Do NOT write in the area above this line.

Midterm Exam

Course: DS 1000B Winter 2026

Sections: 002, 003

Date: March 1st, 2026

2:00 – 4:00 pm (120 minutes)

Instructors:

Marieke Mur

Pavel Shuldiner

Allowed aids:

A calculator (non-programmable, non-graphing)

Formula sheet provided for you with the exam.

Full Name (print) *(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. Do **not** detach the pages of the exam. You may ask for scrap paper if needed.
3. The space at the top of each page is reserved for the scanner. Do **not** write on or near the barcode.
4. Do **not** use highlighters, coloured pens, pencils or markers.
5. The exam has 22 pages. The last three pages may be used for additional workspace or scrap paper.
6. When applicable, enter your final answer in the provided answer box rounded to 2 decimal places.

Section	**Marks**
Multiple Choice (1 mark each)	15
Miscellaneous	10
Section 1	5
Section 2	13
Section 3	7
Section 4	6
Section 5	25
Section 6	19
Total	**100**
Bonus Question	*3*

Multiple Choice (1 mark each) **(15 marks)**

Each of the following multiple-choice questions is worth 1 mark. **Circle** the best answer for each question on the exam paper. You may select only one option per question.

No justification is required and no partial marks will be awarded.

MC1. (1 mark) A dataset containing house prices in a city is found to be right-skewed. Which of the following relationships between the mean and median is **generally** true?

A. Mean $<$ Median

B. Mean $\approx$ Median

C. **Mean $>$ Median**

D. Mean $=$ Mode

E. The relationship cannot be determined.

MC2. (1 mark) What type of variable is "Student ID Number" (e.g., 251123456)?

A. Response **B.** Quantitative **C.** **Categorical** **D.** Explanatory **E.** Confounding

MC3. (1 mark) Consider the dataset: $5, 2, 8, 1, 9, 6$. What is the median?

A. 5 **B.** **5.5** **C.** 6 **D.** 5.17 **E.** 4.5

MC4. (1 mark) A dataset has $Q_1 = 20$ and $Q_3 = 50$. Using the $1.5 \times$ IQR rule, which of the following values would be considered a potential outlier?

A. 0 **B.** 30 **C.** -20 **D.** **96** **E.** 55

MC5. (1 mark) Which of the following statistics is non-resistant (sensitive) to outliers?

A. Median

B. Interquartile Range (IQR)

C. First Quartile (Q_1)

D. **Variance**

E. Mode

MC6. (1 mark) If a variable has a variance of $s^2 = 0$, what can we conclude?

A. The mean is 0.

B. The data has outliers.

C. **All of the observations are identical.**

D. The distribution is symmetric.

E. The data contains only one observation.

MC7. (1 mark) A correlation coefficient of $r = -0.85$ suggests:

A. A weak negative linear relationship.

B. A strong positive linear relationship.

C. **A strong negative linear relationship.**

D. A weak positive linear relationship.

E. No relationship.

MC8. (1 mark) Which of the following values is a valid correlation coefficient?

A. 1.05 **B.** -2.0 **C.** **-0.3** **D.** 100 **E.** π

MC9. (1 mark) What do the columns of a dataset represent?

A. Individuals **B.** Samples **C.** **Variables** **D.** Observations **E.** Categories

MC10. (1 mark) The coefficient of determination of a linear regression model is $R^2 = 0.64$. What is the correlation coefficient r?

A. 0.16 **B.** 0.36 **C.** 0.80 **D.** 0.64 **E.** **Impossible to tell**

MC11. (1 mark) The ecological fallacy occurs when:

A. Individual-level data is used to make inferences about group-level patterns.

B. A confounding variable is mistakenly identified as causal.

C. **Group-level data is used to make inferences about individuals.**

D. Drawing conclusions based on a small or non-representative sample.

E. A correlation is mistaken for causation.

MC12. (1 mark) A residual plot would have x-axis representing:

A. **The explanatory variable x (or predicted values $\hat{y}$)**

B. The response variable y

C. The residuals

D. The squared residuals

E. The sample size

MC13. (1 mark) An observation is considered influential if:

A. It is extreme in both the x and y coordinates.

B. It is extreme in the $x-$coordinate.

C. **Removing it substantially changes the regression line explaining y through x.**

D. It lies exactly on the regression line explaining y through x.

E. It is extreme in the $y-$coordinate.

MC14. (1 mark) The following is a stem-and-leaf plot of test scores. Describe the shape of the distribution.

