Do NOT write in the area above this line. Doing so may cause problems when grading your exam.

Midterm Exam

Course: **DS 1000A Fall 2025**

Sections: **002, 003**

Date: **Nov 2, 2025 1:00 – 3:00 pm (120 minutes)**

Instructors:

Daniel Santaren Guerrero
Pavel Shuldiner

Allowed aids: **A calculator (non-programmable, non-graphing)**

Full Name *(e.g. Tom Marvolo Riddle):*	
Western ID *(e.g. baldemort13):*	
Student Number *(e.g. 251123456):*	

1. Legibly print your Western User ID, full name, and student number in the spaces provided above.
2. The space at the top of each page is reserved for grading. **Do not write on or near the barcode/information there**.
3. The exam has **22 pages**. The last three pages may be used for additional work space. **Clearly indicate** on the original question if you use extra pages, and label your work with the question number.
4. **Write legibly**. Illegible responses may not receive full credit. Post-exam interpretations will not be accepted. **Do not use highlighters or coloured pens as they will not be read by the scanner.**
5. Please round your answers to **2 decimal places.**

Section	**Marks**
Multiple Choice	25
Question 1	12
Question 2	9
Question 3	8
Question 4	10
Question 5	16
Question 6	14
Question 7	6
Total	**100**

Part A – Multiple Choice

Each multiple choice question is worth 1 mark. For each question, select **one** answer only. There are no penalties for incorrect answers, so it is to your advantage to answer every question.

1. What is `Python` in the context of this course?

   A. A large reptile, my instructor keeps one as a pet.

   B. [✓ **A high-level programming language used for data science and statistical analysis.**]

   C. A British comedian from the 70's, I think his first name is Monty?

   D. That's one of the TA's names, right?

   E. That's the title of the DS1000 required textbook.

2. Which of the following best describes the purposes of the `pandas` library?

   A. [✓ **Data manipulation and analysis**]

   B. Emotional support package for Python programmers

   C. Data visualization

   D. A library for storing data about bear species used in laboratory exercises

   E. A tool for analyzing ecological datasets

3. Which of the following Python code snippets import data into a dataframe named `df` from a CSV file named `data.csv`?

   A. `df = pd.import_csv('data.csv')`

   B. `df = pd.load('data.csv', format='csv')`

   C. [✓ `df = pd.read_csv('data.csv')`]

   D. `df = import('data.csv')`

   E. `df = pd.open_csv('data.csv')`

4. A student records daily study times (in hours) over 5 days:

   3 5 6 5 4

   What is the median?

   A. 4.5 hours **B.** [✓ 5 hours] **C.** 6 hours **D.** 5.5 hours **E.** 3 hours

5. What is the type of variable represented by the answer choices in the question below?

> **How many times have you accessed social media today?**
>
> i. zero
> ii. once or twice
> iii. three or four times
> iv. more than four times

A. quantitative

B. [✓ **categorical**]

C. response

D. cannot be determined from the information provided

E. influential

6. The following four observations have mean 2:

$$x_1 = 0, \quad x_2 = 1, \quad x_3 = 2, \quad x_4 = \_\_\_\_$$

What must be the value of x_4?

A. 0 **B.** 2 **C.** 3 **D.** 4 **E.** [✓ **5**]

7. The following three observations have variance 0

$$x_1 = -1, \quad x_2 = -1, \quad x_3 = \_\_\_\_$$

What must be the value of x_3?

A. $x_3 = 2$ **B.** $x_3 = 0$ **C.** $x_3 = 1$ **D.** [✓ $x_3 = -1$] **E.** $x_3 = -3$

8. Which of the following statements is **false**?

A. The interquartile range (IQR) is a measure of spread that is resistant to outliers.

B. [✓ **The response variable is plotted on the x-axis in a scatterplot.**]

C. The correlation coefficient r measures the strength and direction of a linear relationship between two quantitative variables.

D. The normal distribution $N(0, 1)$ has mean 0 with standard deviation 1.

E. Correlation does not imply causation.

9. A sociologist surveyed first-year undergraduates about their daily study hours and nightly sleep hours. The table below summarizes the results.

