Diagnostic Assessment – Unit 02

Statistics and Data Analysis

(Descriptive Statistics; designed to identify learning gaps)

Instructor: Tofik Ali **Time Limit:** None
Submission Deadline: March 31, 2026
Student Name: _____ **Roll No.:** _____

Instructions

- **This is a diagnostic assessment, not a quiz/test.** The goal is to identify learning gaps so you know what to revise next.
- **Total marks: 100.** Easy: $10 \times 2 = 20$, Medium: $10 \times 3 = 30$, Hard: $10 \times 5 = 50$.
- **No time limit.** Submit on or before **March 31, 2026**.
- **Important (70% rule): To have your assignment marks counted, you must score at least 70% (70/100 or more) in this assessment.** If you score below 70%, you may reattempt and resubmit. Your latest submission will be considered.
- **Marking policy (honest attempt):** Marks are deducted only for (i) not attempting a question and (ii) high similarity with other submissions (copying). Correctness is **not** the main focus; your reasoning and effort matter more.
- **Show your real effort:** if you tried multiple approaches for any question, write them all (Attempt 1, Attempt 2, ...). Partial/incorrect attempts are acceptable.
- **Many questions are open-ended by design:** multiple solutions/variants are acceptable. Use your own example data and explain your choices.
- Unless specified, you may choose **sample** or **population** formulas, but you must mention your choice.
- Round final numeric answers to **2 decimal places**.

Easy (10 Questions)

E1. [2 marks] [Subjective] Create a dataset of **8 numbers** using your roll number digits (you may repeat digits). Compute mean and median. Write 2–3 lines: which measure is more robust to outliers and why?

E2. [2 marks] [MCQ – Select all that apply] Which measures are generally **more robust to outliers**?

(A) Mean

(B) Median

(C) Mode

(D) 10% trimmed mean

(E) Standard deviation

E3. [2 marks] [Subjective] For the dataset $x = \{5, 6, 6, 7, 9, 10\}$ compute mean, median, and mode.

E4. [2 marks] [Subjective] Explain the difference between **population variance** and **sample variance**. Why do we use $(n-1)$ in the sample variance formula? (4–6 lines)

E5. [2 marks] [Subjective] Compute the **range** and **IQR** for $x = \{2, 3, 3, 6, 7, 9, 10, 12\}$. State your quartile method.

E6. [2 marks] [Subjective] A dataset has mean = 60 and standard deviation = 8. Compute the z-score for $x = 70$ and interpret in one line.

E7. [2 marks] [MCQ – Select all that apply] Which statements are **true**?

(A) Covariance is scale dependent.

(B) Correlation is always between -1 and $+1$.

(C) Correlation changes if we convert height from cm to meters.

(D) A correlation of 0 means there is no relationship at all.

(E) Negative correlation means as x increases, y tends to decrease (linearly).

E8. [2 marks] [Subjective] Write 4–6 lines explaining **"correlation does not imply causation"** and give one example.

E9. [2 marks] [Subjective] Identify the likely skewness type (right-skewed / left-skewed / symmetric) and justify briefly:

(a) Salaries in a large company

(b) Marks in a very easy exam (most score high)

(c) Adult heights

E10. [2 marks] [Subjective] In 4–6 lines, explain what **kurtosis** tells you. Mention one common misunderstanding about kurtosis.

Medium (10 Questions)

M1. [3 marks] [Subjective] Compute variance and standard deviation for $x = \{4, 6, 8, 10, 12\}$. Use either sample or population formula (mention your choice).

M2. [3 marks] [Subjective] Given paired data $(x, y) = \{(1, 2), (2, 3), (3, 5), (4, 4), (5, 6)\}$: compute covariance and Pearson correlation and interpret the sign and strength (3–5 lines).

M3. [3 marks] [Subjective] Create two datasets A and B of size 8 such that: $\text{mean}(A) = \text{mean}(B)$ but $\text{std}(A) > \text{std}(B)$. Show your datasets and verify using calculations (many answers possible).

M4. [3 marks] [Subjective] For $x = \{4, 5, 7, 8, 9, 10, 12, 13, 14, 30\}$ compute the five-number summary. Identify outliers using the $1.5\times$IQR rule.

M5. [3 marks] [Subjective] Explain (with a short numeric example) why covariance is scale dependent but correlation is not. (Example: $x' = 10x$, y unchanged.)

M6. [3 marks] [Subjective] Create a small dataset (at least 12 points) where correlation is **positive** overall, but one outlier can strongly change the value. Explain in 4–6 lines what happened.

M7. [3 marks] [Subjective] Using Python (recommended), generate a summary table (count, mean, std, min, quartiles, max) for a dataset you choose (at least 30 numbers). Provide code and write 4–6 lines interpreting the results.

M8. [3 marks] [Subjective] A feature has skewness = +1.1 and kurtosis = 5.0. Interpret what this suggests about the distribution shape and tails (6–8 lines).

M9. [3 marks] [Subjective] Explain in 6–8 lines: when can the mean be misleading? Provide one example dataset and explain which summary you would report instead (median/IQR, trimmed mean, etc.).

M10. [3 marks] [Subjective] Choose an appropriate plot for each task and justify (2–3 lines each):

(a) Compare distributions of marks for two sections

(b) Detect outliers in delivery time data

(c) Check linear relationship between study time and marks

Hard (10 Questions)

H1. [5 marks] [Subjective] Robust reporting: for $x = \{12, 13, 13, 14, 15, 16, 16, 17, 18, 120\}$ compute: mean, median, IQR, and a 10% trimmed mean. Write a 8–12 line paragraph explaining what you would report and why.

H2. [5 marks] [Subjective] Construct a dataset (size ≥ 20) that is clearly right-skewed. Justify using mean vs median and at least one plot (histogram or boxplot). Provide your code or calculations.

H3. [5 marks] [Subjective] Build a correlation matrix for three variables you design (at least 12 observations). Explain in 6–10 lines what the matrix tells you, and mention one limitation of correlation.

H4. [5 marks] [Subjective] Nonlinear relationship challenge: create a dataset (at least 10 points) where x and y have a strong nonlinear relationship but Pearson correlation is near 0. Explain why and show the plot that reveals the pattern.

H5. [5 marks] [Subjective] Confounding variable: Give a realistic example where correlation is high but not causal. Identify at least one confounder and describe how you would test/validate the relationship (8–12 lines).

H6. [5 marks] [Subjective] Dimensional summaries: Suppose you have a dataset with 8 numeric features and 500 rows. List at least 10 per-feature statistics you would compute. Mark which ones help detect outliers and non-normality.

H7. [5 marks] [Subjective] Grouped summaries: Design a summary table for exam marks grouped by branch and gender. Mention 3 pitfalls (e.g., small group size, outliers, imbalance) and how your summary handles them.

H8. [5 marks] [Subjective] Write a short "data story" (10–14 lines) for a stakeholder using a dataset you choose. Include: one central tendency measure, one dispersion measure, one shape measure (skewness or kurtosis), and one plot. Keep it clear and non-technical.

H9. [5 marks] [Subjective] Compare two datasets that have the same mean and standard deviation but differ in distribution shape. Construct the datasets, verify the statistics, and explain (6–10 lines) what the summaries miss.

H10. [5 marks] [Subjective] Mini case study: You are analyzing delivery times (minutes) for an online service. Write an analysis plan (10–14 steps) using Unit 02 measures and visualizations. Include how you will treat outliers and how you will communicate results.

End of Assessment