5	2
6	3 5
7	1 4 6 8
8	2 5 7 8 9
9	3 4 6 7 8 9

A. Right-skewed

B. Symmetric

C. **Left-skewed**

D. Bimodal

E. Multimodal

MC15. (1 mark) Simpson's Paradox occurs when:

A. **A trend appearing in different groups reverses when the groups are combined.**

B. A correlation is approximately 0.

C. An outlier drastically changes the regression slope.

D. Correlation is mistaken for causation.

E. The sample size is too small.

Miscellaneous **(10 marks)**

Q1. The following questions are unrelated to each other.

(a) (2 marks) The following question appeared on a survey:

Which day of the week were you born on?

and it resulted in the following data:

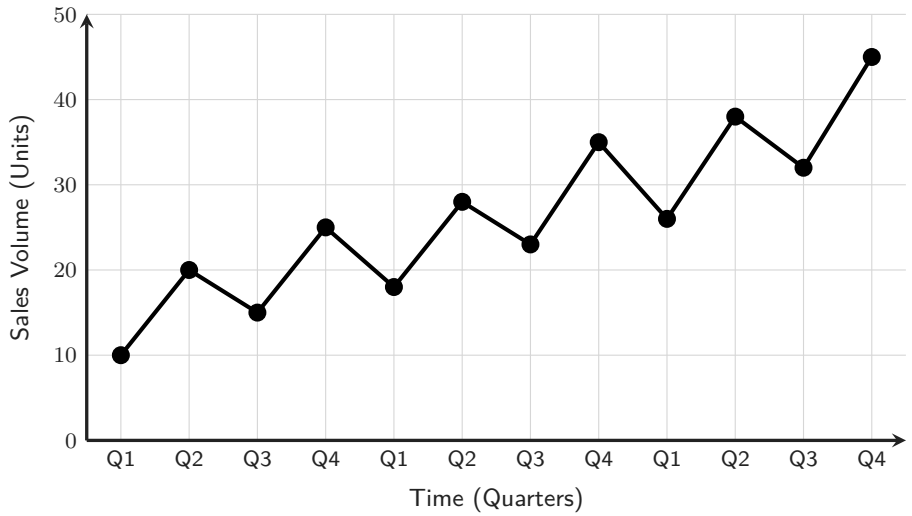
Day of the Week	Number of Participants
Monday	12
Tuesday	15
Wednesday	10
Thursday	9
Friday	14
Saturday	8
Sunday	7

Could a pie chart be used to display this distribution? Explain.

Solution: Yes. A pie chart is appropriate because

- the categories are mutually exclusive (each participant was born on exactly one day), and
- the categories are exhaustive (the seven days cover all possible outcomes).

(b) (3 marks) Interpret the following time series plot.

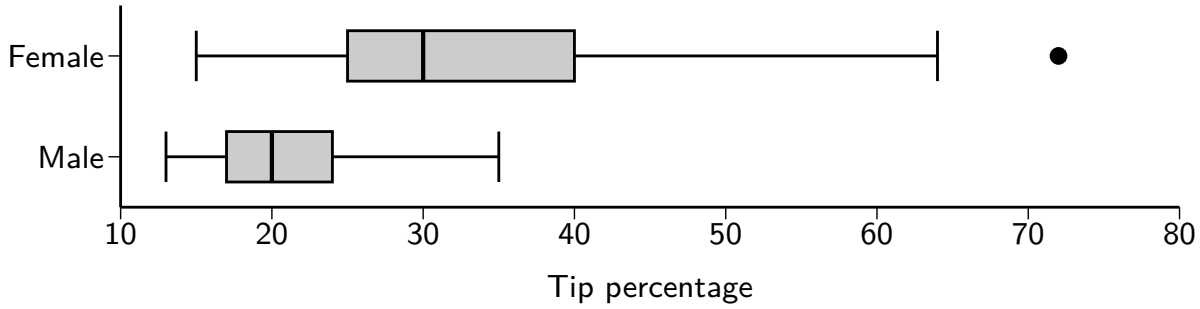


Solution: The plot displays an upward trend over time. There is also a seasonal pattern: within each year, values rise and fall in a regular quarterly cycle.