	Hours Slept			
Hours Studied	**Less than 6**	**6 to 8**	**More than 8**	**Total**
Less than 2	20	40	20	80
2 to 3	10	30	10	50
3 to 4	20	20	20	60
More than 4	20	20	10	50
Total	70	110	60	240

What proportion of the students slept less than 6 hours at night?

A. [✓ **0.29**]

B. 0.3

C. 0.25

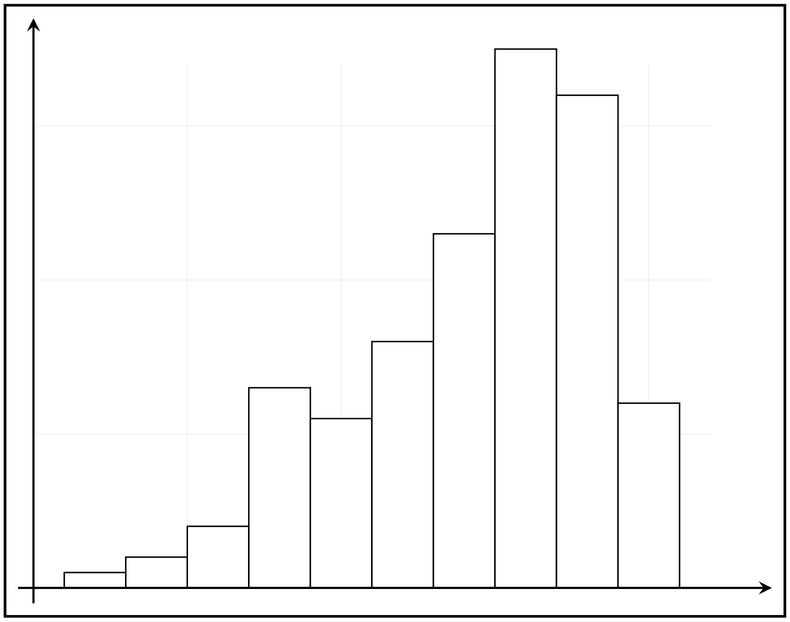
D. 0.33

E. Impossible to deduce from the information provided

10. In Naples, Florida, most houses are priced between $200,000 and $500,000, but a few waterfront properties cost up to $150 million. The distribution of house prices is best described as

A. symmetric **B.** normal **C.** resistant **D.** left skewed **E.** [✓ **right skewed**]

11. Which of the following statements best describes the histogram below?



A. [✓ **It is left-skewed.**]

B. It is approximately symmetric.

C. It is right-skewed.

D. There are several outliers.

E. It appears to be from a normal distribution.

12. Which statement about correlation is correct?

    A. [✓ **Changing the units of measurement does not change the correlation.**]

    B. A correlation of -0.9 indicates weak association between variables.

    C. Correlation values range from 0 to 1.

    D. It depends on which variable is the explanatory variable (x) and which is the response variable (y).

    E. If the correlation is 0, then there is no relationship between the two variables.

13. A researcher studying plant growth calculates the regression equation: height $= 2.5 + 0.8x$, where x is days of sunlight exposure. What can we conclude about the correlation between height and sunlight exposure?

    A. [✓ r **is positive, but the exact value cannot be determined from this information.**]

    B. r is negative, but the exact value cannot be determined from this information.

    C. $r = 0.8$

    D. $r = \dfrac{1}{0.8}$

    E. $r = 2.5$

14. What does `np.random.seed(42)` do?

    A. Generates 42 random numbers and stores them in memory for faster access

    B. Limits all random values to be between 0 and 42

    C. [✓ **Sets the random seed to 42, ensuring reproducible random number generation**]

    D. Creates a random array with 42 elements as the default size

    E. Sets 42 as the starting value for all random number sequences

15. Which of the following outputs is printed by the following Python code snippet?

```
heights = [75, 64.5, 67, 89, 70.4, 63, 50.9]
print(heights[2:4])
```

A. [✓ $[67, 89]$] **B.** $[67, 89, 70.4]$ **C.** $67, 89$ **D.** $70.4, 63, 50.9$ **E.** IndexError

16. Which statement about density curves is **false**?

- **A.** Density curves never fall below the horizontal axis.
- **B.** The area under the curve between two values represents the proportion of observations in that interval.
- **C.** [✓ **The total area under any density curve depends on the curve's shape.**]
- **D.** Density curves represent an idealized model of a distribution.
- **E.** The median divides the area under the curve in half.

