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

Tofik Ali

February 17, 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.

## Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

### Title Slide

`https://github.com/tali7c/Statistics-and-Data-Analysis`

## Quick Links

**Dimensional Summary** **Distribution Shape** **Outliers** **Demo** **Summary**

## Agenda

- Dimensional Summaries
- Distribution Shapes
- Outliers and Robust Summaries
- Demo
- Summary

## Learning Outcomes

- Explain what a dimensional (per-feature) summary is
- Use mean/median/quartiles to get a quick idea of distribution shape
- Recognize common distribution shapes (symmetric, skewed, bimodal)
- Explain why distribution shape matters for interpretation and method choice

## What is a Dimensional Summary?

A **dimensional summary** reports statistics *for each feature*:

- One dataset can have many columns (income, commute, sleep, score)
- We summarize each column separately: center + spread + quartiles
- This helps us quickly spot features that behave differently

## Exercise 1: Mean vs Median (Shape Clue)

Three features (summary only):

| Feature | Mean | Median |
|---|---:|---:|
| A | 50 | 50 |
| B | 80 | 60 |
| C | 60 | 75 |

**Task:** Which looks symmetric? Which is right-skewed? Which is left-skewed?

## Solution 1

- A: mean $\approx$ median $\Rightarrow$ roughly symmetric (likely)
- B: mean $>$ median $\Rightarrow$ right-skewed (high values pull mean up)
- C: mean $<$ median $\Rightarrow$ left-skewed (low values pull mean down)

**Reminder:** this is a clue, not a guarantee. Confirm with a plot.

## Common Distribution Shapes

- **Symmetric (approximately normal)**: mean $\approx$ median; bell-like
- **Right-skewed**: long tail to the right; mean $>$ median (often income)
- **Left-skewed**: long tail to the left; mean $<$ median (often scores near 100)
- **Bimodal**: two peaks (two sub-populations mixed together)

## Exercise 2: Bimodal Example (Mean Can Be Misleading)

Commute times (minutes):

10 12 15 18 20 60 65 70

**Task:** Compute mean and median. Is the mean a "typical" commute time here?

## Solution 2

Sorted data: 10, 12, 15, 18, 20, 60, 65, 70

- mean $= 270/8 = 33.75$
- median $= (18 + 20)/2 = 19$

**Interpretation:** the mean (33.75) is not typical because there are two clusters (short commuters and long commuters). The "middle" has almost no data.

## Exercise 3: Identify the Shape (Quick Reasoning)

**Question:** Which scenario is most likely right-skewed?

1. Heights of students
2. Daily income of individuals
3. Measurement error around zero

## Solution 3

**Daily income** is most likely right-skewed: most values are moderate, with a small number of very high values (long right tail).

## Outliers and Robustness

- Outliers can strongly affect mean and standard deviation
- Median and IQR are more robust (less sensitive to extremes)
- Always ask: error or true extreme?

## Exercise 4: IQR Outlier Check

Dataset:

10 12 13 14 15 16 40

**Task:** Compute $Q_1$, $Q_3$, IQR, and check if 40 is an outlier using the IQR fences.

## Solution 4

Sorted: 10, 12, 13, 14, 15, 16, 40
median = 14; lower half (10,12,13) $\Rightarrow Q_1 = 12$; upper half (15,16,40) $\Rightarrow Q_3 = 16$

- IQR $= 16 - 12 = 4$
- Upper fence $= Q_3 + 1.5 \cdot IQR = 16 + 6 = 22$
- Since $40 > 22$, 40 is an outlier by the IQR rule.

## Robust Options (When Skew/Outliers Exist)

- Report median and IQR instead of mean and std
- Use trimmed mean (remove small % of extremes)
- Transform the feature (e.g., $\log(1 + x)$ for right-skewed positive values)

## Exercise 5: Which Summary Would You Report?

**Question:** For income data (right-skewed), which pair is usually better?

1. mean + standard deviation
2. median + IQR

## Solution 5

**Median + IQR** is usually better for right-skewed income: it represents the typical person and is less distorted by a few extremely high incomes.