(c) (2 marks) Explain the difference between a bar chart and a histogram.

Solution: A bar chart displays the distribution of a categorical variable, with gaps between bars to emphasize distinct categories. A histogram displays the distribution of a quantitative variable, with bars that may have no gaps to represent intervals or ranges of values.

(d) (3 marks) The following boxplots display the distributions of tip percentages (calculated as a percentage of the total bill) earned by waitstaff, grouped by gender. Compare the two distributions by comparing their spread, center, and any potential outliers.



Solution: Comparing the two distributions:

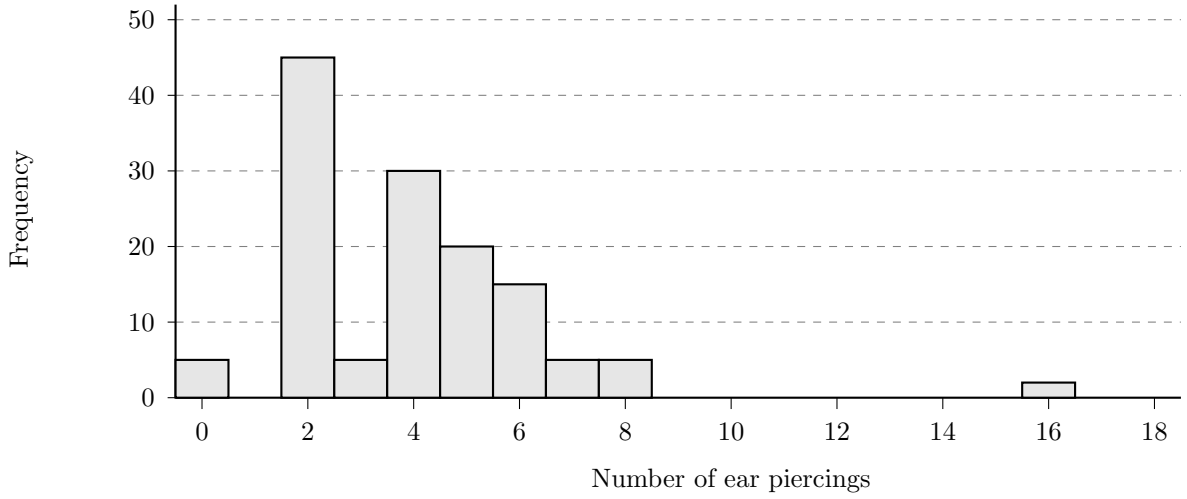
Center: Female waitstaff have a higher median tip percentage (approximately 30%) compared to male waitstaff (approximately 20%).

Spread: The IQR for females (approximately 15%) is larger than that for males (approximately 7%), indicating more variability in tip percentages for females.

Outliers: There is one outlier in the female distribution (approximately 78%), while the male distribution has no outliers.

Section 1 **(5 marks)**

Q2. The number of ear piercings of a random sample of female university students is recorded. The following visualization displays the distribution of the data collected.



(a) (1 mark) What type of variable is "number of ear piercings"?

Solution: Quantitative / numeric (discrete)

Note: The labels "independent" and "dependent" are relational terms that require context about a study design; without that context, neither label applies. Hence, those answers would not be accepted.

(b) (1 mark) Describe the shape of the distribution.

Solution: Right-skewed (positively skewed).

(c) (2 marks) Are there any potential outliers in the data? If so, identify them approximately. If not, write *NONE*.

Solution: Yes, there appears to be potential outliers around 16 piercings (isolated bar far from the main distribution).

