ⓓ 0.3679

Ⓔ 0.6321

2.3. The response time (in milliseconds) of a web application follows a Normal distribution with $\boldsymbol{\mu = 250}$ and $\boldsymbol{\sigma = 20}$. A request is classified as "slow" if it takes more than 230 ms.

Given that a request is slow, what is the probability it takes more than 280 ms? Fractions are shown for readability. In R, these would be written using the "/" operator.

Ⓐ `pnorm(280, mean = 250, sd = 20, lower.tail = FALSE)`

Ⓑ $\dfrac{\texttt{pnorm(280, mean = 250, sd = 20, lower.tail = FALSE)}}{\texttt{pnorm(230, mean = 250,sd = 20, lower.tail =TRUE)}}$

Ⓒ $\dfrac{\texttt{pnorm(280, mean = 250, sd = 20, lower.tail =TRUE)} - \texttt{pnorm(230, mean = 250, sd = 20, lower.tail =TRUE)}}{\texttt{pnorm(230, mean = 250, sd = 20,lower.tail = FALSE)}}$

Ⓓ $\dfrac{\texttt{pnorm(280, mean = 250, sd = 20, lower.tail = FALSE)}}{\texttt{pnorm(230, mean = 250, sd = 20, lower.tail = FALSE)}}$

2.4. A one-way ANOVA with **4 groups** is significant, and the researcher wants to compare **all possible pairs** of means using **Bonferroni**. What is the Bonferroni-adjusted significance level for each comparison if the family-wise error rate is set to 0.05?

Ⓐ 0.0500

Ⓑ 0.0250

Ⓒ 0.0125

Ⓓ 0.0083

Ⓔ 0.0050

2.5. Which of the following is **not required** for a traditional one-way ANOVA?

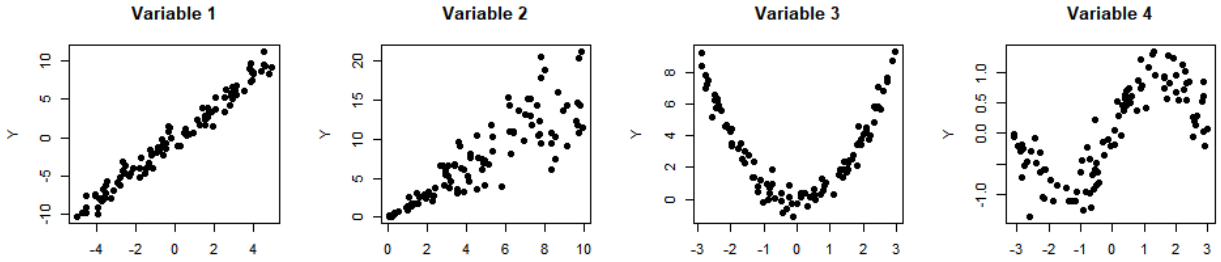
Ⓐ Independent random samples from each population

Ⓑ Equal population variances across groups

Ⓒ Each of the k populations is normally distributed, or sample means are approximately normally distributed

Ⓓ Equal sample sizes in all groups

2.6. A dataset consists of one response variable and four explanatory variables (Variables 1-4). For each explanatory variable, a scatter plot is drawn against the response variable. Select the two explanatory variables whose sample correlation coefficient r with the response variable is closest to zero.



Ⓐ Variables 1 & 2

Ⓑ Variables 2 & 3

Ⓒ Variables 2 & 4

Ⓓ Variables 3 & 4

Ⓔ None — all four variables have strong sample correlations with the response variable.

Free Response Questions 3-5. Show all work, clearly label your answers, and use **four decimal places**.

3. **(24 points)** A utility company, Earl Energy, is known for long customer service wait times. Let X denote the waiting time (in hours) until a customer is connected to the next available representative. The probability density function (pdf) and cumulative distribution function (cdf) of X are given below.

$$f_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}x^2 e^{-x}, & x \geq 0 \end{cases}$$

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - \frac{1}{2}e^{-x}(x^2 + 2x + 2), & x \geq 0 \end{cases}$$

a) (10 points) What is the probability that a customer waits for more than 30 minutes?

