Statistics and Data Analysis
Unit 02 – Lecture 05 Notes
Dimensional Summaries and Distributions

Tofik Ali

February 9, 2026

What You Will Learn (Beginner-Friendly)

In earlier lectures we learned measures of center (mean/median/mode) and spread (IQR, variance, std). In this lecture we scale up that idea:

- A dataset usually has many columns (dimensions/features).
- Each feature can have a different distribution shape.
- We need per-feature (dimensional) summaries and distribution thinking.

By the end, you should be able to:

- compute and interpret per-feature summaries,
- recognize common distribution shapes (symmetric, skewed, bimodal),
- explain why shape matters for choosing the right summary statistic.

1. Dimensional (Per-Feature) Summaries

1.1 Definition

A **dimensional summary** means summarizing each feature/column separately using:

- center: mean/median,
- spread: std/IQR,
- range: min/max,
- quartiles: Q_1, Q_3.

Why it helps. If you have 20 columns, you can quickly identify:

- which features have large variability,
- which features have outliers,
- which features are likely skewed,
- which features might need transformation (like log).

Exercise 1 (solution)

Given:

- A: mean=50, median=50 $\Rightarrow$ roughly symmetric (likely)
- B: mean=80, median=60 $\Rightarrow$ right-skewed (high values pull mean upward)
- C: mean=60, median=75 $\Rightarrow$ left-skewed (low values pull mean downward)

This rule is a **heuristic**. Always confirm using a histogram or boxplot.

2. Distribution Shapes

2.1 Symmetric distributions

For symmetric distributions (often approximately normal):

- mean $\approx$ median,
- left and right tails are similar,
- mean and std are often reasonable summaries.

2.2 Right-skewed distributions

Right-skewed means there is a long tail on the right. Example: income. Most people have moderate incomes, but a few people have very high incomes. This pulls the mean upward, so mean $>$ median is common.

2.3 Left-skewed distributions

Left-skewed means there is a long tail on the left (a few very low values). Example: marks on an easy exam where many students score very high. Mean $<$ median can occur.

2.4 Bimodal distributions

Bimodal means two peaks. This often happens when the data mixes two sub-populations. Example: commute times might be short for hostel students and long for day scholars.

Exercise 2 (solution)

Commute times: 10, 12, 15, 18, 20, 60, 65, 70
Mean:

$$\frac{270}{8} = 33.75$$

Median:

$$\frac{18+20}{2} = 19$$

Interpretation: the mean is not typical because the data has two clusters and very few values around 34.

Exercise 3 (solution)

Daily income is most likely right-skewed.

3. Outliers and Robust Summaries

3.1 Outliers

Outliers are values that are unusually far from the rest. They can be:

- errors (wrong entry, sensor fault),
- or true extremes (rare but real cases).

So we should detect them and think, not blindly delete them.

3.2 IQR rule (recap)

Compute:

$$\text{IQR} = Q_3 - Q_1$$

Fences:

$$Q_1 - 1.5\text{IQR}, \quad Q_3 + 1.5\text{IQR}$$

Values outside fences are flagged as potential outliers.

Exercise 4 (solution)

Dataset: 10, 12, 13, 14, 15, 16, 40
Median = 14; $Q_1 = 12$; $Q_3 = 16$; IQR = 4
Upper fence = 16 + 1.5(4) = 22
So 40 is an outlier by the IQR rule.

Exercise 5 (solution)

For income (right-skewed), median + IQR is usually better than mean + std because it is robust.

Exercise 6 (solution)

Mean(hours) = 4 and mean(score) = 60.

4. Mini Demo (Python)

Run from the lecture folder:

```
python demo/dimensional_summaries_distributions_demo.py
```

It uses `data/multi_feature_distributions.csv` and prints a dimensional summary:

- mean, median, std, min/max, quartiles, and a simple skewness estimate.

If matplotlib is installed, it also saves `images/hists_grid.png`.

References

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