(d) (1 mark) Approximately how many participants had exactly 2 ear piercings?

Solution: 45

Answers between 43-47 should be accepted

Section 2 **(13 marks)**

Q3. As a data analyst for Spotify, you are investigating the listening habits of university students. You have collected data on the number of minutes spent listening to music per day for a random sample of 8 students.

20	40	50	60
70	100	100	185

(a) (5 marks) Calculate the five-number summary of the daily listening times.

Solution: First, order the data: $20, 40, 50, 60, 70, 100, 100, 185$.

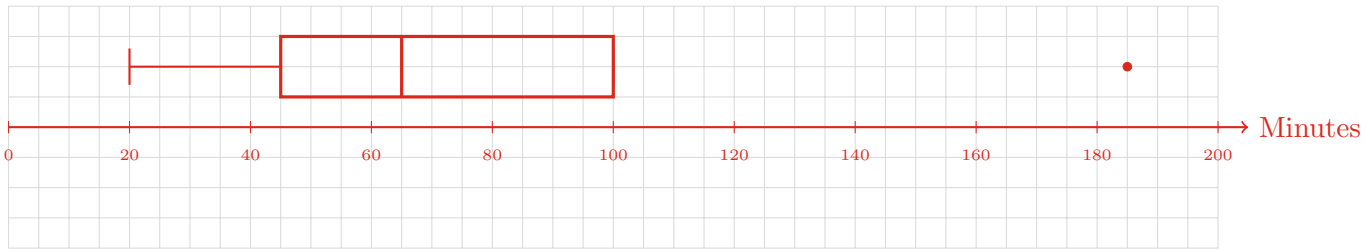
With $n = 8$:

- Minimum: 20
- Q_1 (median of lower half: $20, 40, 50, 60$): $\frac{40+50}{2} = 45$
- Median (average of 4th and 5th values): $\frac{60+70}{2} = 65$
- Q_3 (median of upper half: $70, 100, 100, 185$): $\frac{100+100}{2} = 100$
- Maximum: 185

$\boxed{\text{Min} = 20,\ Q_1 = 45,\ \text{Median} = 65,\ Q_3 = 100,\ \text{Max} = 185}$

(b) (4 marks) Draw the modified boxplot for the data above.

Solution:



$\text{IQR} = Q_3 - Q_1 = 100 - 45 = 55$

Lower fence: $Q_1 - 1.5 \times \text{IQR} = 45 - 82.5 = -37.5$

Upper fence: $Q_3 + 1.5 \times \text{IQR} = 100 + 82.5 = 182.5$

The value 185 exceeds the upper fence of 182.5, so it is marked as an outlier (plotted as an individual point). The upper whisker should extend to 100 (the largest non-outlier value) which is also Q_3, hence it does not extend further.

(c) (4 marks) Which descriptive statistics should one use to report the center and spread of this data? Justify your answer.

Solution: Use the median and IQR. The distribution has a significant outlier (185 minutes) and is right-skewed. The median and IQR are resistant (robust) to outliers, making them more appropriate than the mean and standard deviation.

Section 3 **(7 marks)**

Q4. A teaching assistant records the raw quiz scores (out of 10) for a random sample of 3 students:

2	4	6

(a) (4 marks) Compute the mean and standard deviation of the quiz scores. Show your work.

