

Type of Data

Categorical Data /
Qualitative Data

Numerical Data /
Quantitative data

Nominal

no order / No Rank

Ordinal

order / Rank

Discrete data
khok. no.

Continuous data

ex - Gender
male / female

ex - General

SL	5*
3AC	4*
2AC	3*
1AC	2*
	1*
	Bad.

Winner :- First
Second
Third.

↓ Age
↓ T-shirt size
↓ no. of children

↓ Height
↓ weight
↓ Length.

Statistical Measures

① Measure of Central Tendency.

Zero
mile

Center

A measure of Central tendency is a statistical measure that represents a typical or central value for a dataset.

① Mean \Rightarrow It represent average Value of the dataset.. Sum of all values to the no. of Value.

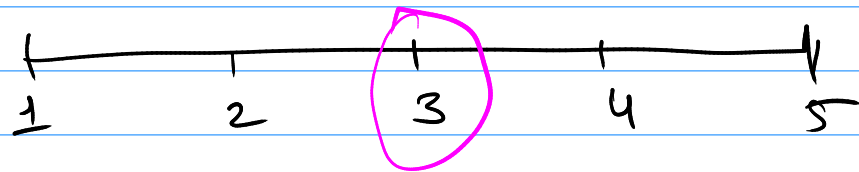
Notation $\mu = \frac{\sum}{\text{Population mean}}$ $\bar{X} = \frac{\sum}{\text{Sample mean.}}$

ex = data = [1, 2, 3, 4, 5]

$N = \text{no. of value}$

$N = 5$

$\mu = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$



Use $\mu \rightarrow$ To replace the missing values in the numerical dataset.

ex = ¹⁻18⁺ year class Age

108 [17, 18, 17, 18, 19, 20, 18, 19, 20, 20, ...]
 - - - 8 - n - - -)

8 missing \rightarrow replace = mean

$\frac{18 \cdot 20}{18} \quad \frac{18 \cdot 49}{18} \quad | \quad \frac{18 \cdot 50, 18 \cdot 51}{19}$

Disadvantage \Rightarrow It Robbust to outlier

◊ It affected by the outlier.

1st year data = [17, 18, 19, 18, 19, 17, 19, 18, 20, 18]

$$\text{mean} = \frac{17+18+19+18+19+17+19+18+20+18}{10}$$

$$= 183/10 = 18.3 \approx \underline{\underline{18}}$$

mean

outlier = [17, 18, 19, 18, 19, 17, 19, 18, 20, 81]

$$\text{mean} = \frac{17+18+19+18+19+17+19+18+20+8}{10}$$
$$= 246/10 = 24.6 \approx 25$$

 x

② median

o = The middle value after sorting the data.

$Q = [5, 1, 3, 2, 4] \xrightarrow{\text{Sort}} [1, 2, 3, 4, 5]$

The element 3 is highlighted as the median.

$$\text{median} = \left(\frac{n+1}{2} \right)$$

$$\text{mean} = \frac{17+18+19+18+19+17+19+18+20+8}{10} \quad \text{outlier}$$

$$\text{Sort} = 17, 17, 18, 18, 18, 19, 19, 19, 20, 8$$

$$\therefore \text{median} = \frac{n+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$$

$$\frac{18+19}{2} = 18.5 = 18.5$$

Replace with median

③ mode \Rightarrow The most frequent Value in the dataset.

Gender

M - 1
F - 2
M - 3
M - 4
F - 5
F - 6

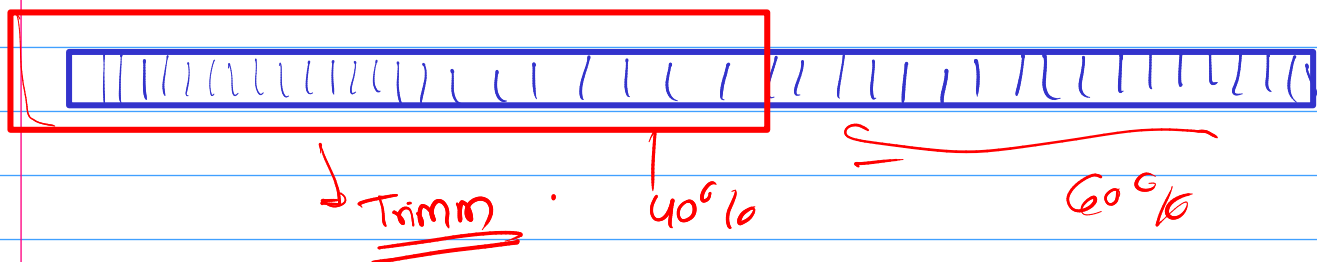
$$\begin{aligned} m &= 6 \\ f &= 4 \end{aligned}$$

mode = Male

\hookrightarrow To replace missing Value in categorical feature.

$$m=6, f=4 \rightarrow$$

4. Trimmed mean



⑤ Weighted Mean (ML)

Hourly price

$$= \begin{array}{l} \boxed{\text{RF}} \Rightarrow 0.8 \Rightarrow 20L \\ \boxed{\text{LR}} \Rightarrow 0.9 \Rightarrow 22L \end{array}$$

$$\boxed{\text{Xgboost}} \Rightarrow 0.85 \Rightarrow 21L$$

Weight

$$\frac{0.8 \times 20L + 0.9 \times 22L + 0.85 \times 21L}{0.8 + 0.9 + 0.85}$$

Measure of Dispersion