Measure of central Tendency

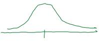
Measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a single value that summarizes a set of data by identifying the central position within that dataset.

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)

Sample mean ($\bar{X}$)

Population (N)

$n \leq N$

Sample(n)

Sample n is a subset of population.

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

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

Here X is a random variable {N → population size}

n → is sample size

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

$$N = 5$$

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

Characteristics

1) Affected by extreme outliers.

2) Used for interval and ratio data

without outlier

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

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

with outlier

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

$$\mu = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.166$$

2) Median

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

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

The numero de elements is 5, 5 is odd

Median = 3

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

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

No of elements = 6
6 is even

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

Characteristics

i) Not affected by extreme outliers.

ii) Used for ordinal interval and ratio data.

3) Mode

Definition: 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)

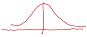
Mode = 5, 6 {bimodal}

*) Characteristics

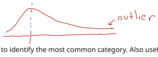
1) Not effected by extreme values.
2) used for nominal, ordinal interval, and ratio data.

Choosing the appropiate measure

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



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



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

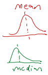
Real word application

Feacture engineering => EDA

for missing Mode

Mode {Nominal + ordinal}

Handling the missing value



	Age	Weight	Salary	Gender	Degree
1	24	70	40k	M	BE
2	25	80	70k	F	-
3	27	95	45k	F	-
4	24	-	50k	M	PHD
5	32	-	60k	-	BE
6	-	60	-	-	Master
7	-	65	55k	-	BSC
8	40	72	-	M	BE