Solution: The mean is calculated as

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

Sum of squared deviations: $4 + 0 + 4 = 8$. Hence,

$$s^2 = \frac{8}{3-1} = \frac{8}{2} = 4$$

and so

$$s = \sqrt{4} = \boxed{2}$$

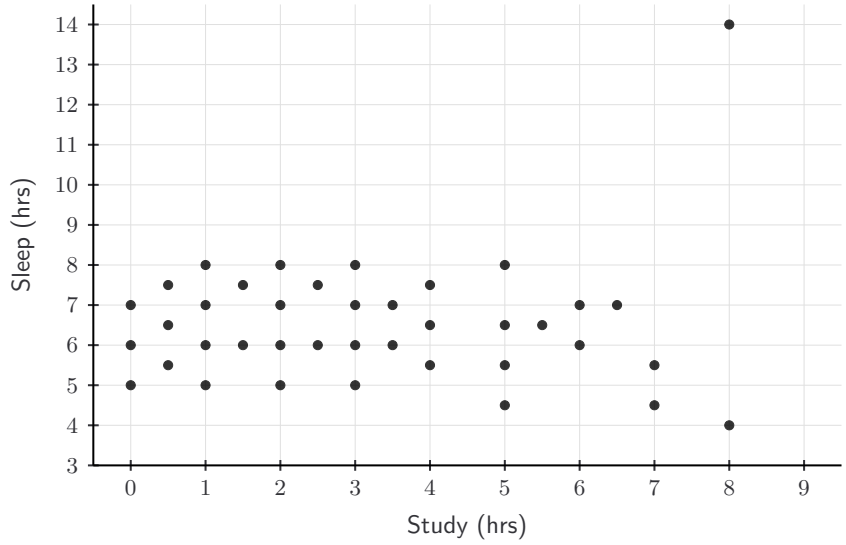
(b) (3 marks) If the quiz scores were converted to be out of 100 (by multiplying each score by 10), what would the new standard deviation be?

Solution: Multiplying every value by a constant $c = 10$ multiplies the standard deviation by $|c|$.

$$s_{\text{new}} = 10 \times s = 10 \times 2 = \boxed{20}$$

Section 4 **(6 marks)**

Q5. The following scatterplot displays the relationship between the daily number of hours spent studying and the daily number of hours spent sleeping for a sample of university students.



(a) (1 mark) Are there any outliers in this scatterplot? If so, identify them by writing their coordinates. If not, write *NONE*.

Solution: Yes, the point approximately at $(8, 14)$ is an outlier: it lies far above the general pattern of other points.

(b) (1 mark) How many students get at most 4.5 hours of sleep per day?

Solution: $\boxed{3}$

(c) (4 marks) A researcher claims that the correlation is approximately $r = 0.95$. Does the scatterplot support this claim? Justify your answer.

Solution: No, the scatterplot does not support this claim for two reasons:

i. The overall trend appears to be negative (more study hours associated with fewer sleep hours), so if anything, r should be negative, not positive.

ii. The relationship is weak and has substantial scatter, not the tight linear pattern that would produce $|r| \approx 0.95$.

