Date:

3. MODE

- unimode
- no mode
- multiple mode

SURVEY → favourite FRUIT among APPLE, BANANA, ORANGES :-
∴ MODE = FRUIT mentioned ↑frequently

- most frequent value.
- Applicable Best FOR CATEGORICAL VARIABLES.
- In case of NUMERICAL VARIABLES → unique Count SHOULD be less.

E.g.:- $X = \{2, 4, 3, 4, 2, 6, 7, 2\}$

$\Rightarrow$ (2) ≡ MODE.

II MEASURES OF DISPERSION

↳ also known as SPREAD or VARIABILITY.

- It refers to statistics that describes how data points in a dataset are **distributed** or **spreadout**.

↳ Measures OF Dispersion like

- RANGE
- INTERQUARTILE RANGE (IQR)
- VARIANCE
- STANDARD DEVIATION, help us understand how much data values diverge or spread FROM central values like mean or median.

∴ ↑SPREAD ⟹ ↑VARIABILITY ⟹ ↓CONSISTENCY or PREDICTIBILITY

- Higher variability can be useful in some scenarios and less desirable in OTHERS

1. RANGE

- Range = {maximum value of dataset − minimum value of dataset}

↳ SENSITIVE to EXTREME VALUES / OUTLIERS.

E.g.- {1, 2, 3, 4}
Range = 4 − 1
= 3

E.g.- {1, 2, 3, 4, 50}
⇒ Range = {50 − 1}
= 49

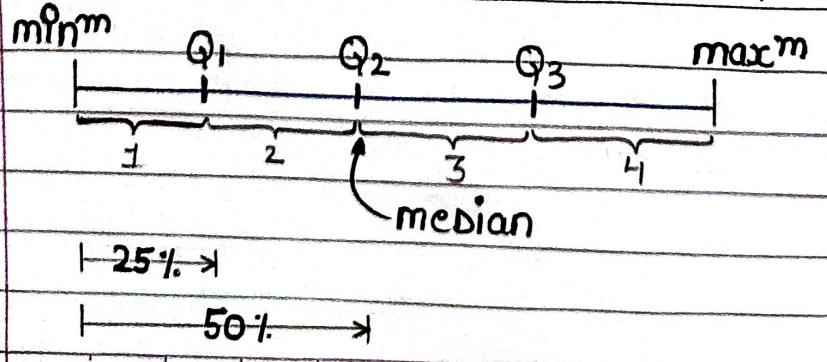
↳ Does NOT give any INFORMATION on HOW data is DISTRIBUTED around the mean.

↳ Suitable for small datasets.

2. QUARTILES [Q_1, Q_2 & Q_3]

- Quartiles are values that divide a dataset into 4 equal parts each containing a quarter of the data. They help describe the distribution.

↳ FIRST QUARTILE, Q_1 – value that separates lowest 25% of data
↳ SECOND QUARTILE, Q_2 – MEDIAN or the midpoint of the data.
↳ THIRD QUARTILE, Q_3 – value that separates the lowest 75% of data from highest 25%



- For finding quartiles try to arrange your dataset in increasing order.

Page No.:
Date:

→ $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ term

→ $Q_2 = \left(\frac{n+1}{2}\right)^{th}$ term

→ $Q_3 = \left(\frac{3(n+1)}{4}\right)^{th}$ term

E.g. - $x = (4, 6, 7, 8, 10, 23, 34)$

- $Q_1 = (7+1/4) = (8/4) \equiv Q_1 = 6$
- $Q_2 = (7+1/2) = 4th$ Term $\equiv Q_2 = 8$
- $Q_3 = (3(7+1)/4) = 6^{th}$ Term $\equiv Q_3 = 23$

3. PERCENTILES :

→ Percentiles divide a dataset into 100 equal parts. For example the 25 th percentile is the value below which 25% of the data falls.

→ QUARTILES = 4 Equal parts
PERCENTILES = 100 Equal parts.

Position, $P = \frac{p}{100} \times (n+1)$

- p = percentile
- n = no. of observations.

