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)

$$\mu = \sum_{i=1}^{N} \frac{X_i}{N}$$ { N = Population size }.

Here X is a random variable

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

$N = 5$

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

note: $n \leq N$

Sample mean ($\bar{x}$)

Sample (n)

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

$n \rightarrow$ Sample size

(*) Characteristics

(*) Affected by extreme outliers

(*) Used for interval And Ratio Data

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

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

adding outlier 100

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

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

note : outlier is an no. which does not belong to original set of random variables .

thus , we can see that means has changed from 3 to 23 i.e affected by extreme outlier.

(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

Median = 3

adding outlier

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

No. of element = 6

6 is even

$\text{Median} = \frac{3+4}{2} = 3.5$ → not getting affected

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)

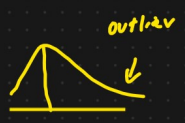
(*) Characteristics

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.

hence , we can replace the missing values with mean/median/mode as per the req.