Statistics and Data Analysis

Unit 02 – Lecture 01: Measures of Central Tendency

Tofik Ali

School of Computer Science, UPES Dehradun

February 2, 2026

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

Quick Links

Core Concepts Visuals Interactive

Agenda

Learning Outcomes

- Differentiate mean, median, and mode

Learning Outcomes

- Differentiate mean, median, and mode
- Select an appropriate measure for context

Learning Outcomes

- Differentiate mean, median, and mode
- Select an appropriate measure for context
- Explain robustness and outlier effects

Learning Outcomes

- Differentiate mean, median, and mode
- Select an appropriate measure for context
- Explain robustness and outlier effects
- Apply measures to real datasets

Why Central Tendency?

- Summarizes a dataset with one representative value

Why Central Tendency?

- Summarizes a dataset with one representative value
- Enables comparison across groups or time

Why Central Tendency?

- Summarizes a dataset with one representative value
- Enables comparison across groups or time
- Foundation for dispersion and inference

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Sensitive to extreme values

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Sensitive to extreme values
- Best for symmetric distributions

Median

- Middle value after sorting

Median

- Middle value after sorting
- Robust to outliers and skew

Median

- Middle value after sorting
- Robust to outliers and skew
- Preferred for income or skewed data

Mode

- Most frequent value

Mode

- Most frequent value
- Useful for categorical data

Mode

- Most frequent value
- Useful for categorical data
- Can be multi-modal

Categorical Example (Mode)

Dataset (Major Specialization):

CSE	CSE	AI	CSE	DS	AI
CSE	DS	AI	AI	CSE	DS

Mode: CSE (most frequent category).

Decimal Real-Value Example

Dataset (CGPA values):

7.2	7.5	7.8	8.1	8.3	8.6
7.4	7.9	8.0	8.2	8.5	8.7

Use: Mean/median for central tendency of continuous data.

Multi-Dimensional Example

Dataset (Student Attributes):

Student	Attendance (%)	Quiz Score	Project Score
A	92	8.5	9.0
B	85	7.8	8.2
C	88	9.1	8.7
D	76	6.9	7.5

Use: Compute mean/median per feature (column-wise).

Multi-Dimensional Categorical Example

Dataset (Student Profile):

Student	Major	Section	Club
A	CSE	A	Robotics
B	AI	B	AI
C	CSE	A	Coding
D	DS	B	Robotics
E	CSE	A	AI

Use: Mode per column (most common major/section/club).

Dataset for Calculation Exercise

Use this sample income data (in thousands) for mean, median, and mode:

12	14	15	15	16	18
19	20	22	25	27	30
32	35	40			

Exercise: Compute mean, median, and mode. Comment on skew.

Exercise Solution (Summary)

Sorted data (n=15): 12, 14, 15, 15, 16, 18, 19, 20, 22, 25, 27, 30, 32, 35, 40

- **Mean** = 22.67
- **Median** = 20
- **Mode** = 15
- **Skew** = right-skewed (mean > median)

Comparison Table

Measure	Mean	Median	Mode
Symmetric data	✓	✓	✓
Skewed data		✓	✓
Outliers present		✓	✓

Outlier Effect (Step 1)

Dataset (in thousands):

12	14	15	15	16	18
19	20	22	25	27	30
32	35	40			

Add a large outlier to the dataset: 200

Outlier Effect (Step 2)

- **Mean:** 22.67 $\rightarrow$ 33.75 *Mean shifts strongly*
- **Median:** 20 $\rightarrow$ 21 *Median changes little*

Equation Sample

$$\bar{x}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

Trimmed mean for robustness.

Checkpoint Question

Which measure best describes income data, and why?

Mini Demo (Python)

- Load sample dataset
- Compute mean, median, mode
- Visualize distribution

Summary

- Mean, median, mode capture different centers

Exit question: What would you report first and why?

Summary

- Mean, median, mode capture different centers
- Robustness matters for skew/outliers

Exit question: What would you report first and why?

Summary

- Mean, median, mode capture different centers
- Robustness matters for skew/outliers
- Always check the distribution shape

Exit question: What would you report first and why?