Statistics and Data Analysis

Unit 02 – Lecture 05: Dimensional Summaries and Distributions

Tofik Ali

School of Computer Science, UPES Dehradun

February 9, 2026

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

Quick Links

Dimensional Summary Distribution Shape Outliers Demo Summary

Agenda

Learning Outcomes

- Explain what a dimensional (per-feature) summary is

Learning Outcomes

- Explain what a dimensional (per-feature) summary is
- Use mean/median/quartiles to get a quick idea of distribution shape

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)

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)

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

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

Common Distribution Shapes

- **Symmetric (approximately normal)**: mean $\approx$ median; bell-like
- **Right-skewed**: long tail to the right; mean $>$ median (often income)

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)

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

Outliers and Robustness

- Outliers can strongly affect mean and standard deviation
- Median and IQR are more robust (less sensitive to extremes)

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:

$$10 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 40$$

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

Robust Options (When Skew/Outliers Exist)

- Report median and IQR instead of mean and std
- Use trimmed mean (remove small % of extremes)

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

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

Mini Demo (Python)

Run from the lecture folder:

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

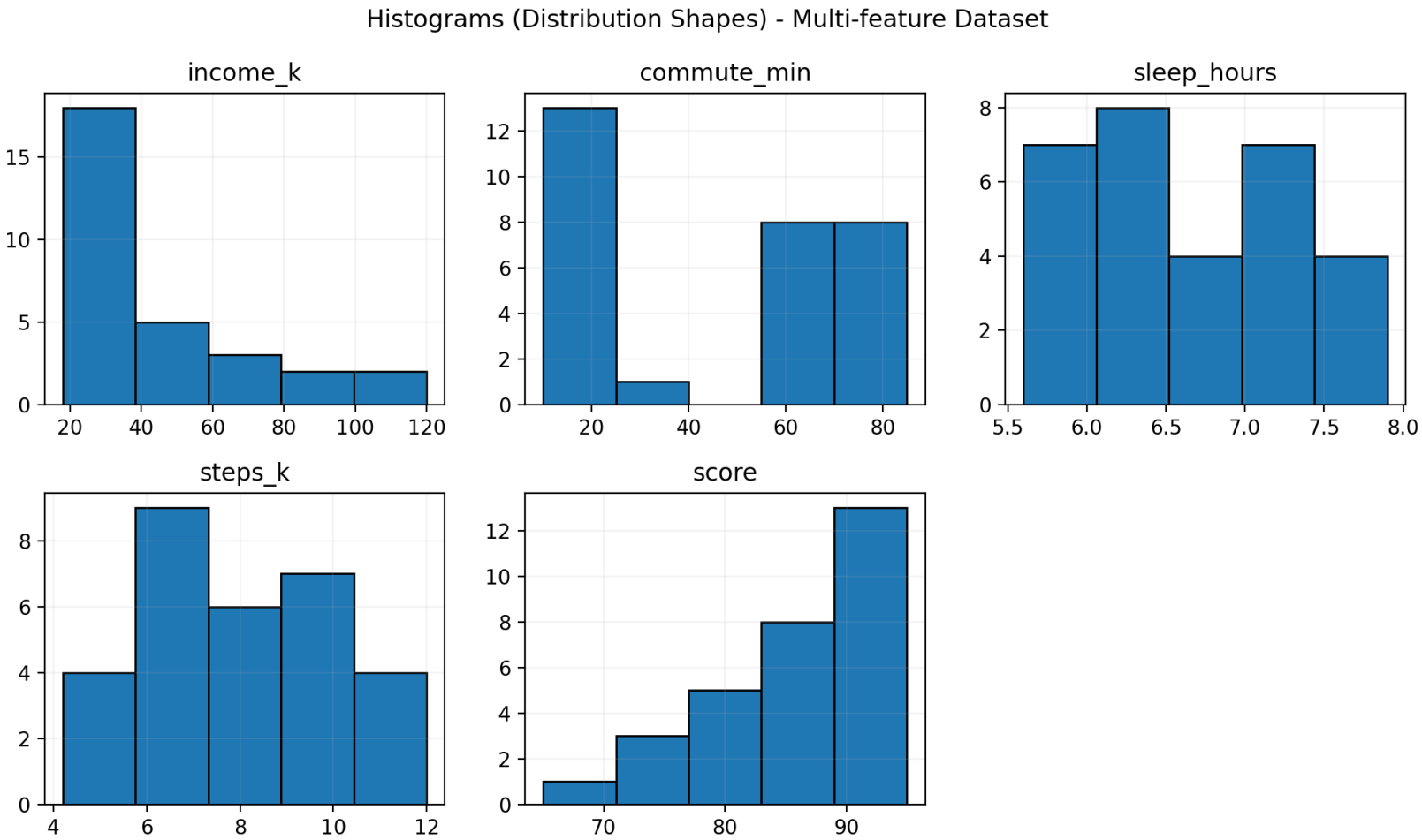
Uses:

- `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

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

Summary

- Dimensional summaries describe each feature (column) separately
- Mean vs median gives a fast clue about skewness; confirm with plots

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

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"

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

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?