17. What is the coefficient of determination for the linear regression model obtained based on the following Python code snippet and its output?

```
model_modified = stats.linregress(x = possum_modified['age'], y = possum_modified['headL'])
model_modified.intercept
model_modified.slope
model_modified.rvalue
```

```
90.72
0.36
0.26
```

A. 0.9072 **B.** 0.36 **C.** 0.26 **D.** 0.13 **E.** [✓ **0.07**]

18. In the following stemplot, stems are whole percentages and leaves are tenths. Which of the following statements is true?

4	2
5	
6	
7	
8	
9	7 9
10	0 8
11	1 5 5 6 6
12	0 1 2 2 2 3 4 4 4 4 5 7 8 8 8 9 9 9
13	0 1 2 3 3 3 3 3 4 4 4 8 9 9
14	0 2 6 6 6
15	2 3
16	8

- **A.** The distribution is right-skewed
- **B.** 16.8 is an outlier
- **C.** The mean is approximately 11.1
- **D.** The mode is 12.3
- **E.** [✓ **4.2 is an outlier**]

19. Suppose we have a variable X that follows a standard Normal distribution. Which of the following best approximates the proportion of values of X have z-scores between 0 and 2?

A. 68% **B.** 50% **C.** 99.7% **D.** 95% **E.** [✓ 47.7%]

20. A least-squares regression line has slope 450. After removing one data point and the least-squares regression line, the new slope is -10,000. The removed point is best described as

A. [✓ **influential**]

B. resistant

C. low leverage point

D. lowkey based

E. an outlier in only the y-direction

21. A variable has the following five-number summary: 0, 10, 20, 30, 70. How many potential outliers does this variable contain?

A. [✓ **At least one**] **B.** 1 **C.** 0 **D.** 2 **E.** 3

22. Which of the following best describe the direction, form and strength of the relationship in the scatterplot below?

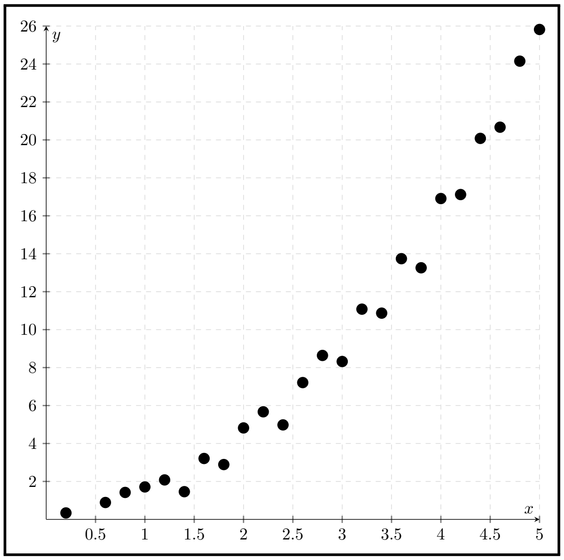
A. positive, linear, strong

B. negative, linear, strong

C. [✓ **positive, non-linear, strong**]

D. negative, non-linear, strong

E. no relationship



23. Suppose it was found that the least-squares regression line predicting house price (y) from size in square feet (x) is given by

$$\hat{y} = 20000 + 150x$$

Which of the following provides the correct interpretation regarding the slope of the regression line?

A. **As house size increases by 1 square foot, the expected (or average) increase in house price is $150.**

B. As house size increases by 150 square feet, the expected (or average) increase in house price is $20,000.

C. As house size increases by 20,000 square feet, the expected (or average) increase in house price is $150.

D. As house size increases by 1 square foot, the expected (or average) increase in house price is $20,000.

E. As house price increases by $1, the expected (or average) increase in house size is 150 square feet.

24. Continuing from question 23: Which of the following gives the least-squares regression line predicting square footage (y) based on house price (x)?

