Measures of Central Tendency

1. Introduction

Measures of Central Tendency are statistical metrics used to identify the center or typical value of a dataset.

They help summarize large amounts of data with a single representative value, making it easier to understand the dataset's general characteristics.

The three most common measures are:

- Mean
- Median
- Mode

Understanding these measures is crucial in data science and analytics, as they provide insights into the distribution and tendencies of data.

2. Mean (Arithmetic Average)

The mean is the sum of all values in the dataset divided by the number of values. It is the most widely used measure of central tendency.

- Formula:

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

- Example

Dataset: 4, 6, 8, 10, 12

Mean= (4+6+8+10+12) / 5 = 40/5 = 8

- When to Use

Works well for symmetrical distributions without extreme outliers.

- Limitations

Sensitive to outliers (e.g., a very large or very small value can significantly change the mean).

3. Median

The median is the middle value when the dataset is ordered from smallest to largest.

If there is an even number of values, the median is the average of the two middle values.

- Steps to Calculate

    1. Sort the dataset in ascending order.
    2. Identify the middle position:
        - If n is odd, → middle value is the median.
        - If n is even → median = average of the two middle values.

- Example

Dataset: 3, 7, 8, 12, 14

Median = 8 (middle value).

Even dataset: 3, 7, 8, 12

Median = (7 + 8) / 2 = 7.5

- When to Use

Useful for skewed distributions or when there are extreme outliers.

- Limitations

Does not consider all data values in calculation (only the position).

4. Mode

The mode is the most frequently occurring value in the dataset.

A dataset can have:

- o Unimodal: one mode
- o Bimodal: two modes
- o Multimodal: more than two modes
- o No mode: all values occur with the same frequency

- Example

Dataset: 2, 4, 4, 6, 8

Mode = 4

Dataset: 1, 1, 2, 2, 3

Modes = 1 and 2 (bimodal)

- When to Use
    - Best for categorical data (e.g., most common category).
    - Also useful for discrete numerical data.

- Limitations
    - May not exist or may not be unique.
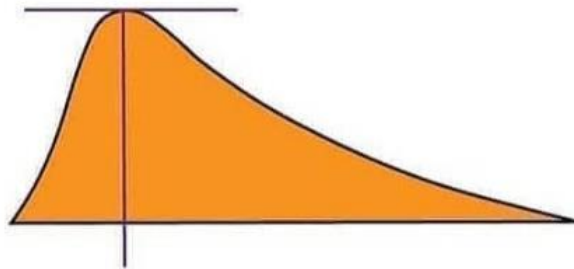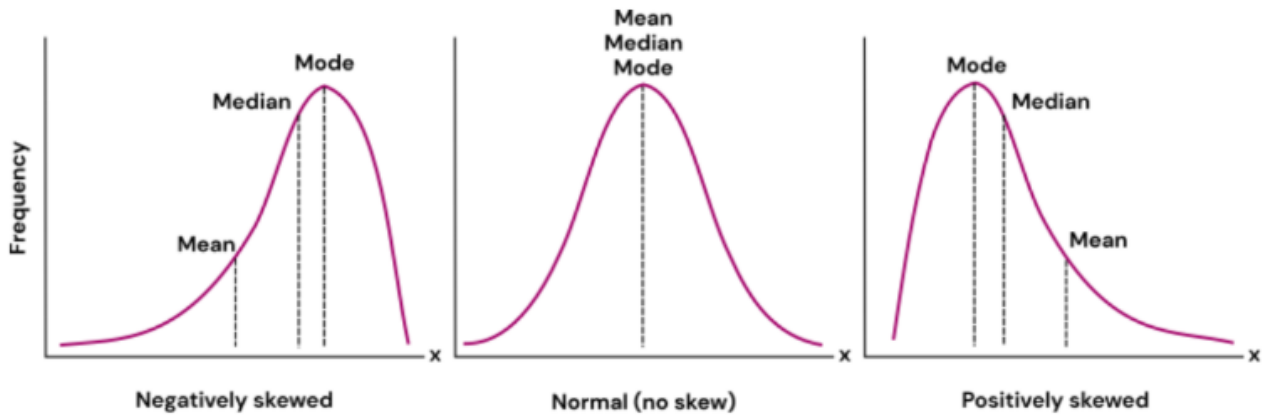   - Less useful for continuous data.

5. Comparison

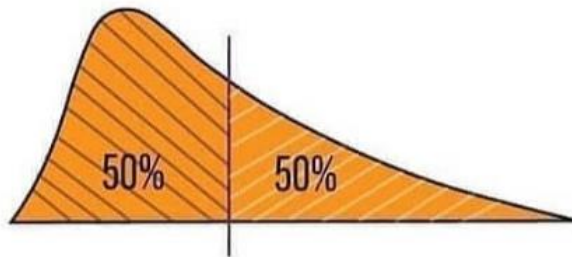| Feature	Mean	Median	Mode
Uses all values?	☑ Yes	✖ No	✖ No
Sensitive to outliers?	☑ Yes	✖ No	✖ No
Best for	Symmetrical numeric data	Skewed numeric data	Categorical/discrete data

6. Common Mistakes

o Using mean with heavily skewed data (gives misleading center).
o Forgetting to sort the data when calculating the median.
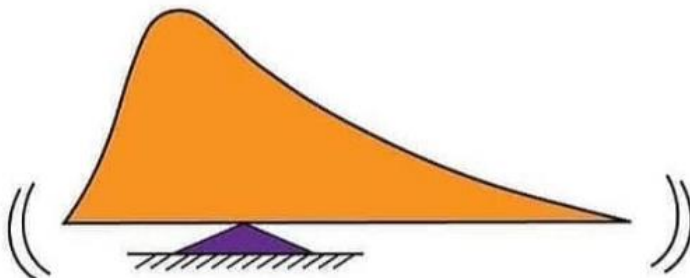o Assuming there is always one unique mode.

7. Visual Example