# Statistics and Data Analysis
## Unit 02 – Lecture 04 Notes
## Statistical Summaries for Data

Tofik Ali

February 9, 2026

## What You Will Learn (Beginner-Friendly)

In many real problems, the dataset is too large to read row-by-row. So we ask:

1. What is a "typical" value of a feature (center)?

2. How much do values vary (spread)?

3. How can we compare groups (CSE vs ECE) or time (week 1 vs week 2)?

By the end of this lecture, you should be able to:

- interpret a common summary table (count, mean, std, quartiles, min/max),

- compute and interpret the five-number summary,

- produce and interpret grouped summaries,

- and explain what information is lost when we compress data into a few numbers.

## 1. Why Summaries Are Needed

If a dataset has 10,000 rows, you cannot communicate it in a report by printing the raw table. A **summary** compresses the data into a small set of numbers that still capture the important story.

### 1.1 "Comparison across groups or time"

Summaries let us compare:

- **groups**: e.g., mean/median final score in CSE vs ECE,

- **time periods**: e.g., average weekly sales in January vs February.

Instead of comparing 10,000 raw values, we compare a few summary values.

## 2. Standard Summary Table (What Each Column Means)

For a numeric variable (say `final_score`), a typical summary contains:

- **count** ($n$): how many values exist (after excluding missing values),

- **mean**: arithmetic average (sensitive to outliers),

- **std**: sample standard deviation (typical distance from mean),

- **min/max**: extremes,

- **25%**, **50%**, **75%**: quartiles ($Q_1$, median, $Q_3$).

**Important warning.** A summary table does **not** show the full distribution shape. Two datasets can have similar mean/std but look very different (skewed vs bimodal).

## 3. Five-Number Summary

The five-number summary is:

$$\min(x), \ Q_1, \ \text{median}, \ Q_3, \ \max(x)$$

It is used to create a **boxplot** and is often more robust than mean/std.

### 3.1 Quartile interpretation

- $Q_1$ (25th percentile): about 25% of values are at or below $Q_1$.

- Median (50th percentile): about 50% of values are at or below the median.

- $Q_3$ (75th percentile): about 75% of values are at or below $Q_3$.

So, the middle 50% of the data lies between $Q_1$ and $Q_3$.

### Exercise 1 (solution)

Dataset: 4, 5, 7, 8, 9, 10, 25
Five-number summary:

- $\min = 4, \ \max = 25$

- median $= 8$

- lower half (4, 5, 7) $\Rightarrow Q_1 = 5$

- upper half (9, 10, 25) $\Rightarrow Q_3 = 10$

- IQR $= Q_3 - Q_1 = 5$

# 4. Mean vs Median (Quick Skewness Clue)

Mean and median are both measures of center, but:

- mean uses all values and is pulled by outliers,
- median uses only ordering and is robust to outliers.

**Rule of thumb (not a proof).**

- mean ≈ median: might be roughly symmetric,
- mean > median: often right-skewed,
- mean < median: often left-skewed.

Always confirm with a plot.

## Exercise 2 (solution)

Given summary:

- 25% = 65 means about 25% scored 65 or less.
- 75% = 82 means about 75% scored 82 or less.
- std = 12 means a typical score is roughly 12 points away from the mean (spread).

## Exercise 3 (solution)

Group A: 60, 62, 65, 95
Group B: 70, 72, 73, 74

- Group A mean = 70.5; median = 63.5
- Group B mean = 72.25; median = 72.5

Interpretation: Group A has an outlier (95) that inflates its mean. Typical performance (median) is much higher in Group B.

# 5. Grouped Summaries (Stratification)

Sometimes one global summary is misleading. We compute summaries **within groups** (by program, section, gender, etc.).

## 5.1 Weighted mean (why it matters)

If groups have different sizes, the overall mean must weight by group size.

**Exercise 4 (solution)**

Section A: $n_A = 10$, mean=70; Section B: $n_B = 5$, mean=80

$$\bar{x} = \frac{70 \cdot 10 + 80 \cdot 5}{15} \approx 73.33$$

Simple average of means (75) is incorrect here.

**Exercise 5 (solution)**

Means:

- CSE: 72.5
- ECE: 62.5
- AIML: 82.5

**Exercise 6 (solution)**

When we only report mean and std, we can miss:

- outliers and skewness,
- multi-modality (two peaks),
- differences between subgroups.

## 6. Mini Demo (Python)

Run from the lecture folder:

`python demo/statistical_summaries_demo.py`

It uses `data/student_summary.csv` and prints:

- an overall summary per numeric column,
- a grouped summary of `final_score` by `program`.

It also saves:

- `data/overall_summary.csv`
- `data/summary_by_program.csv`
- `images/mean_final_by_program.png` (if matplotlib is installed)

## References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley, 7th ed., 2020.
- McKinney, W. *Python for Data Analysis*, O'Reilly, 2022.
- Freedman, D., Pisani, R., & Purves, R. *Statistics*, W. W. Norton, 4th ed., 2007.