A. **The line cannot be determined from the information provided.**

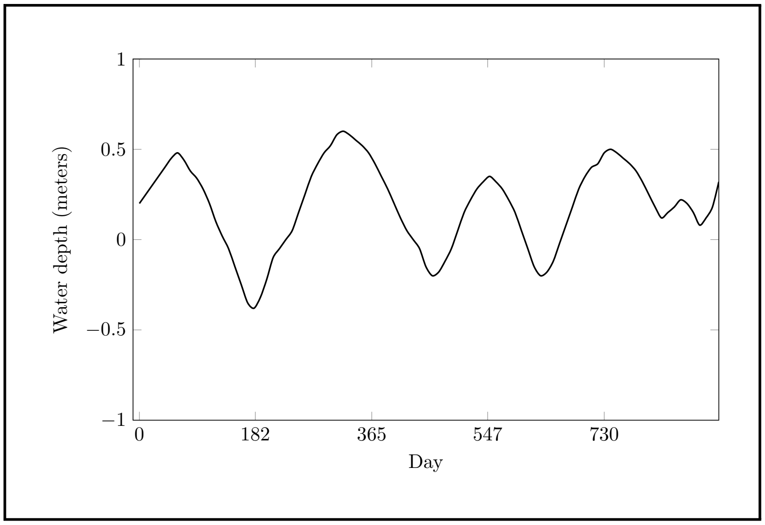
B. $\hat{y} = -133.33 + 0.0067x$

C. $\hat{y} = 20000 + 150x$

D. $\hat{y} = -20000 + 150x$

E. $\hat{y} = 20000 + \frac{1}{150}x$

25. The following time plot shows the water depth (in meters) measured in a well over a period of three years. Which of the following statements best describes what we can infer from the plot?



A. The water depth trends downwards.

B. The water depth trends upwards.

C. **The water depth shows a seasonal pattern with no overall trend.**

D. The water depth fluctuates randomly with no discernible pattern.

E. A linear model of water depth over time would be a good fit.

Short Answer

Question 1 (7 marks)

A random sample of 7 graduate students tracked their weekly coffee consumption (in cups), and their results are as follows

3	6	6	7	10	10	10

(a) (2 marks) What is/are the mode(s) of weekly coffee consumption of graduate students?

Solution: The mode is $\boxed{10}$ as it appears 3 times.

(b) (5 marks) Determine the five-number summary for the students' weekly coffee consumption.

Solution:

$$\begin{aligned} \text{Minimum:} &\quad 3 \\ Q_1: &\quad \frac{3+6}{2} = 4.5 \text{ or } 6 \text{ (different methods)} \\ Q_2 \text{ (Median)}: &\quad 7 \\ Q_3: &\quad \frac{10+10}{2} = 10 \\ \text{Maximum:} &\quad 10 \end{aligned}$$

Five-number summary: $\boxed{3, 6, 7, 10, 10}$

Question 2 (9 marks)

A random sample of 3 undergraduate students tracked their average daily hours of sleep, and the results are as follows:

2	4	6

(a) (1 mark) What kind of variable is the average daily hours of sleep? You do not have to justify your answer.

Solution: Quantitative variable

(b) (1 mark) What are the individuals described by this dataset? You do not have to justify your answer.

Solution: The individuals are the 3 undergraduate students.

(c) (2 marks) Name two visualizations that are appropriate for a variable of this type. You do not have to draw or justify your answer.

Solution: Any two of: histogram, stemplot, boxplot, density plot

(d) (1 mark) Describe the shape of this distribution. You do not have to draw or justify your answer.

Solution: The shape is $\boxed{\text{symmetric}}$.

(e) (4 marks) Find the sample mean and standard deviation for the undergraduate students' average daily hours of sleep.

