Do NOT write in the area above this line.



Practice 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. The exam has 18 pages. The last three pages may be used for additional workspace or scrap paper.
5. When applicable, enter your final answer in the provided answer box rounded to 2 decimal places.

Section	**Marks**
Multiple Choice (1 mark each)	10
Miscellaneous	17
Chapter 1: Data Visualization	3
Chapter 2: Median, Quartiles, and Boxplots	20
Chapter 2: Mean and Standard Deviation	7
Chapter 4: Scatterplot and Correlation	15
Chapter 5: Regression	20
Chapter 6: Contingency Tables	8
Total	**100**

Multiple Choice (1 mark each) **(10 marks)**

Each of the following multiple-choice questions is worth 1 mark. Circle a single answer for each question. No justification is required and no partial marks will be awarded.

MC1. (1 mark) A dataset of employee commute times has a mean of 42 minutes and a median of 28 minutes. The shape of the distribution is most likely:

A. Bimodal

B. Symmetric

C. Left-skewed (negatively skewed)

D. **Right-skewed (positively skewed)**

E. Uniform

MC2. (1 mark) Which variable is best classified as categorical?

A. Weight (kg) **B.** Height (cm) **C.** **Blood type** **D.** Temperature (°C)

MC3. (1 mark) A dataset contains 23 observations sorted from smallest to largest. The median is the value in position:

A. 11 **B.** **12** **C.** 11.5 **D.** 13 **E.** 23

MC4. (1 mark) A dataset has $Q_1 = 30$ and IQR $= 20$. What is the lower fence for detecting outliers using the $1.5 \times$ IQR rule?

A. -10 **B.** **0** **C.** 10 **D.** 20 **E.** 30

MC5. (1 mark) Which of the following statements about the standard deviation is false?

A. It is always greater than or equal to zero.

B. It equals zero only when all observations are identical.

C. It is measured in the same units as the data.

D. **It is resistant to outliers.**

E. It measures spread around the mean.

MC6. (1 mark) If $r = 0.80$, what is R^2?

A. -0.80 **B.** -0.64 **C.** **0.64** **D.** 0.80 **E.** 0.40

MC7. (1 mark) A regression line is $\hat{y} = 120 - 3.5x$. The slope tells us that:

A. **For each 1-unit increase in x, predicted y decreases by 3.5.**

B. For each 1-unit increase in x, predicted y increases by 3.5.

C. For each 3.5-unit increase in x, predicted y decreases by 1.

D. The correlation is -3.5.

E. The intercept is -3.5.

MC8. (1 mark) Using a regression model to predict y for an x-value far beyond the range of observed data is called:

A. Interpolation

B. **Extrapolation**

C. Standardization

D. Confounding

E. Residual analysis

MC9. (1 mark) In a two-way table, a conditional distribution:

A. **Shows proportions within a specific row or column.**

B. Shows the raw counts for all cells.

C. Always sums to the total sample size.

D. Is the same as the marginal distribution.

E. Requires knowing the population size.

MC10. (1 mark) Given $\hat{y} = 15 + 2.5x$, what is $\hat{y}$ when $x = 8$?

A. **35** **B.** 20 **C.** 27.5 **D.** 37.5 **E.** 40

Miscellaneous **(17 marks)**

Q1. (2 marks) You have data on students' "Favourite Color" and "Favourite Season." Explain why a scatterplot would be inappropriate for displaying the relationship between these two variables.

Solution: Both of the variables are categorical variables. Scatterplots require two quantitative variables with meaningful numerical coordinates.

Q2. (2 marks) Sketch a time series plot that exhibits an upward trend with seasonal fluctuations. (No precise data points are needed; show the overall shape.)

Solution: Any time series which rises overall from left to right, with regular fluctuations repeating at consistent intervals.

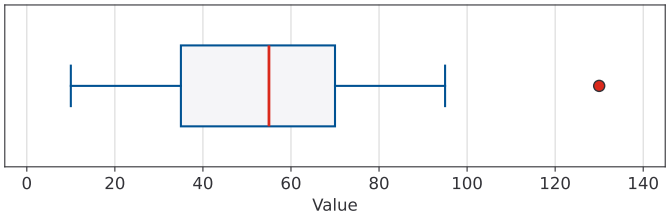
Q3. (1 mark) What does it mean for a statistic to be "resistant" (or "robust")?