$$P\left(X > \frac{1}{2}\right) = 1 - F_X\left(\frac{1}{2}\right)$$

$$= 1 - \left(1 - \frac{1}{2}e^{-\frac{1}{2}} \cdot \left(\left(\frac{1}{2}\right)^2 + 2 \cdot \left(\frac{1}{2}\right) + 2\right)\right)$$

$$= e^{-\frac{1}{2}} \cdot \left(\frac{13}{8}\right) \approx 0.9856$$

b) (14 points) Find the variance of a rate $\frac{1}{X}$ given that $E\left[\frac{1}{X}\right] = 0.5$.

$$\text{Var}\left(\frac{1}{X}\right) = E\left[\frac{1}{X^2}\right] - \left(E\left[\frac{1}{X}\right]\right)^2$$

$$E\left[\frac{1}{X^2}\right] = \int_0^\infty \frac{1}{x^2} \cdot \frac{1}{2}x^2 e^{-x}\, dx$$

$$= \frac{1}{2} \cdot \int_0^\infty e^{-x}\, dx = \frac{1}{2}$$

$$\text{Var}\left(\frac{1}{X}\right) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

4. **(26 points)** A software team uses Claude Code to assist with code commits. Each commit is independently classified as either routine or novel. A **routine commit** is routed to **Configuration A**, and a **novel commit** is routed to **Configuration B**. Each commit **independently** has a **20% probability** of **being novel (Configuration B)** and an **80% probability** of being **routine (Configuration A)**.
During a particular week, the team makes **25** commits. Each commit is automatically and **independently** tested for **bugs**. The probability that a **Configuration A commit** contains a **bug** is **0.05**, and the probability that a **Configuration B commit** contains a **bug** is **0.30**.

   a) **(10 points)** A **single commit** is selected at random from the week's 25 commits. What is the probability that it contains a bug?

$$P(\text{Bug}) = P(\text{Bug}|\text{A}) \cdot P(A) + P(\text{Bug}|\text{B}) \cdot P(B)$$
$$= 0.05 \cdot 0.8 + 0.3 \cdot 0.2 = 0.1$$

   b) **(10 points)** A **single commit** from the week is found to contain a bug. What is the probability it was handled by **Configuration B**?

$$P(\text{B}|\text{Bug}) = \frac{P(\text{Bug}|\text{B}) \cdot P(B)}{P(\text{Bug})}$$
$$= \frac{P(\text{Bug}|\text{B}) \cdot P(B)}{P(\text{Bug})}$$
$$= \frac{0.3 \cdot 0.2}{0.1} = \frac{0.06}{0.1} = 0.6$$

   c) **(6 points)** Find the expected number and standard deviation of bugs found in **Configuration B** during the week.

Let X_B denote the number of bugs found in **Configuration B** during the week

$$X_B \sim \text{Binomial}(\text{n} = 25, \text{p} = 0.06)$$
$$E[X_B] = 25 \cdot 0.06 = 1.5$$
$$\text{Var}(X_B) = 25 \cdot 0.06 \cdot (1 - 0.06) = 1.41$$
$$\text{SD}_{X_B} = \sqrt{1.41} \approx 1.1874$$

5. **(40 points)** A agronomist wants to compare the average plant height increase (in cm) produced by **four fertilizers**. A random sample of plants was assigned to each fertilizer treatment. The summary information is given below.

Group	n_i	$\bar{x}_i$	s_i
Fertilizer 1	10	17.80	2.05
Fertilizer 2	10	20.70	2.31
Fertilizer 3	10	19.25	2.18
Fertilizer 4	10	24.00	2.42

a) (2 points) Using the summary statistics, assess whether the equal variance (homogeneity of variance) assumption appears reasonable. Show your work and state your conclusion clearly. For the rest of the problem, assume all other ANOVA assumptions are satisfied.

$$\frac{s_{\max}}{s_{\min}} = \frac{2.42}{2.05} \approx 1.1805 < 2$$

By the rule of thumb the equal variance assumption is valid.

b) (14 points) Complete the ANOVA table below. Show your work in the box below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F statistic	Pr(>F)
Factor	3	211.2687	70.423	13.9815	3.351353e-06
Error	36	181.3266	5.0369		
Total	39	392.5953			

c) (4 points) Provide the first two steps of the four-step one-way ANOVA hypothesis testing procedure.

Step 1 Identify and describe the parameter(s):

Let $\mu_1, \mu_2, \mu_3, \mu_4$ denote the true mean height increase in cm treated with fertilizers 1, 2, 3, and 4 respectively.

