# Statistics and Data Analysis
# Unit 02 – Lecture 01 Notes

Tofik Ali

February 17, 2026

## Measures of Central Tendency

### Learning Outcomes

- Differentiate mean, median, and mode.
- Choose the most appropriate measure for a given context.
- Explain robustness and the effect of outliers.
- Apply these measures to real datasets.

## Slide-by-slide Notes

### Title Slide

Introduce the unit and the focus on central tendency as the "center" of data. Explain that many real datasets are large and messy, so we need a simple way to summarize what a typical value looks like. Central tendency gives us a single number that represents the dataset.

### Quick Links / Agenda

Explain the flow: core concepts, examples and visuals, then a short demo and exercise. Tell students the aim is to understand when each measure is useful, not just how to calculate it.

### Learning Outcomes

Central tendency is the foundation for descriptive statistics and later inference. By the end of the lecture, students should be able to pick the correct measure for a given context and justify their choice.

### Why Central Tendency?

Central tendency compresses data into a representative value for comparison and communication. In practice, people ask: "What is a typical salary?" or "What is the typical score?" The answer depends on the data distribution. No single measure is "best" in all cases because different datasets have different shapes and outliers. When we say it "enables comparison across groups or time," we mean we can reduce each group (e.g., Section A vs. Section B) or each time period (e.g., this year vs. last year) to a single representative value and then compare those values. For example, if the median score in Section A is 72 and in Section B is 68, we can quickly see which group performed better. Similarly, if the mean weekly sales were 50 units in January and 65 units in February, we can say sales increased over time. Without a central measure, comparing large datasets is slower and harder to communicate.

### Mean

The arithmetic mean is the average:

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

Explain each symbol: $n$ is the number of data points, $x_i$ is the $i$-th value, and $\bar{x}$ is read as “x-bar.” The mean adds all values and divides by the count. It uses every data point, so it changes when any value changes.
**Strengths:** stable for symmetric distributions, easy to compute, and used in later formulas (variance, covariance, regression).
**Limitations:** sensitive to outliers. A single extremely large or small value can pull the mean away from the “typical” value.

### Median

The median is the middle value after sorting the data. If there are an odd number of values, the median is the middle one. If there are an even number, the median is the average of the two middle values.
**Strengths:** robust to extreme values; preferred for skewed data (e.g., income, house prices).
**Limitations:** ignores the magnitude of extreme values, so it is less sensitive to tail changes.

### Mode

The mode is the most frequent value or category. A dataset can have one mode, multiple modes (bi-modal, multi-modal), or no clear mode.
**Strengths:** useful for categorical or discrete data (e.g., most common major, most common defect type).
**Limitations:** can be unstable and may not represent the “center” for continuous data.

### Categorical Example (Mode)

Use a categorical dataset like “Major Specialization” to show that mean and median are not meaningful for categories, while mode is meaningful. The mode answers: which category appears most often?

### Decimal Real-Value Example

For continuous values such as CGPA, mean and median are both valid. Emphasize that these values are real numbers and represent typical performance. The mode may be less informative if values are mostly unique.

### Multi-Dimensional Example

For a table with multiple columns (e.g., attendance, quiz score, project score), central tendency is computed column-wise. This shows that a dataset can have several "centers," one for each feature.

### Dataset for Calculation Exercise

Provide the small income dataset. Ask students to compute mean, median, and mode manually. This helps them practice sorting and calculating.

### Exercise Solution

Explain the solution step-by-step:

- Count values: $n = 15$.
- Sort the data and identify the middle value (median).
- Compute the mean by summing all values and dividing by 15.
- Identify the most frequent value (mode).

Highlight that the mean is larger than the median, which indicates right-skew.

### Comparison Table

Selection rules:

- Symmetric data: mean/median/mode are all acceptable.
- Skewed data or outliers: median (or mode) is more stable.

Explain that this is a practical checklist for choosing the correct measure.

### Outlier Effect

Demonstrate with a simple example:

- Dataset A: 10, 12, 12, 13, 14 $\rightarrow$ mean $\approx 12.2$, median $= 12$
- Dataset B: add outlier 100 $\rightarrow$ mean $\approx 30.2$, median $= 12$

Conclusion: median is robust. This is why reports for skewed data often use median instead of mean.

### Trimmed Mean

The trimmed mean removes a small percentage of the smallest and largest values before computing the mean. It is a compromise between mean and median and is useful when there are mild outliers but we still want to use most of the data.

### Checkpoint Question

Which measure best describes income data, and why?
Expected: median, because income distributions are typically right-skewed. Right-skewed means a long tail to the right: a small number of very high incomes pull the mean upward, while most values remain lower. The median resists this pull and better represents the "typical" income for the majority of people.

### Mini Demo (Python)

Walk through:

1. Load dataset
2. Compute mean/median/mode
3. Visualize histogram and show effect of outlier

Explain each step briefly, and remind students to compare mean and median after adding the outlier.

Demo assets:

- `demo/mini_demo.py`
- `data/income_small.csv`

### Summary / Exit Question

Reinforce the selection logic and ask students to justify a choice in context. Encourage them to always look at the distribution shape before deciding on a single summary value.

## Demo (Python)

Run:

```
python demo/mini_demo.py
```

What it shows:

- Mean/median/mode on a small income dataset
- Mean shift after injecting an outlier
- Histogram of values

## References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley, 7th ed., 2020.
- Gupta, S. C., & Kapoor, V. K. *Fundamentals of Applied Statistics*, Sultan Chand & Sons, 4th rev. 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

**Core Concepts** **Visuals** **Interactive**

## Agenda

- Overview
- Core Concepts
- Visuals
- Interactive
- Summary

## 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
- 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
- Best for symmetric distributions

## Median

- Middle value after sorting
- Robust to outliers and skew
- Preferred for income or skewed 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 → 33.75 *Mean shifts strongly*
- **Median:** 20 → 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
- Robustness matters for skew/outliers
- Always check the distribution shape

**Exit question:** What would you report first and why?