Statistics and Data Analysis

Unit 02 – Lecture 04: Statistical Summaries for Data

Tofik Ali

School of Computer Science, UPES Dehradun

February 14, 2026

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

Quick Links

Overview Summary Tables Grouped Summary Demo Summary

Agenda

1. Overview
2. Summary Tables
3. Grouped Summary
4. Demo
5. Summary

Learning Outcomes

- Explain why we summarize data (communication and comparison)

Learning Outcomes

- Explain why we summarize data (communication and comparison)
- Interpret a standard summary table (count, mean, std, quartiles, min/max)

Learning Outcomes

- Explain why we summarize data (communication and comparison)
- Interpret a standard summary table (count, mean, std, quartiles, min/max)
- Compute and interpret a five-number summary (min, Q1, median, Q3, max)

Learning Outcomes

- Explain why we summarize data (communication and comparison)
- Interpret a standard summary table (count, mean, std, quartiles, min/max)
- Compute and interpret a five-number summary (min, Q1, median, Q3, max)
- Produce grouped summaries (mean/median by category)

Learning Outcomes

- Explain why we summarize data (communication and comparison)
- Interpret a standard summary table (count, mean, std, quartiles, min/max)
- Compute and interpret a five-number summary (min, Q1, median, Q3, max)
- Produce grouped summaries (mean/median by category)
- Explain what is lost when we compress data into a few numbers

Why Summaries?

A summary answers: “What does the dataset look like in one page?”

- We cannot read thousands of rows one-by-one

Why Summaries?

A summary answers: "What does the dataset look like in one page?"

- We cannot read thousands of rows one-by-one
- We need quick **comparison** across groups (CSE vs ECE) or time (week 1 vs week 2)

Why Summaries?

A summary answers: “What does the dataset look like in one page?”

- We cannot read thousands of rows one-by-one
- We need quick **comparison** across groups (CSE vs ECE) or time (week 1 vs week 2)
- Summaries are used in reports, dashboards, and as a first step in analysis

A Standard Summary Table (Common Columns)

For a numeric feature (example: `final_score`), a typical summary includes:

- **count** (n): number of non-missing values
- **mean**, **std** (sample standard deviation)
- **min**, **max**
- **25% (Q1)**, **50% (median)**, **75% (Q3)**

Idea: these numbers quickly describe center + spread + typical range.

Five-Number Summary (Very Important)

For a dataset x (sorted), the five-number summary is:

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

- It is the foundation of the boxplot

Five-Number Summary (Very Important)

For a dataset x (sorted), the five-number summary is:

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

- It is the foundation of the boxplot
- It is more robust than mean/std when outliers exist

Exercise 1: Five-Number Summary

Dataset:

$$4 \quad 5 \quad 7 \quad 8 \quad 9 \quad 10 \quad 25$$

Task: Compute min, Q_1, median, Q_3, max and IQR.

Solution 1

Sorted: 4, 5, 7, 8, 9, 10, 25 (n=7)

- 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 $= Q_3 - Q_1 = 10 - 5 = 5$

Reading Quartiles (Interpretation)

Quartiles are percentiles:

- Q_1 (25%): 25% of values are *at or below* Q_1

Checkpoint: the middle 50% of values lie between Q_1 and Q_3.

Reading Quartiles (Interpretation)

Quartiles are percentiles:

- Q_1 (25%): 25% of values are *at or below* Q_1
- median (50%): 50% of values are at or below the median

Checkpoint: the middle 50% of values lie between Q_1 and Q_3.

Reading Quartiles (Interpretation)

Quartiles are percentiles:

- Q_1 (25%): 25% of values are *at or below* Q_1
- median (50%): 50% of values are at or below the median
- Q_3 (75%): 75% of values are at or below Q_3

Checkpoint: the middle 50% of values lie between Q_1 and Q_3.

Exercise 2: Interpret a Summary Row

Suppose a feature has summary:

	count	mean	std	min	25%	50%	75%	max
final_score	24	71.0	12.0	40	65	74	82	92

Task: What do 25% and 75% mean? What does std tell us?

Solution 2

- 25% (Q1)=65: about 25% of students scored 65 or less.
- 75% (Q3)=82: about 75% of students scored 82 or less.
- std=12: a typical score is about 12 points away from the mean (rough idea of spread).

Important: summaries do not show the full distribution shape.

Mean vs Median in Summaries

- If mean $\approx$ median, distribution may be roughly symmetric

Rule of thumb: always confirm with a plot (histogram/boxplot).

Mean vs Median in Summaries

- If mean $\approx$ median, distribution may be roughly symmetric
- If mean $>$ median, data is often right-skewed (high outliers pull mean up)

Rule of thumb: always confirm with a plot (histogram/boxplot).