Solution: Sample mean:

$$\overline{x} = \frac{2+4+6}{3} = \frac{12}{3} = \boxed{4}$$

Sample standard deviation:

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{3-1}} \\ &= \sqrt{\frac{4+0+4}{2}} \\ &= \sqrt{\frac{8}{2}} = \sqrt{4} = \boxed{2} \end{aligned}$$

Question 3 (5 marks)

The variable x has the following observations:

5	10	10	10	11	12	12	13

It is known that the quartile values of x are

$$Q_1 = 10, \quad Q_2 = 10.5, \quad Q_3 = 12$$

(You may use these values to answer the following questions without recalculating them.)

(a) (2 marks) Compute the interquartile range for this variable.

Solution: The interquartile range (IQR) is given by

$$\text{IQR} = Q_3 - Q_1 = 12 - 10 = \boxed{2}$$

(b) (3 marks) Draw a modified boxplot of this data.

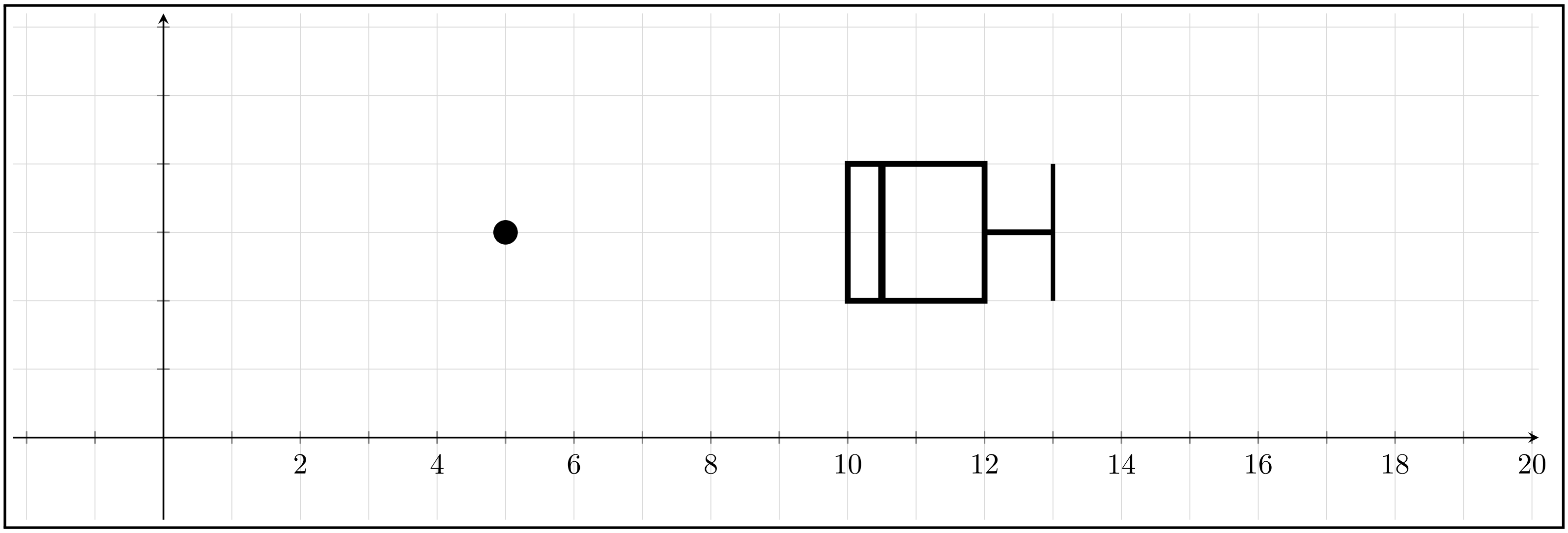
Solution: First, we need to identify outliers using the 1.5 × IQR rule:

Lower fence: $Q_1 - 1.5 \times \text{IQR} = 10 - 1.5(2) = 10 - 3 = 7$

Upper fence: $Q_3 + 1.5 \times \text{IQR} = 12 + 1.5(2) = 12 + 3 = 15$

The value 5 is below the lower fence of 7, so it is an outlier.

Modified boxplot should show: - Outlier at 5 (marked with a point or asterisk) - Left whisker extends to 10 (the smallest non-outlier value) - Box from $Q_1 = 10$ to $Q_3 = 12$ with median line at $Q_2 = 10.5$ - Right whisker extends to 13 (the maximum value)



Question 4 (13 marks)

The first two Starbucks cafes which opened in Canada are located in Vancouver and Toronto. The ratings of nearby cafes were collected and it was found that cafe ratings are **normally distributed** in both cities.