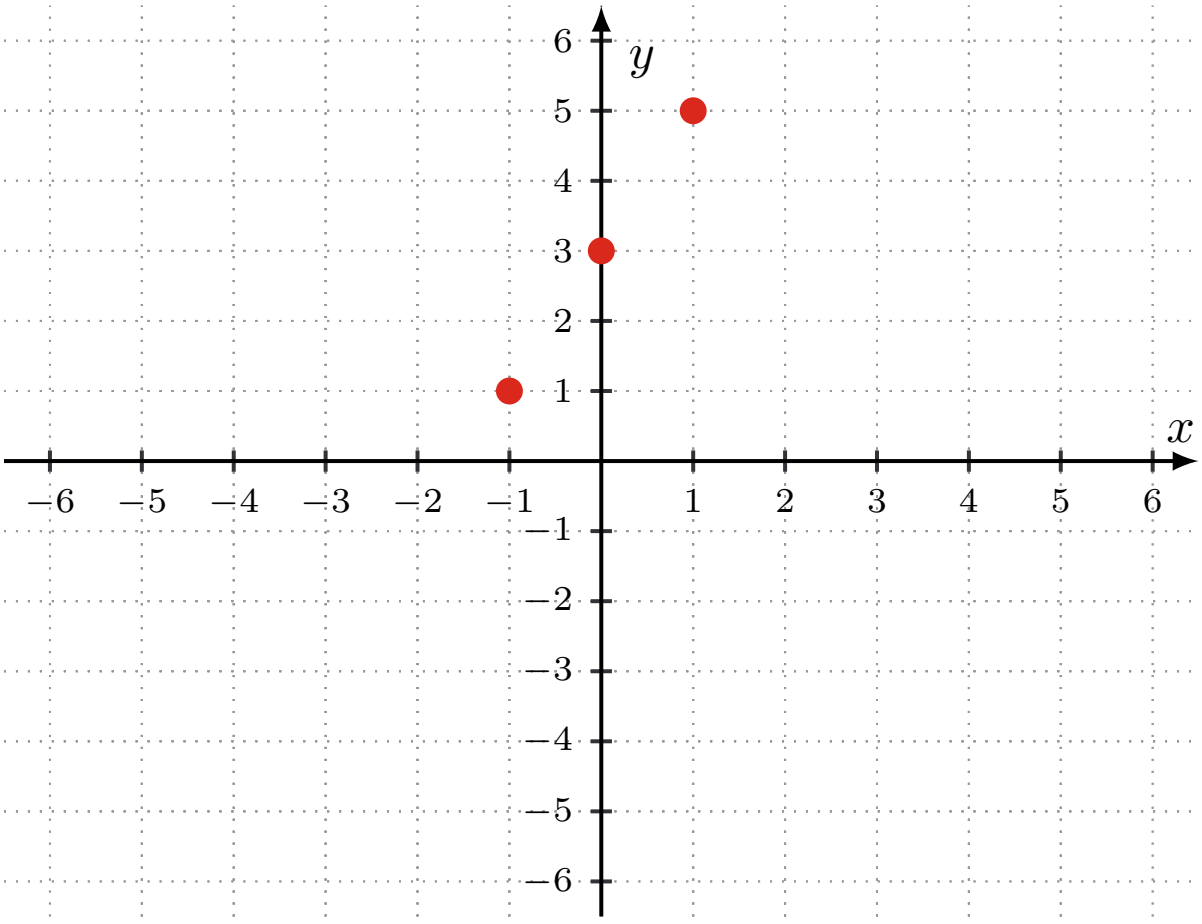
Section 5 **(25 marks)**

Q6. The table below shows temperature deviation from average (x, in °C) and heating energy usage (y, in hundreds of kWh) for three winter days:

x	y
-1	1
0	3
1	5

(a) (2 marks) Plot the three points on the scatterplot below.

Solution:



Recall that x and y are given by the following table:

x	y
-1	1
0	3
1	5

(b) (6 marks) Find the equation of the least-squares regression line for predicting heating energy usage from temperature deviation.

Solution: First, compute the mean and standard deviation of x and y.

$$\text{For } x: \quad \bar{x} = \frac{-1+0+1}{3} = 0 \qquad \text{For } y: \quad \bar{y} = \frac{1+3+5}{3} = 3$$

$$s_x = \sqrt{\frac{(-1)^2+0^2+1^2}{3-1}} = \sqrt{\frac{2}{2}} = 1 \qquad s_y = \sqrt{\frac{(1-3)^2+(3-3)^2+(5-3)^2}{3-1}} = \sqrt{\frac{8}{2}} = 2$$

Now, compute the z-scores:

$$z_{x,1} = \frac{-1-0}{1} = -1, \quad z_{x,2} = \frac{0-0}{1} = 0, \quad z_{x,3} = \frac{1-0}{1} = 1$$
$$z_{y,1} = \frac{1-3}{2} = -1, \quad z_{y,2} = \frac{3-3}{2} = 0, \quad z_{y,3} = \frac{5-3}{2} = 1$$

Correlation coefficient:

$$r = \frac{1}{n-1}\sum z_x z_y = \frac{(-1)(-1)+(0)(0)+(1)(1)}{2} = \frac{2}{2} = 1$$

Slope:

$$b = r \cdot \frac{s_y}{s_x} = 1 \times \frac{2}{1} = 2$$

Intercept:

$$a = \bar{y} - b\bar{x} = 3 - 2 \cdot 0 = 3$$

$$\boxed{\hat{y} = 2x + 3}$$

Q7. An instructor develops a regression model to predict a student's final exam score (y, as a percentage 0–100) from the number of hours studied (x). Using data from students who studied between 0 and 15 hours, the instructor obtains the regression line:

$$\hat{y} = 55 + 3x$$

(a) (2 marks) A student studies for 8 hours. Predict their exam score using the regression model.

