

Type of Data

Categorical Data /

Qualitative Data

Nominal

NO Order/No Rank

Ordinal

Order / Rank

Numerical Data /

Quantitative data

Discrete data
Whole no.

Continuous
data

Age

T-shirt size

no. of children

Height
Weight
Length.

Ex - Gender

male/female

ex - General

SL

SAC

2AC

1AC

Bad.

Kinner :- First
Second
Third.

Statistical Measures

① Measure of Central Tendency.

Zero
milk

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.

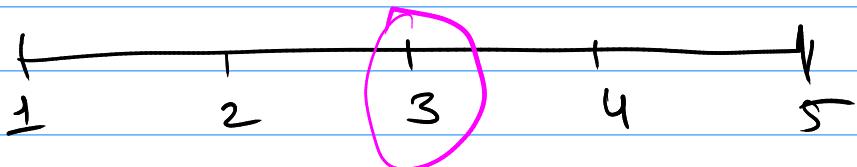
$$\begin{array}{l} \text{Notation} \quad \underline{\underline{\mu}} = \underline{\underline{\bar{x}}} \\ \text{Population mean} \quad \text{Sample mean.} \end{array}$$

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

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

$$N = 5$$

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



Use \Rightarrow To replace the missing values in the numerical dataset.

ex = $\frac{1}{18+}$ year class Age

$$\underline{\underline{108}} \left[17, 18, 17, 18, 19, 20, 18, 19, 20, 20, \dots \right]$$

8 missing → replace = mean

$$\frac{18.20}{18} \quad --- \quad \frac{18.49}{18} \quad | \quad \frac{18.50, 18.51}{19} \rightarrow$$

Disadvantage

⇒ It Robust to outlier

⇒ It affected by the outlier.

$$\text{1st year} \underset{\text{data}}{=} [17, 18, 19, 18, 19, 17, 19, 18, 20, 18]$$

1 2 3 4 5 6 7 8 9 10

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

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

mean

$$\text{outlier} = [17, 18, 19, 18, 19, 17, 19, 18, 20, 81]$$

1 2 3 4 5 6 7 8 9 10
 outlier

$$\text{mean} = \frac{17+18+19+18+19+17+19+18+20+81}{10}$$

$$= 246/10 = 24.6 \approx \underline{\underline{25}}$$

→ —

② median

⇒ The middle value after sorting the data.

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

1 2 3 4 5
median

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

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

Sort = 17, 17, 18, 18, 18, 19, 19, 19, 20, 8
 ↓ 2 ↓ 3 ↓ 4 ~~5~~ 6

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

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

→ —

Replace with median

③ mode ⇒ The most frequent value in the dataset.

Gender

M - 1

F

M - 2

M - 3

F

M - 4

F

M - 5

M - 6

F

—

—

$$M = 6$$

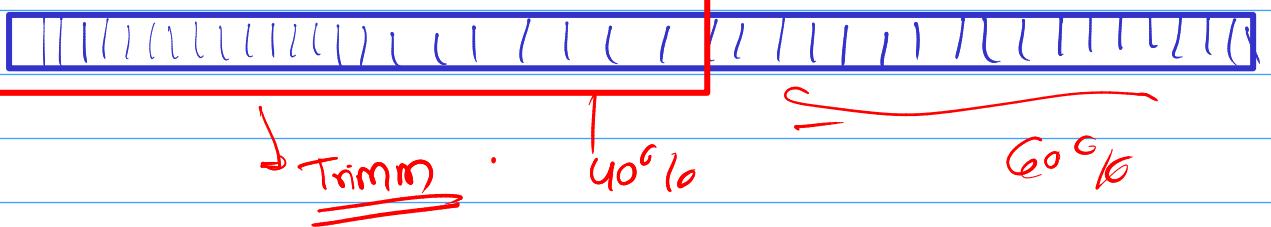
$$F = 4$$

$$\boxed{\text{mode} = \text{Male}}$$

↑ To replace missing value in categorical feature.

$$\underline{\underline{M=6}}, \underline{\underline{F=4}} \rightarrow$$

4. Trimmed Mean



5 Weighted Mean (ML)

= $\boxed{RF} \Rightarrow 0.8 \Rightarrow 20L$

$\boxed{LR} \Rightarrow 0.9 \Rightarrow 22L$

$\boxed{Xgpost} \Rightarrow 0.85 \Rightarrow 21L$

~~weight~~ $0.8 \times 20L + 0.9 \times 22L + 0.85 \times 21L$
 $0.8 + 0.9 + 0.85$

~~Measure of Dispersion~~