Step 2 Define the hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: **At least one mean is different**

d) (3 points) Which of the following R code statements returns the correct p-value?

Ⓐ
```
pf(F_ts/2, df1=4, df2 = 36, lower.tail = FALSE)
```

Ⓑ
```
pf(F_ts, df1=3, df2 = 36, lower.tail = FALSE)
```

Ⓒ
```
pf(F_ts, df1=4, df2 = 37, lower.tail = TRUE)
```

Ⓓ
```
pf(F_ts, df1=3, df2 = 36, lower.tail = TRUE)
```

Ⓔ
```
2*pf(F_ts, df1=4, df2 = 40, lower.tail = FALSE)
```

e) (8 points) The calculated p-value is 3.35e-06. At a significance level of $\alpha = 0.05$, state your formal decision and conclusion in the context of the problem.

Since the p-value $= 3.35e-06 < 0.05 = \alpha$, therefore we have evidence to reject H_0.

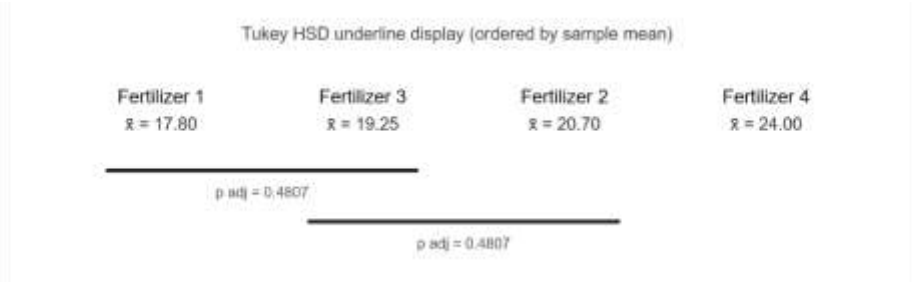
The data **does** give **strong** support (p-value = $3.35e-06$) to the claim that **at least one of the fertilizers** produces a **true mean height increase** (in centimeters) that differs from **at least one other**.

f) **(4 points)** Based on your conclusion in part **e)**, is it appropriate to proceed to pairwise comparisons such as Tukey's HSD? Briefly explain.

Since we rejected the null hypothesis we need to determine which means are different and thus a post-hoc analysis should be conducted.

g) **(5 points)** The following Tukey HSD results were obtained. Construct a graphical display based on these results, and briefly state which fertilizer appears to have the largest population mean plant height increase and provide justification.

Comparison	diff	lwr	upr	p adj
Fertilizer 2 - Fertilizer 1	2.9000	0.1969	5.6031	0.0306
Fertilizer 3 - Fertilizer 1	1.4500	-1.2531	4.1531	0.4807
Fertilizer 4 - Fertilizer 1	6.2000	3.4969	8.9031	0.0000
Fertilizer 3 - Fertilizer 2	-1.4500	-4.1531	1.2531	0.4807
Fertilizer 4 - Fertilizer 2	3.3000	0.5969	6.0031	0.0118
Fertilizer 4 - Fertilizer 3	4.7500	2.0469	7.4531	0.0002



6. **(37 points)** A driving school wants to estimate the monthly car insurance premium for teenage drivers who are at least 16 but under 20 years old (all with minimum-coverage policies). They randomly selected 100 teen drivers and recorded each driver's monthly premium (in dollars) and age (in years) at enrollment. Preliminary analysis indicates a linear relationship between monthly premium **(y)** and age **(x)**. The school plans to fit a simple linear regression model to provide statistical estimates of monthly premiums based on age.

$S_{xx} = 63.797$	$S_{xy} = -538.3375$	$S_{yy} = 24069$
$\bar{x} = 17.5535$	$\bar{y} = 204.2009$	$n = 100$

a) **(10 points)** The simple linear regression model requires four assumptions. Not all assumptions are needed at every stage of the analysis pipeline.
   i. State the four assumptions.
   ii. For each assumption, identify the stage at which it is first required: model fitting/estimation, statistical inference, or prediction intervals.
   iii. Explain why prediction intervals are not robust to the violation of the assumption identified in ii.

Linearity: The **true relationship** between x and y is linear, meaning $E[Y|x] = \beta_0 + \beta_1 \cdot x$. Part of model fitting/estimation.
Independent Errors: The error terms $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are independent of one another. Needed for proper statistical inference and prediction intervals.
Normality of errors: The error terms $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are normally distributed with **0** mean. Needed for prediction intervals though it is wanted for standard statistical inference but we can rely on CLT.
Homogeneity of Variance: The variance of the errors is constant across all values of x. Needed for statistical inference and prediction intervals.

