

Ch-5 Statistical Representⁿ of Data

→ Statistical Description of Data.

- Descriptive statistics is a branch of statistics that involves summarizing, organizing & presenting data meaningfully and concisely.

1. Central Tendency

= Mean, Mode, Median, Midrange

2. Dispersion

= range, variance, standard deviation.

3. Shape of distribution

= skewness, kurtosis.

4. Distributive Measure.

= sum(), count(), max(), min()

partⁿ data into subsets & merge values obtained for each subset.

5. Algebraic Measures

= average() or mean()

computed as $\text{sum}() / \text{count}()$

6. Holistic Measure

= median()

• computed on entire dataset as a whole

Mean, Median, Mode

Empirical Relation

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

___/___/___

→ Example 1

Suppose you have scores of 20 students in exam.

= 85, 90, 75, 92, 88, 79, 83, 95, 87, 91,
78, 86, 89, 94, 82, 80, 84, 93, 88, 81.

① Mean

= add all values / no. of values

$$= 1770 / 20$$

$$= 1720 / 20$$

$$= 88.5$$

$$86$$

② Median

= arrange in ascending & find middle value

$$= 86.5$$

③ Mode

= identify value occurring most frequently.

$$= 88$$

④ Range, Midrange = avg $\frac{75+95}{2} = 85$

= diff = highest value - lowest value

$$= 95 - 75 = 20$$

⑤ Variance

$$= \frac{[(85-88.5)^2 + (90-88.5)^2 + \dots + (81-88.5)^2]}{20}$$

$$= 33.25 \quad 30.7$$

⑥ Standard deviation = $\sqrt{\text{Variance}} = \sqrt{33.25} = 5.77 = 5.5$

→ Measuring Dispersion of Data

① Dispersion / Variance

- Degree to which numerical data tend to spread.

② Range

$$\text{diff} = \text{max value ()} - \text{min value ()}$$

③ Quartile

1st $Q_1 = 25^{\text{th}}$ percentile
3rd $Q_3 = 75^{\text{th}}$ "

④ IQR = Interquartile Range.
 $Q_3 - Q_1$

⑤ Five-number summary

min, Q_1 , M , Q_3 , max

⑥ Outlier

value higher or lower than $1.5 \times \text{IQR}$

→ Measuring Dispersion by Boxplot Analysis.

- Way of visualizing distribution.
- Box plot = works on 5-number summary.
- Data is represented with a box.

- Ends of box = 1st and 3rd quartile.
- i.e = height of box = IQR
- line in box = median
- 2 lines extending to Max & Min = whiskers.

→ Variance & Standard Deviation

- N observations x_1, x_2, \dots, x_N

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

where x_i = data value

\bar{x} / μ = mean of data.

- Variance = σ^2

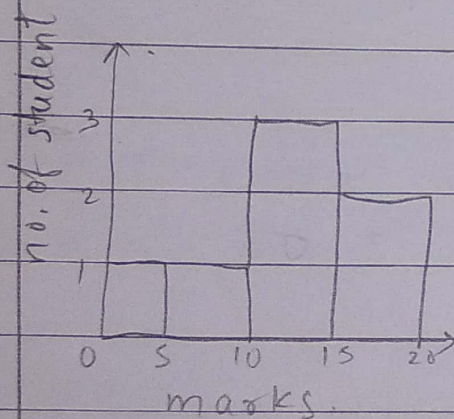
- Standard Deviation = $\sqrt{\sigma^2} = \sigma$

NOTE $\sigma = 0$, no spread of data, all obs have same value
otherwise, $\sigma > 0$, always.

→ Histogram Analysis

Histogram

- Used for Numeric attributes.
- It depicts the frequency distribution of variables.
- eg = height of individual (ranging from shorter to taller heights)
- Here, there is ^{no} space between bars.



Bar Graph

- Used for Categorical attributes.
- It is used to compare categorical variables.
- eg = types of fruits = job titles.
- Here, there is space between bars.