In Vancouver, the average cafe rating is 3.8 with a standard deviation of 0.2, while in Toronto, the average cafe rating is 4.1 with a standard deviation of 0.3. The first Vancouver Starbucks cafe has a rating of 4.2, while the first Toronto Starbucks cafe has a rating of 4.4.

	Rating	**Mean**	**Standard Deviation**
Vancouver	4.2	3.8	0.2
Toronto	4.4	4.1	0.3

(a) (4 marks) What proportion of Vancouver cafes have ratings at or below the first Vancouver Starbucks cafe's rating of 4.2?

Solution: Let $X \sim N(3.8, 0.2)$ denote the rating of a cafe in Vancouver.
To find the desired proportion, we calculate the z-score:

$$z = \frac{4.2 - 3.8}{0.2} = \frac{0.4}{0.2} = 2.0$$

Using the standard normal table, $P(Z \leq 2.0) \approx 0.9772$ or $\boxed{0.98}$ rounded to two decimal places or (they may also leave it as $\boxed{97.72\%}$).

Solution: Alternatively, using the empirical rule: approximately 95% of data lies within 2 standard deviations of the mean in a normal distribution, and since 4.2 is exactly 2 standard deviations above the mean, we can estimate that about $\boxed{97.5\%}$ of cafes have ratings at or below this value.

(b) (3 marks) Suppose that the first Toronto Starbucks cafe ranks at the 84th percentile among Toronto cafes. Based on part (a), which cafe's rating is higher relative to its city? Briefly explain.

Solution: The Vancouver Starbucks cafe has a higher relative rating because its percentile (97.72% or approximately 98th percentile) is greater than the Toronto Starbucks cafe's percentile (84th percentile).

This means the Vancouver cafe ranks higher compared to other cafes in its city than the Toronto cafe does in its city.

Solution: One can also appeal to $z-$scores: The Vancouver cafe has a $z-$score of 2.0, while the Toronto cafe's $z-$score is 1.0. Since $2.0 > 1.0$, the Vancouver cafe has a higher relative rating.

(c) (5 marks) What proportion of cafes in Vancouver have ratings between 3.4 and 4.2?

Solution: Let $X \sim N(3.8, 0.2)$ represent the rating of a cafe in Vancouver. We are interested in

$$P(3.4 < X < 4.2).$$

Recall that $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, we can standardize the variable X to find the desired proportion.

$$\begin{aligned} P(3.4 < X < 4.2) &= P\left(\frac{3.4-3.8}{0.2} < \frac{X-3.8}{0.2} < \frac{4.2-3.8}{0.2}\right) \\ &= P(-2.0 < Z < 2.0). \end{aligned}$$

Calculating z-scores:

$$\begin{aligned} z_1 &= \frac{3.4-3.8}{0.2} = \frac{-0.4}{0.2} = -2.0 \\ z_2 &= \frac{4.2-3.8}{0.2} = \frac{0.4}{0.2} = 2.0 \end{aligned}$$

Therefore:

$$P(-2.0 < Z < 2.0) = P(Z < 2.0) - P(Z < -2.0) \approx 0.9772 - 0.0228 = 0.9544$$

The desired proportion or probability is $\boxed{0.95 \text{ or } 95.44\%}$.

Solution: Alternatively, using the empirical rule: approximately 95% of data lies within 2 standard deviations of the mean in a normal distribution.

Since 3.4 and 4.2 are exactly 2 standard deviations below and above the mean respectively, we can estimate that about $\boxed{95\%}$ of cafes have ratings between these two values.

(d) (1 mark) What is the probability that a randomly selected cafe in Vancouver has a rating exactly equal to 4.1?