Solution: A resistant statistic is not strongly affected by outliers.

Q4. (1 mark) Within the context of our course, what does the term **shape** of a distribution refer to?

Solution: It describes the symmetry of the distribution: whether it is symmetric, skewed left, or skewed right.

Q5. (5 marks) Consider the boxplot below.



Estimate each of the following from the plot: (i) the median, (ii) Q_1 and Q_3, (iii) the IQR, and (iv) the approximate value of any outlier(s).

Solution: From the plot:

(i) Median (thick center line) ≈ 55.

(ii) Q_1 (left edge of box) ≈ 35; Q_3 (right edge of box) ≈ 70.

(iii) IQR $= Q_3 - Q_1 \approx 70 - 35 = 35$.

(iv) Outlier (isolated dot) ≈ 130.

Q6. (2 marks) What is Simpson's paradox?

Solution: Simpson's paradox occurs when a trend that appears in separate groups reverses when the groups are combined.

This happens when a lurking variable confounds the relationship between the variables of interest.

Q7. (1 mark) What is an outlier?

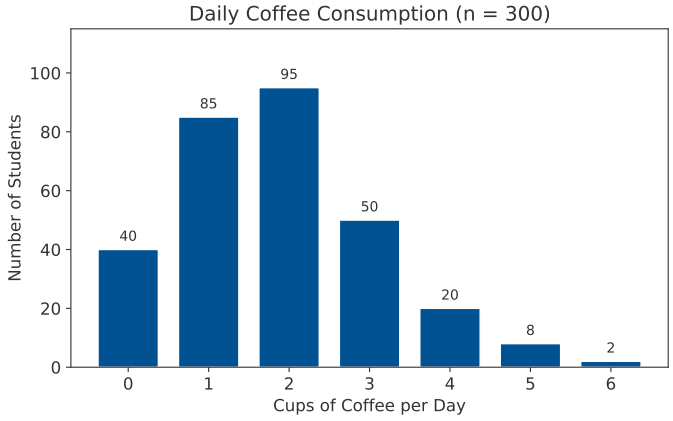
Solution: An outlier is an observation that lies far from the rest of the data.

Q8. (3 marks) Is it possible to visualize more than two variables in a scatterplot? If so, describe one method for doing this. If not, explain why not.

Solution: Yes. You can use colour to represent a third variable (e.g., different colours for different categories) or use point size to represent a fourth variable (e.g., larger points for higher values).

Chapter 1: Data Visualization **(3 marks)**

Q9. A campus coffee shop surveyed 300 university students about their daily coffee consumption. The bar chart below shows the results.



(a) (1 mark) What type of variable is "Cups of Coffee per Day"?

Solution: Quantitative

(b) (1 mark) Describe the shape of the distribution shown in the chart.

Solution: Right-skewed (positively skewed).

(c) (1 mark) The bar for "0 cups" has a height of 40 and the bar for "1 cup" has a height of 85. What proportion of the 300 students drink at most 1 cup of coffee per day? Round to 2 decimal places.

Solution:

$$\text{Proportion} = \frac{40 + 85}{300} = \frac{125}{300} \approx \boxed{0.42}$$

Chapter 2: Median, Quartiles, and Boxplots **(20 marks)**

Q10. You are analyzing the daily commute times (in minutes) for 10 employees at a tech company. The sorted data are:

5	12	15	22	25	30	35	42	55	105

(a) (7 marks) Calculate the five-number summary (Min, Q_1, Median, Q_3, Max) for this data. Show your work.

Solution: Five-number summary: 5, 15, 27.5, 42, 105

(b) (5 marks) Determine whether there are any outliers.

Solution:

105 is outside the inner fence (-25.5, 82.5), hence it is an outlier.

(c) (4 marks) Draw a modified boxplot for this data.