A **confidence interval** for the **mean response** ($E[Y|x_0]$) benefits from the CLT: since the estimates of the slope and intercept are averages (weighted) of many observations their sampling distributions become approximately normal even when the errors aren't, provided n is reasonably large.

A **prediction interval,** however must account for the variability of a **single future observation** $Y_0 = \beta_0 + \beta_1 \cdot x_0 + \epsilon_0$ as well as that of the **estimate of the mean response.** The interval's coverage depends directly on the distribution of that individual error term ϵ_0 there is no averaging and no CLT to rescue us. If the errors are skewed or heavy-tailed, the interval endpoints (constructed assuming normality) will be in the wrong places, and the stated coverage probability (e.g., 95%) will be incorrect.

b) (8 points) Assuming all assumptions are met, compute the slope $\boldsymbol{b_1}$ and the intercept $\boldsymbol{b_0}$. Write the fitted regression line $\hat{\boldsymbol{y}}$.

$$\boldsymbol{b_1 = \frac{S_{xy}}{S_{xx}} = \frac{-538.3375}{63.797} = -8.4383}$$

$$\boldsymbol{b_0 = \bar{y} - b_1\bar{x} = 204.2009 - (-8.4383) \times 17.5535 = 352.3226}$$

Regression line:

$$\boldsymbol{\hat{y} = 352.3226 - 8.4383x}$$

c) (8 points) Predict the monthly premium for a 17-year-old teen and a 13-year-old teen, respectively. Discuss the statistical validity of these predictions.

Plug in $\boldsymbol{x = 17}$ and $\boldsymbol{x = 13}$ to the regression line:

$$\boldsymbol{\hat{y} = 352.3226 - 8.4383x}$$

$$\boldsymbol{\hat{y}_{17} = 352.3226 - 8.4383 \cdot 17 = 208.6998}$$

$$\boldsymbol{\hat{y}_{13} = 352.3226 - 8.4383 \cdot 13 = 242.4934}$$

The prediction for a 17-year-old is statistically valid, because 17 falls within the range of observed ages. This is an **interpolation**.

The prediction for a 13-year-old is not statistically valid, because 13 falls outside the observed age range. This is an **extrapolation**, and the linear model **may not accurately reflect premiums** for ages not represented in the data.

d) (8 points) Construct a 95% confidence interval for the mean monthly premium of all 17-year-old drivers. Use the R output below, along with the summary statistics from the problem introduction and your fitted regression model.

```
Residual standard error: 14.12 on 98 degrees of freedom
Multiple R-squared: 0.1887, Adjusted R-squared: 0.1804
F-statistic: 22.8 on 1 and 98 DF, p-value: 6.296e-06
```

`qt(0.025, 98, lower.tail=FALSE)` `[1] 1.984467`	`qf(0.025, 1, 98, lower.tail=FALSE)` `[1] 5.181823`
`qt(0.05, 98, lower.tail=FALSE)` `[1] 1.660551`	`qf(0.05, 1, 98, lower.tail=FALSE)` `[1] 3.938111`

Compute the **standard error** of the **mean prediction**

$$SE(\hat{y}_{17}) = s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}} = 14.12 \times \sqrt{\frac{1}{100}+\frac{(17-17.5535)^2}{63.797}} \approx 1.7178964$$

The **95% CI** is $\hat{y}_{17} \pm t_{0.025,df=98} * SE(\hat{y}_{17}) = 208.6998 \pm 1.984467 * 1.7178964 \approx (205.29, 212.11)$

e) (3 points) Which of the following statements is reasonable regarding an interval estimate for a **new response** x^*?

Ⓐ The confidence interval for y^* becomes wider if x^* moves farther away from the sample mean $\bar{x}$.

Ⓑ The confidence interval for y^* becomes narrower if x^* moves farther away from the sample mean $\bar{x}$.

Ⓒ The prediction interval for y^* becomes wider if x^* moves farther away from the sample mean $\bar{x}$.

Ⓓ The prediction interval for $\boldsymbol{y}^*$ becomes narrower if $\boldsymbol{x}^*$ moves farther away from the sample mean $\overline{\boldsymbol{x}}$