Solution: For any continuous distribution (such as the normal distribution), the probability of any exact value is 0.

$$\begin{aligned} P(X = 4.1) &\approx P(4.1 \leq X \leq 4.1) \\ &= P(X \leq 4.1) - P(X \leq 4.1) \\ &= 0 \end{aligned}$$

Question 5 (15 marks)

A study tracked 1000 days at an ice cream shop to examine the relationship between daily temperature (x in °C) and daily ice cream sales (y, in cones sold).

$$r = 0.8, \quad \overline{x} = 20\,°\text{C}, \quad s_x = 5\,°\text{C}, \quad \overline{y} = 150\,\text{cones}, \quad s_y = 25\,\text{cones}$$

(a) (2 marks) What is the proportion of variation in ice cream sales that is explained by the least-squares regression on the daily temperature?

Solution: The coefficient of determination is:

$$R^2 = r^2 = (0.8)^2 = \boxed{0.64}$$

(b) (3 marks) What are the slope and the intercept of the least-squares regression line?

Solution: Slope:

$$b = r \cdot \frac{s_y}{s_x} = 0.8 \cdot \frac{25}{5} = 0.8 \cdot 5 = \boxed{4.0}$$

Intercept:

$$a = \bar{y} - b\bar{x} = 150 - 4.0 \times 20 = 150 - 80 = \boxed{70}$$

(c) (2 marks) Write down the least-squares regression line using the coefficients you found in part (b).

Solution: The least-squares regression line is

$$\hat{y} = 70 + 4.0x$$

where x is the daily temperature in °C and $\hat{y}$ is the predicted ice cream sales in cones.

(d) (4 marks) What is the predicted ice cream sales when the daily temperature is 60°C, using the regression line you found in part (b)? Is there a problem with this prediction? Briefly explain.

Solution: When $x = 60$°C, the predicted ice cream sales is

$$\hat{y} = 70 + 4.0 \times 60 = 70 + 240 = \boxed{310} \text{ cones}$$

This prediction is not reliable because:

- 60°C is far outside the observed data range (extrapolation).
- 60°C is not a realistic daily temperature for humans.

(e) (4 marks) Suppose the regression line is $\hat{y} = 100 + 10x$. If $(x, y) = (10, 100)$ is a data point in the dataset, what would be its residual?

Solution: The residual is the difference between the observed value and the predicted value.

Given the regression line $\hat{y} = 100 + 10x$ and the data point $(x, y) = (10, 100)$:

Predicted value:

$$\hat{y} = 100 + 10 \times 10 = 100 + 100 = 200$$

Residual:

$$e = y - \hat{y} = 100 - 200 = \boxed{-100}$$

Question 6 (14 marks)

A university is comparing two tutors (Tutor A and Tutor B) to determine which is more effective at helping students pass their exams. Data was collected from 80 students across beginner-level and advanced-level courses. Their combined results are summarized in the table below.

Result	**Tutor A**	**Tutor B**	**Total**
Fail	18	8	26
Pass	42	12	54
Total	60	20	80

(a) (1 mark) Find the probability that a student was tutored by Tutor A given that they failed.

Solution: There are 26 students who failed, and 18 of them were tutored by Tutor A. Thus, the desired probability is

$$\frac{18}{26} \approx \boxed{0.69} \text{ or } 69.23\%$$

(b) (2 marks) Find the odds of passing for students tutored by Tutor A.

Solution: The odds of passing for students tutored by Tutor A is the ratio of the number who passed to the number who failed:

$$\text{Odds} = \frac{42}{18} \approx \boxed{2.33}$$

So, for every student tutored by Tutor A who failed, about 2.33 passed.

(c) (3 marks) Calculate the pass rate (as a proportion) for each tutor. Based on the combined data, which tutor is more effective?

Solution: Pass rate for Tutor A:

$$\frac{42}{60} = 0.70 \text{ or } 70\%$$

Pass rate for Tutor B:

$$\frac{12}{20} = 0.60 \text{ or } 60\%$$

Based on the combined data, Tutor A is more effective.

The results are broken down by course level in the tables below.

Beginner Level			
Result	**Tutor A**	**Tutor B**	**Total**
Fail	10	1	11
Pass	40	9	49
Total	50	10	60

Advanced Level			
Result	**Tutor A**	**Tutor B**	**Total**
Fail	8	7	15
Pass	2	3	5
Total	10	10	20

(d) (6 marks) Calculate the pass rate (as a percentage) for each tutor at each course level. Which tutor is more effective for beginner-level courses? Which is more effective for advanced-level courses?