Solution:

(d) (4 marks) Which statistics should be used to describe the center and spread of this data? Explain your reasoning.

Solution: Use the median and IQR. The distribution is right-skewed with an outlier (105), so the mean and standard deviation would be pulled by the extreme value. The median and IQR are resistant to outliers.

Chapter 2: Mean and Standard Deviation **(7 marks)**

Q11. A researcher records the daily screen time (in hours) for 5 teenagers:

3	5	6	8	13

(a) (4 marks) Calculate the sample mean $\bar{x}$ and the sample standard deviation s.

Solution: Mean:

$$\bar{x} = 7$$

Squared deviations:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 58.$$

Therefore, the variance is

$$s^2 = \frac{58}{5-1} = \frac{58}{4} = 14.5$$

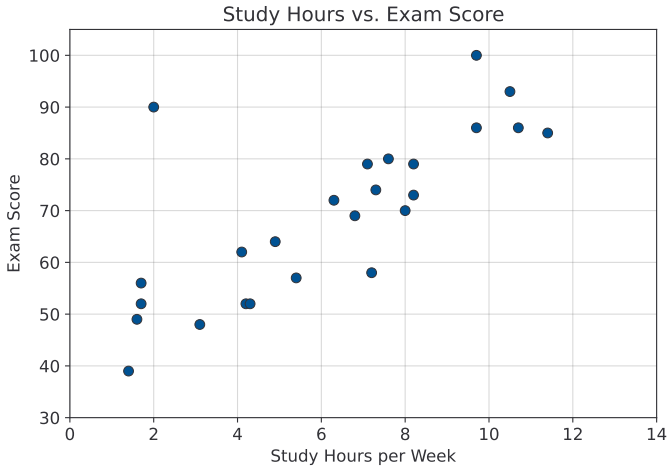
And the standard deviation is $s = \sqrt{14.5} \approx \boxed{3.81 \text{ hours}}$

(b) (3 marks) Each teenager reduces their screen time by exactly 1 hour per day. Find the new variance of daily screen time among the teenagers.

Solution: Subtracting a constant shifts the data but does not change the spread. The variance remains $s^2 = \boxed{14.5}$.

Chapter 4: Scatterplot and Correlation **(15 marks)**

Q12. A study examines the relationship between weekly study hours (x) and exam scores (y) for 25 university students.



The correlation coefficient for this data is $r = 0.72$.

(a) (3 marks) Describe the direction, form, and strength of the association shown in the scatterplot.

Solution: The association is positive (points trend upward), linear in form, and (moderately) strong ($r = 0.72$).

(b) (1 mark) One student studied only 2 hours per week but scored 90 on the exam. If this point were removed, would the correlation r likely increase (move closer to 1) or decrease (move closer to 0)? Explain briefly.

Solution: The correlation would likely increase (move closer to 1). This point contradicts the positive trend (low x, high y), so removing it strengthens the linear pattern.

(c) (3 marks) Calculate R^2 and interpret it in the context of this study.

Solution:

$$R^2 = (0.72)^2 = \boxed{0.5184 \approx 51.84\%}$$

About 51.84% of the variation in exam scores is explained by the linear relationship with weekly study hours.

(d) (3 marks) Suppose every student's exam score is increased by 5 points (a curved exam). What would happen to the correlation r? Explain.

Solution: r would remain unchanged at 0.72. Adding a constant to every y-value shifts the data vertically but does not change the linear association or the spread. Correlation is invariant under translation.

(e) (5 marks) The table below shows weekly study hours (x) and exam scores (y) for a sample of four students from a different class. Calculate the correlation coefficient r for these students. Compare the strength of the linear relationship between these points to that exhibited in the plot in part (a).

Student	x	y
A	2	50
B	4	60
C	6	70
D	8	65

Solution: Check that the correlation is approximately $r = 0.83$.

Since the correlation for the 4 students is larger than the correlation for the 25 students ($r = 0.72$), the linear relationship between study hours and exam scores is stronger for this smaller sample.

