

Measure of Central Tendency

Measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a single value that summarizes a set of data by identifying the central position within that dataset.

- ① Mean or Average
- ② Median
- ③ Mode

$$Ages = [24, 32, 12, 48, 16, 20]$$

↓
Center point

Central position



① Mean

Mean is the sum of all values divided by the number of values

Population Mean (μ)

Population (N)

$$\mu = \sum_{i=1}^N \frac{X_i}{N} \quad \left\{ N = \begin{array}{l} \text{Population} \\ \text{size} \end{array} \right\}$$

Sample mean (\bar{x})

Sample (n)

$$n \leq N$$

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

$n \rightarrow$ Sample size

Here X is a random variable

$$X = \{5, 8, 12, 15, 20\}$$

$$N = 5$$

$$\mu = \frac{5+8+12+15+20}{5} = \frac{60}{5} = 12 //$$

* Characteristics

- * Affected by extreme outliers
- * Used for Interval And Ratio Data

$$X = \{1, 2, 3, 4, 5\}$$

$$\mu = \frac{1+2+3+4+5}{5} = 3$$

$$X = \{1, 2, 3, 4, 5, 100\}$$

$$\mu = \frac{1+2+3+4+5+100}{5} = \frac{115}{5} = 23 //$$

② Median

The median is the middle value in a dataset when the values are arranged in ascending or descending order.

$$X = \{1, 2, \boxed{3}, 4, 5\}$$

No. of elements = 5

5 is odd

$$\text{Median} = \underline{\underline{3}}$$

$$X = \{3, 4, 1, 5, 2, 100\} \Rightarrow \{1, 2, \boxed{3}, \boxed{4}, 5, 100\}$$

No. of element = 6

6 is even

$$\text{Median} = \frac{3+4}{2} = 3.5$$

Characteristics

- * Not affected by extreme outliers
- * Used for ordinal, interval and ratio data.

③ Mode

Defn: The mode is the value that appears most frequently in a dataset.

Dataset: 2, 4, 4, 6, 7, 7, 7, 9

Mode = 7 (most frequent value)

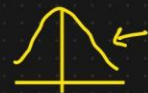
3, 5, 5, 6, 6, 8

Mode = 5, 6 (bimodal)

④ Characteristics

- ① Not affected by extreme values.
- ② Used for Nominal, ordinal, interval and ratio data.

Choosing the Appropriate Measure



1. Mean: Best used when data is symmetrically distributed without outliers. Provides a mathematical average, which is useful for further statistical calculations.



2. Median: Best used when data is skewed or contains outliers. Provides the middle value, which better represents the center of a skewed dataset.

3. Mode: Best used for categorical data to identify the most common category. Also useful for identifying the most frequent value in ordinal, interval, or ratio data.

Real World Application

EDA ← Feature Engineering

Mode
↑

Gender
↓
M Nominal
F Dat

Mode
↑

Degree
↓

BE

Handling
the missing
value

{ Nominal
+
Ordinal }

PHD

B.E

Master

Bsc

B.E

Mean



Median



Age

24

25

27

24

32

-

-

40

Weight

70

80

95

-

-

60

65

72

Salary

40K

70K

45K

50K

60K

-

55K

-

Gender

M

F

F

M

-

-

-

M

Degree

BE

-

-

PHD

B.E

Master

Bsc

B.E