## Exercise 6: Dimensional Summary (Tiny Table)

| Student | hours | score |
|---|---|---|
| 1 | 2 | 50 |
| 2 | 4 | 60 |
| 3 | 6 | 70 |

**Task:** Compute mean(hours) and mean(score). This is a 2-feature dimensional summary.

## Solution 6

- mean(hours) $= (2 + 4 + 6)/3 = 4$
- mean(score) $= (50 + 60 + 70)/3 = 60$

## Mini Demo (Python)

Run from the lecture folder:

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

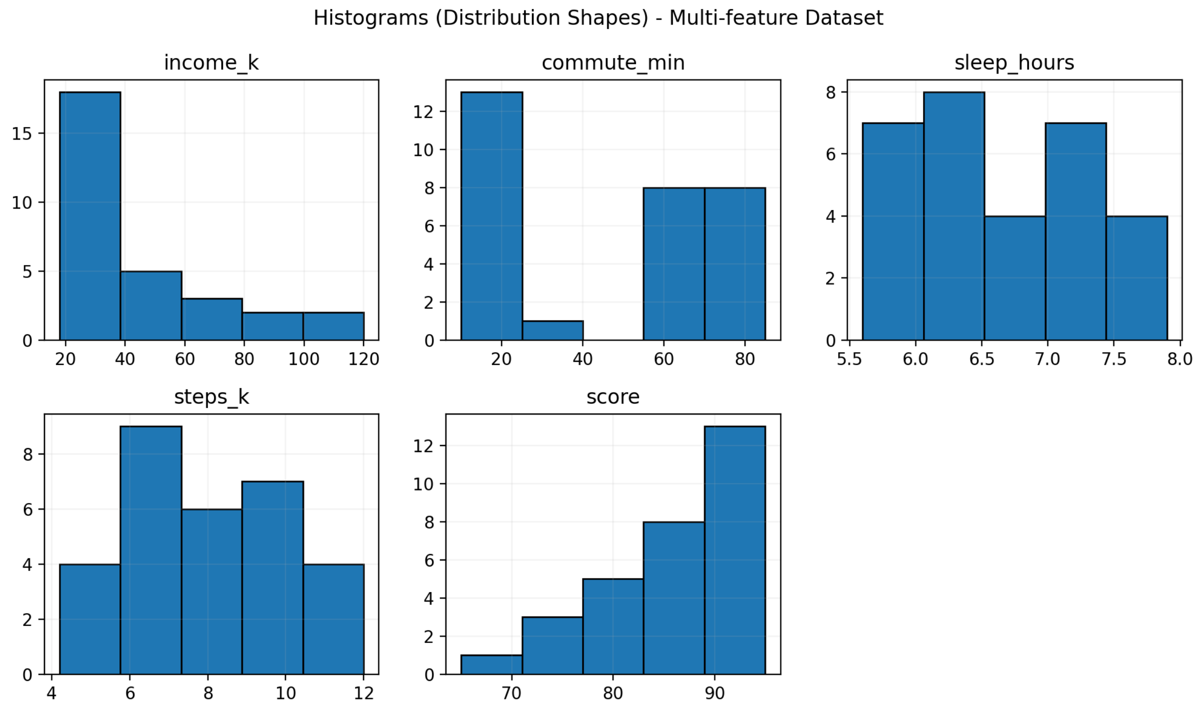
Uses:

- `data/multi_feature_distributions.csv`

Outputs:

- prints a per-feature summary (mean, median, std, quartiles, skewness)
- saves `images/hists_grid.png` (if matplotlib is installed)

## Demo Output (Histogram Grid)



## Summary

- Dimensional summaries describe each feature (column) separately
- Mean vs median gives a fast clue about skewness; confirm with plots
- Bimodal data can make the mean "not typical"
- For skew/outliers, use robust summaries (median/IQR) or transformations

**Exit question:** Why can the mean be misleading for a bimodal distribution?