Chapter 5: Regression **(20 marks)**

Q13. A researcher is studying the relationship between hours of sleep per night (x) and reaction time in milliseconds (y) for university students. Summary statistics:

$$\bar{x} = 6, \quad \bar{y} = 260, \quad s_x = 2, \quad s_y = 50, \quad r = -0.60.$$

(a) (8 marks) Calculate the slope b and intercept a of the least-squares regression line $\hat{y} = a + bx$. Show your work.

Solution: Slope:

$$b = r \cdot \frac{s_y}{s_x} = -15$$

Intercept:

$$a = \bar{y} - b\bar{x} = 260 - (-15)(6) = 260 + 90 = 350$$

$$\boxed{\hat{y} = 350 - 15x}$$

(b) (3 marks) Interpret the slope in the context of sleep and reaction time.

Solution: For each additional hour of sleep, predicted reaction time decreases by 15 milliseconds.

Q14. An urban planner models monthly rent (y, in dollars) from apartment size (x, in square feet).

Model: $\hat{y} = 200 + 1.5x$. Data range: apartment sizes from 400 to 1600 sq ft. $R^2 = 0.72$.

(a) (3 marks) Predict the monthly rent for an apartment that is 1000 square feet.

Solution:

$$\hat{y} = 200 + 1.5(1000) = 200 + 1500 = \boxed{\$1700}$$

(b) (3 marks) A tenant pays \$1900 per month for a 1000 sq ft apartment. Calculate the residual for this observation and interpret it.

Solution: Predicted: $\hat{y} = 1700$ (from part (a)).

$$\text{Residual} = y - \hat{y} = 1900 - 1700 = \boxed{\$200}$$

The tenant pays \$200 more than the model predicts for this apartment.

(c) (3 marks) Provide an interpretation of the intercept in the context of this problem, if it has practical meaning. If not, explain why.

Solution: The intercept predicts a rent of \$200 for a 0 sq ft apartment. This has no practical meaning because no apartment can have zero size.

Chapter 6: Contingency Tables **(8 marks)**

Q15. A university wellness study examined the relationship between exercise habits and stress levels among 100 students.

	Low Stress	High Stress	Total
Regular Exercise	48	12	60
No Regular Exercise	16	24	40
Total	64	36	100

(a) (1 mark) Calculate the risk (probability) of low stress given that a student exercises regularly.

Solution:

$$P(\text{Low Stress} \mid \text{Regular Exercise}) = \frac{48}{60} = \boxed{0.80}$$

(b) (1 mark) Calculate the odds of low stress for those who do not exercise regularly.

Solution:

$$\text{Odds} = \frac{16}{24} \approx \boxed{0.67}$$

(c) (2 marks) Are regular exercise and low stress independent? Explain your reasoning.

Solution: $P(\text{Low Stress} \mid \text{Regular Exercise}) = 48/60 = 0.80$.

$P(\text{Low Stress} \mid \text{No Regular Exercise}) = 16/40 = 0.40$.

These conditional probabilities differ ($0.80 \neq 0.40$), so the variables are not independent. Regular exercise is associated with lower stress.

(d) (2 marks) Calculate the relative risk of low stress for those who exercise regularly compared to those who do not. Interpret the result.

Solution:

$$\text{RR} = \frac{P(\text{Low Stress} \mid \text{Regular})}{P(\text{Low Stress} \mid \text{No Regular})} = \frac{0.80}{0.40} = \boxed{2.0}$$

Students who exercise regularly are 2 times as likely to have low stress compared to those who do not.

Another solution:
Students who exercise regularly are 100% more likely to have low stress compared to those who do not.

(e) (2 marks) Calculate the odds ratio of having low stress for those who exercise regularly compared to those who do not. Interpret the result.

Solution: $\text{Odds}(\text{Low Stress} \mid \text{Regular}) = \frac{48}{12} = 4.0$.
$\text{Odds}(\text{Low Stress} \mid \text{No Regular}) = \frac{16}{24} \approx 0.667$.

$$\text{OR} = \frac{4.0}{0.667} = \boxed{6.0}$$

The odds of low stress are 6 times higher for students who exercise regularly compared to those who do not.

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.