# Assignment – Unit 02

Statistics and Data Analysis

(Central Tendency, Dispersion, Covariance, Correlation, Skewness, Kurtosis, Summaries, Distributions)

**Instructor:** Tofik Ali **Submission Deadline:** March 31, 2026
**Student Name:** _____ **Roll No.:** _____

---

**Instructions**

- Answer **all** questions. Show calculations and write assumptions.
- There is **no time limit**. Submit on or before **March 31, 2026**.
- Unless specified, you may choose **sample** or **population** formulas, but you must clearly mention your choice.
- Round final numeric answers to **2 decimal places**.
- You may use a calculator or Python for verification, but show key steps/formulas in your solution.

## Easy (10 Questions)

E1. **Mean/Median/Mode:** For the data

$$x = \{7, 9, 9, 10, 12, 13\}$$

compute mean, median, and mode. Which measure is most representative here, and why (2–3 lines)?

E2. **Effect of an outlier:** Consider

$$x = \{10, 11, 12, 13, 14\}$$

Compute the mean and median. Now add an outlier value 100 and recompute mean and median. Write one line: which measure changed more and why?

E3. **Trimmed mean:** Compute the 20% trimmed mean for:

$$x = \{2, 3, 4, 7, 9, 10, 12, 50, 60, 80\}$$

(Sort first. Remove equal number of values from both ends.)

E4. **Range and IQR:** For the data

$$x = \{3, 5, 7, 8, 8, 10, 12, 15\}$$

compute range, $Q_1$, $Q_3$, and IQR. State the method you used for quartiles.

E5. **Variance and standard deviation:** For

$$x = \{4, 6, 8, 10\}$$

compute variance and standard deviation (sample or population, mention your choice).

E6. **Z-scores:** Using your mean and standard deviation from E5, compute the z-score of $x = 10$. Interpret the z-score in one line.

E7. **Covariance sign:** Without calculating exact numbers, predict whether covariance is positive, negative, or near zero for each pair, and justify in 1–2 lines:

(a) Height and weight

(b) Price and demand (in general)

(c) Shoe size and exam marks

E8. **Correlation vs causation:** Write 4–6 lines explaining why "correlation does not imply causation", and give one real-world example.

E9. **Skewness (conceptual):** For each dataset, identify whether it is likely **right-skewed**, **left-skewed**, or **approximately symmetric**. Justify briefly.

(a) Salaries in a large company

(b) Marks in a very easy test

(c) Heights of adult humans

E10. **Kurtosis (conceptual):** In 4–6 lines, explain what **high kurtosis** indicates about a distribution. Mention one common misunderstanding about kurtosis.

## Medium (10 Questions)

M1. **Five-number summary + outliers:** For the dataset:

$$x = \{4, 5, 7, 8, 9, 10, 12, 13, 14, 30\}$$

compute the five-number summary and identify outliers using the 1.5×IQR rule.

M2. **Compare two groups:** Section A marks:

$$A = \{55, 60, 62, 65, 70, 72, 75\}$$

Section B marks:

$$B = \{45, 50, 58, 60, 68, 80, 90\}$$

Compute mean and standard deviation for both and write 4–6 lines comparing central tendency and dispersion.

M3. **Covariance and correlation:** Given paired data:

$$(x, y) = \{(1, 2), (2, 3), (3, 5), (4, 4), (5, 6)\}$$

compute covariance and Pearson correlation. Interpret the result in 3–5 lines.

M4. **Scale dependence:** Show (with a short calculation) how covariance changes if we transform $x' = 10x$ while keeping $y$ same. Use the dataset from M3.

M5. **Correlation matrix (small):** Consider the dataset below (5 observations):

$$x = \{1, 2, 3, 4, 5\}, \quad y = \{2, 4, 6, 8, 10\}, \quad z = \{5, 4, 3, 2, 1\}$$

Compute the correlation matrix among $(x, y, z)$ and interpret the signs/magnitudes.

M6. **Interpret skewness/kurtosis values:** A feature has skewness = +1.2 and kurtosis = 4.8. Write 5–8 lines interpreting what this suggests about the distribution shape and tails.

M7. **Summary table (by hand):** For the data:

$$x = \{2, 4, 4, 5, 7, 9, 10, 10\}$$

create a summary table containing: count, mean, std, min, $Q_1$, median, $Q_3$, max. (You may compute quartiles using your chosen method; state it.)

M8. **Python summary (recommended):** Using Python (pandas), compute `describe()` for the same dataset in M7. Write the code and paste the output table (or write the key values). Mention one value that surprised you.

M9. **When summaries mislead:** Give an example (you may invent small data) where two datasets have the same mean and std but look very different. Explain in 5–8 lines what statistics miss and what plot you would use.

M10. **Choice of measure:** For each scenario, choose an appropriate measure of central tendency and dispersion and justify (3–4 lines each):

(a) Monthly income data

(b) Daily temperature in a city

(c) Delivery times for an online service (with occasional delays)

## Hard (10 Questions)

H1. **Robust summary report:** You are given the dataset:

$$x = \{12, 13, 13, 14, 15, 16, 16, 17, 18, 120\}$$

Compute mean, median, IQR, and a 10% trimmed mean. Write a 6–10 line "data summary" paragraph explaining what you would report and why.

H2. **Design a dataset:** Construct two datasets $A$ and $B$ of size 8 such that:

- $\text{mean}(A) = \text{mean}(B)$
- $\text{std}(A) > \text{std}(B)$

Show your datasets and verify the conditions with calculations.

H3. **Covariance matrix:** For the 3-variable dataset (6 observations) below:

$$(x, y, z) = \{(1, 2, 3), (2, 1, 4), (3, 3, 2), (4, 5, 1), (5, 4, 2), (6, 6, 0)\}$$

compute the sample covariance matrix. Identify which pair has the strongest linear relationship.

H4. **Correlation pitfalls:** Give a real or realistic example where correlation is high but the relationship is **not** causal. Explain a confounding variable in 6–10 lines.

H5. **Nonlinear relationship:** Create a small dataset (at least 8 points) where $x$ and $y$ have a strong nonlinear relationship but Pearson correlation is near zero. Explain why this happens and what plot reveals it.

H6. **Dimensional summaries:** Suppose you have a dataset with 8 features (columns) and 500 rows. List the key summary statistics you would compute per feature (at least 10 items). Mention which ones you would use to detect outliers and non-normality.

H7. **Grouped summaries:** You have exam marks for two branches (CSE and AI) and two genders. Design a table (columns) that summarizes the data fairly and avoids misleading conclusions. Mention at least 3 pitfalls and how your table avoids them.

H8. **Skewness vs median/mean:** Explain the relationship between skewness and the relative positions of mean and median. Provide two small example datasets: one right-skewed and one left-skewed, and verify mean vs median.

H9. **Kurtosis comparison:** Two distributions have the same mean and variance but different kurtosis. In 8–12 lines, explain what might differ in their shape and why it matters for outliers/risk.

H10. **Mini case study:** You are analyzing delivery times (in minutes) for an online service. Suggest a complete analysis plan (8–12 steps) using Unit 02 measures and visualizations. Include which statistics you will compute, how you will treat outliers, and how you will communicate results.

*End of Assignment*