Mean vs Median in Summaries

- If mean $\approx$ median, distribution may be roughly symmetric
- If mean $>$ median, data is often right-skewed (high outliers pull mean up)
- If mean $<$ median, data is often left-skewed (low outliers pull mean down)

Rule of thumb: always confirm with a plot (histogram/boxplot).

Exercise 3: Group Comparison (Outlier Effect)

Two groups:

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

Task: Compute mean and median for both groups. Which group is “better”?

Solution 3

- Group A mean $= (60 + 62 + 65 + 95)/4 = 70.5$; median $= (62 + 65)/2 = 63.5$
- Group B mean $= (70 + 72 + 73 + 74)/4 = 72.25$; median $= (72 + 73)/2 = 72.5$

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

Grouped Summaries (Stratification)

Instead of one summary for the entire dataset, we summarize **by group**:

- mean/median `final_score` by `program`

Why? A single global average can hide important group differences.

Grouped Summaries (Stratification)

Instead of one summary for the entire dataset, we summarize **by group**:

- mean/median `final_score` by `program`
- mean attendance by section or batch

Why? A single global average can hide important group differences.

Grouped Summaries (Stratification)

Instead of one summary for the entire dataset, we summarize **by group**:

- mean/median `final_score` by `program`
- mean attendance by section or batch
- revenue by category, etc.

Why? A single global average can hide important group differences.

Exercise 4: Weighted Mean (Correct Overall Average)

Suppose:

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

Task: Compute the overall mean score (all 15 students together).

Solution 4

Overall mean is a **weighted mean**:

$$\bar{x} = \frac{70 \cdot 10 + 80 \cdot 5}{10 + 5} = \frac{700 + 400}{15} = \frac{1100}{15} \approx 73.33$$

Note: $(70 + 80)/2 = 75$ is wrong because group sizes are different.

Exercise 5: Mean by Program (Small Table)

Program	final_score values
CSE	70, 75
ECE	60, 65
AIML	80, 85

Task: Compute mean final_score for each program.

Solution 5

- CSE mean $= (70 + 75)/2 = 72.5$
- ECE mean $= (60 + 65)/2 = 62.5$
- AIML mean $= (80 + 85)/2 = 82.5$

Interpretation: group summaries let us compare programs directly.

Exercise 6: What Is Lost in a Summary Table?

Question: If we only report mean and std for a dataset, what could we miss?

- Think about outliers, skewness, and multi-modal distributions.

Solution 6

A small set of numbers can hide:

- outliers (one extreme value can distort mean/std)
- skewness (mean $\neq$ median) and long tails
- multi-modality (two peaks) where “average” is not typical
- subgroup differences (one group high, one group low)

Takeaway: summaries are useful, but always validate with plots.

Mini Demo (Python)

Run from the lecture folder:

```
python demo/statistical_summaries_demo.py
```

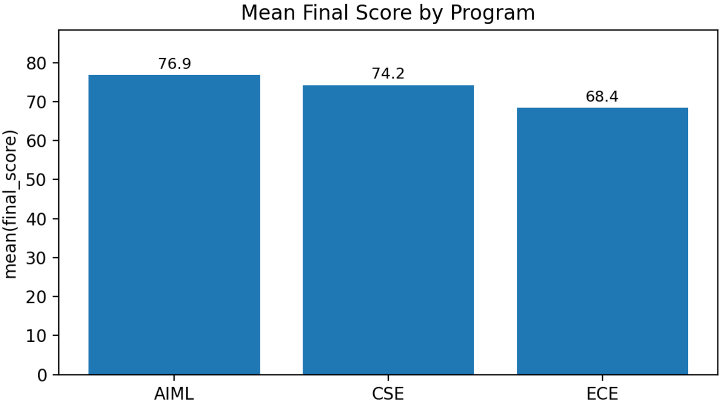
Uses:

- `data/student_summary.csv`

Outputs:

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

Demo Output (Example Plot)



Summary

- Summary tables compress data into center + spread + typical range (quartiles)

Exit question: Why is a weighted mean needed when groups have different sizes?

Summary

- Summary tables compress data into center + spread + typical range (quartiles)
- Five-number summary is robust and supports boxplot thinking

Exit question: Why is a weighted mean needed when groups have different sizes?

Summary

- Summary tables compress data into center + spread + typical range (quartiles)
- Five-number summary is robust and supports boxplot thinking
- Grouped summaries (stratification) reveal differences hidden by global averages

Exit question: Why is a weighted mean needed when groups have different sizes?

Summary

- Summary tables compress data into center + spread + typical range (quartiles)
- Five-number summary is robust and supports boxplot thinking
- Grouped summaries (stratification) reveal differences hidden by global averages
- Summaries can hide distribution shape and outliers $\Rightarrow$ use plots too

Exit question: Why is a weighted mean needed when groups have different sizes?