→ E.g.:- $x = 4, 6, 7, 8, 10, 23, 34$
Find 5 percentile of the given dataset.

We know :- Position, $P = \frac{p}{100} \times (n+1) = \frac{5}{100} \times (7+1) = \left(\frac{1}{20} \times 8\right)$

$= 0.4$

This indicates a position just before the 1st data point. Therefore, the 5% percentile is closest to the lowest value which is (4).

Page No.:
Date:

NOTE :-

To find the value of a datapoint that falls between two datapoints, we would interpolate betwn them.

For Example :- 2.5 position value means the percentile value is halfway between 2nd and 3rd data point, so take average of these two datapoints (value) to get the answer.

For Example : If the position value comes out to be 2.7? This means we're 70% of the way between the second and third data point.
Here's how to calculate percentile :-

Let say,
Value at position 2 = V_2
Value at position 3 = V_3

Then the interpolated value at position 2.7 is =

$$V = V_2 + 0.7(V_3 - V_2)$$
$$\Rightarrow V = 6 + 0.7(10 - 6)$$
$$= 6 + 0.7 \times 4$$
$$= 8.8$$

∴ The percentile value would be around 8.8.

4. IQR (Interquartile Range)

$IQR = Q_3 - Q_1$ = 50% of data from our dataset.

↳ IQR is less sensitive to extreme value as it finds the central 50% of data leaving first and last 25% called as extreme values.

5. VARIANCE

↳ Variance measures how spread out the values in a dataset are from the mean.

↳ Variances are the average of the squared differences from the mean.

↳ FORMULA :-

For a dataset with value $x_1, x_2, x_3, \dots, x_n$ and mean μ.

- POPULATION VARIANCE ;

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

- SAMPLE VARIANCE :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where,

- $\mu \equiv$ Population mean
- $\bar{x} \equiv$ Sample mean
- $n \equiv$ no. of data points.

x_i →

* Example : Dataset :- [4, 7, 10]

- Mean, $\mu = \left(\frac{4+7+10}{3}\right) = 7$
- Squared Difference ⇒ $(4-7)^2 = 9$
  $(7-7)^2 = 0$
  $(10-7)^2 = 9$

∴ Population Variance = $\sigma^2 = \left(\frac{9+0+9}{3}\right) = \frac{18}{3} = 6$ (Ans)

6. Standard Deviation (Std)

- S.D. = $\sqrt{\text{variance}}$
- The unit of standard deviation is same as of the dataset while the unit of variance is squared of dataset. That's why most of the time we prefer calculating S.D over variance.

Frequency

↳ no. of times a value of data occurs.
↳ We prefer to calculate Frequency in case of Categorical variables.
↳ We create Frequency Distribution Table.

Page No.:
Date:

* Ex :- {1,3,3,2,4,1,2,2,1,2,3,5,4,1,2,1,3,1,4,1}

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
1	7	7/20
2	5	5/20
3	4	4/20
4	3	3/20
5	1	1/20
	20	

RELATIVE FREQUENCY :-

↳ Percentage or portion of data values present in complete dataset.

↳ FREQUENCY X 100
Total number of OBSERVATIONS.

CUMULATIVE FREQUENCY :-

• Cumulative Frequency is the running total of Frequencies in a dataset.

Example : Customer complaints at a company —
Imagine working in a customer service department of an e-commerce company and you've collected data on types of complaints received over a month.
You have categorizes the complain into main 4-Types :-

Page No.:
Date:

Shipping Delays - 15 complaints
Product Quality - 10 complaints
Billing Issues - 8 complaints
Returns - 12 complaints

∴ CUMULATIVE FREQUENCY TABLE :-

COMPLAINT TYPE	COMPLAINT FREQUENCY	CUMULATIVE FREQUENCY
SHIPPING DELAYS	15	15
PRODUCT QUALITY	10	15+10 = 25
BILLING ISSUES	8	25+8 = 33
RETURNS	12	33+12 = 45