Solution:

Course Level	**Tutor A Pass Rate**	**Tutor B Pass Rate**
Beginner	$\frac{40}{50} = 0.80\ (80\%)$	$\frac{9}{10} = 0.90\ (90\%)$
Advanced	$\frac{2}{10} = 0.20\ (20\%)$	$\frac{3}{10} = 0.30\ (30\%)$

Therefore, Tutor B is more effective at both course levels.

- Beginner level: 90% (Tutor B) vs 80% (Tutor A)
- Advanced level: 30% (Tutor B) vs 20% (Tutor A)

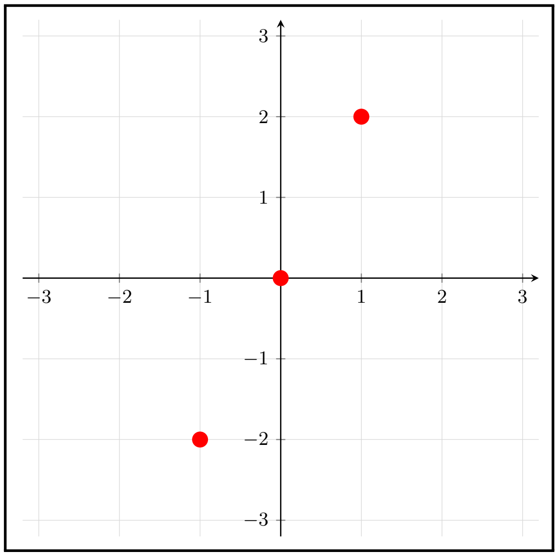
(e) (2 marks) What statistical phenomenon explains why the better tutor in part (c) differs from the better tutor at each individual course level in part (d)?

Solution: This is an example of Simpson's Paradox.

Question 7 (12 marks)

(a) (3 marks) Plot the following data on the scatterplot provided.

x	y
-1	-2
0	0
1	2



(b) (5 marks) Find the correlation of x and y. Be sure to justify your answer.

Solution: Method 1 (Visual)

The data points lie on the line $y = 2x$, showing a perfect positive linear relationship. Hence, the correlation coefficient r is equal to 1.

Solution: Method 2 (Direct):

$$\overline{x} = \frac{-1+0+1}{3} = 0 \qquad \overline{y} = \frac{-2+0+2}{3} = 0$$

$$s_x = \sqrt{\frac{(-1-0)^2+(0-0)^2+(1-0)^2}{3-1}} = \sqrt{\frac{2}{2}} = 1 \quad s_y = \sqrt{\frac{(-2-0)^2+(0-0)^2+(2-0)^2}{3-1}} = \sqrt{\frac{8}{2}} = 2$$

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \\ &= \frac{1}{2}\left[\left(\frac{-1}{1}\right)\left(\frac{-2}{2}\right)+\left(\frac{0}{1}\right)\left(\frac{0}{2}\right)+\left(\frac{1}{1}\right)\left(\frac{2}{2}\right)\right] \\ &= \frac{1}{2}\left[(-1)(-1)+0+(1)(1)\right] = \frac{1}{2}[1+0+1] = 1 \end{aligned}$$

(c) (2 marks) Find the correlation of y and x. Be sure to justify your answer.

y	x
-2	-1
0	0
2	1

Solution: As correlation is symmetric, $r = 1.00$.

(d) (2 marks) Find the correlation of x and z, where $z = 2y + 1$. Be sure to justify your answer.

$\boldsymbol{x}$	$\boldsymbol{z}$
-1	-3
0	1
1	5

Solution: The correlation between x and z is $r = 1.00$.

Correlation is not affected by changes in units. Hence, multiplying y by 2 and adding 1 to get z does not change the correlation.

This page was intentionally left blank.

You may use this space if you run out of room for a particular question.

If you do, be sure to indicate this clearly here on this page and also on the question page.

Scratch Work

This page will not be scanned.