Solution: $\hat{y} = 55 + 3(8) = 55 + 24 = \boxed{79}$ percent

(b) (2 marks) Comment on the validity of the prediction from part (a).

Solution: The prediction is valid. The input $x = 8$ hours lies within the observed range of the data (0–15 hours), so the prediction is an interpolation rather than an extrapolation. No additional justification is required.

Comments on specific residuals are not relevant.

Bonus mark: Award one additional mark if the student **also** notes that the reliability of the prediction depends on the quality of the model fit (e.g., a high R^2 or well-behaved residual plot).

(c) (2 marks) Suppose a student who studied for 8 hours scored 99% on the exam, calculate their residual.

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

$$e = y - \hat{y} = 99 - 79 = 20$$

(d) (2 marks) Provide an interpretation of the slope of the regression line in this context.

Solution: For each additional hour studied, the predicted exam score increases by 3 percentage points, on average.

(e) (2 marks) If possible, provide an interpretation of the intercept. If it is impractical, explain why the intercept does not have a meaningful interpretation in this context.

Solution: The intercept is 55. It represents the predicted exam score for a student who studies $x = 0$ hours.

The intercept has practical meaning here: it represents the baseline score a student is expected to achieve without studying.

(f) (3 marks) The correlation between hours studied and exam score was $r = 0.75$. Determine and interpret the coefficient of determination R^2 in this context.

Solution: $R^2 = r^2 = 0.75^2 = 0.5625$

Interpretation: Approximately 56.25% of the **variation** in exam scores can be explained by the number of hours studied using the regression model.

Note, it is not inaccurate to state that it explains exam scores themselves, only the variation in exam scores.

(g) (2 marks) The correlation between hours studied and exam score was $r = 0.75$. If study time is converted from hours to minutes, what will the new correlation be? Explain.
Recall that 1 hour = 60 minutes.

Solution: The correlation remains $\boxed{r = 0.75}$. Correlation is unitless and invariant under linear transformations of either variable.

(h) (2 marks) Suppose the exam scores were converted from raw scores (0–100) to proportions (0–1) by dividing each raw score by 100. What would the slope of the new regression line be? Justify your answer.

Solution: The slope would be divided by 100:

$$b_{\text{new}} = \frac{b_{\text{old}}}{100} = \frac{3}{100} = \boxed{0.03}$$

This follows from the fact that the response variable has 1/100th of the standard deviation of the original variable, while the explanatory variable remains unchanged.

Section 6 **(19 marks)**

Q8. A study investigates the relationship between coffee consumption and exam performance among university students. Students were classified by whether they drank coffee on the morning of an exam (Yes/No) and whether they passed the exam (Pass/Fail). The results are summarized in the following contingency table:

	Exam outcome	
Group	Pass	Fail
Coffee	36	24
No Coffee	20	40

(a) (1 mark) Calculate the proportion of coffee drinkers among the students that passed the exam.

Solution:

$$\text{Risk}(\text{Coffee} \mid \text{Pass}) = \frac{36}{56} = \boxed{0.643}$$

(b) (1 mark) Calculate the odds of passing the exam for students who did not drink coffee.

Solution:

$$\text{Odds}(\text{Pass} \mid \text{No Coffee}) = \frac{20}{40} = \boxed{0.50}$$

(c) (3 marks) Write down the conditional distribution of exam outcome (Pass/Fail) among coffee drinkers.

Solution: Among coffee drinkers ($n = 60$):

$$P(\text{Pass} \mid \text{Coffee}) = \frac{36}{60} = 0.60 \quad (60\%)$$

$$P(\text{Fail} \mid \text{Coffee}) = \frac{24}{60} = 0.40 \quad (40\%)$$

Recall that:

	Exam outcome	
Group	Pass	Fail
Coffee	36	24
No Coffee	20	40

(d) (3 marks) Calculate the odds ratio comparing coffee drinkers to non-coffee drinkers for passing the exam. Interpret this value.

Solution:

$$\text{Odds}(\text{Pass} \mid \text{Coffee}) = \frac{36}{24} = \frac{3}{2} = 1.50$$
$$\text{Odds}(\text{Pass} \mid \text{No Coffee}) = \frac{20}{40} = \frac{1}{2} = 0.50$$

$$\text{OR} = \frac{\text{Odds}(\text{Pass} \mid \text{Coffee})}{\text{Odds}(\text{Pass} \mid \text{No Coffee})} = \frac{1.50}{0.50} = \boxed{3.00}$$

Interpretation: The odds of passing the exam for coffee drinkers are 3 times higher than the odds for non-coffee drinkers.

(e) (3 marks) Calculate the relative risk of passing the exam for coffee drinkers compared to non-coffee drinkers. Interpret this value.

Solution:

$$\text{Risk}(\text{Pass} \mid \text{Coffee}) = \frac{36}{60} = 0.60$$

$$\text{Risk}(\text{Pass} \mid \text{No Coffee}) = \frac{20}{60} = 0.\overline{3}$$

$$\text{RR} = \frac{\text{Risk}(\text{Pass} \mid \text{Coffee})}{\text{Risk}(\text{Pass} \mid \text{No Coffee})} = \frac{0.60}{0.\overline{3}} = \boxed{1.80}$$

Interpretation: Coffee drinkers are 1.8 times (or 80%) more likely to pass the exam compared to non-coffee drinkers.

There was an error in the data entry process, and the pass/fail entries in the table corresponding to the coffee drinkers group were actually incorrect.

	Exam outcome	
Group	Pass	Fail
Coffee	x	y
No Coffee	20	40

Let x be the number of coffee drinkers who passed and y be the number of coffee drinkers who failed.

(f) (4 marks) Suppose that coffee drinkers are twice as likely to pass the exam as non-coffee drinkers. Given that the total number of coffee drinkers remains 60, determine the values of x and y.

Solution: Let the probability of passing for non-coffee drinkers be p. Then, for coffee drinkers, it is $2p$. Since

$$p = \frac{20}{60} = \frac{1}{3},$$

it follows that

$$\text{Risk}(\text{Pass} \mid \text{Coffee}) = \frac{2}{3}$$

Therefore, since $x + y = 60$, it must be that

$$\frac{x}{x+y} = \frac{x}{60} = \frac{2}{3}$$

and solving for x gives

$$x = 60 \cdot \frac{2}{3} = \frac{120}{3} = 40.$$

Hence, $y = 60 - x = 60 - 40 = 20$. We conclude $\boxed{x = 40,\ y = 20}$.

(g) (4 marks) Given your findings from part (f), is there evidence of an association between coffee consumption and exam performance? Justify your answer.

Solution: Yes, there is evidence of an association between coffee consumption and exam performance.

This follows from the fact that coffee drinkers are twice as likely to pass the exam as non-coffee drinkers, which indicates a positive association between coffee consumption and passing the exam.

(There are other arguments one can make but the key point is that the probabilities of passing differ between the two groups, which indicates an association.)

Bonus Question **(3 bonus marks)**

BQ1. Suppose that x and y are two variables given by the following table.

x	y
-1	0
0	4
1	y_3

(a) (1 mark) Suppose the correlation is 1. What can we conclude about the relationship between x and y?

Solution: There is a perfect positive linear relationship between x and y. All data points lie exactly on a straight line with positive slope.

(b) (2 marks) Suppose the correlation is 1. Determine the possible values of y_3.

Solution: Since $r = 1$, all points must lie exactly on a straight line with positive slope.

Let the three points be $(-1, 0)$, $(0, 4)$, and $(1, y_3)$. The slope between the first two points is:

$$m = \frac{4 - 0}{0 - (-1)} = \frac{4}{1} = 4$$

So the line is $y = 4x + b$. Plug in $(0, 4)$:

$$4 = 4 \cdot 0 + b \implies b = 4$$

So the line is $y = 4x + 4$.

Plug in $x = 1$:

$$y_3 = 4 \cdot 1 + 4 = 8$$

The only possible value for y_3 is $\boxed{8}$.

Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.

Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.

Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.