SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT II – DESCRIBING VARIABILITY

Describing Variability

EMPATHY:

- Variability refers to the **divergence of data** from its mean value and is commonly used in the statistical and financial sectors.

- Variability shows how spread out the scores are in a distribution.

Scores: **10, 10, 10, 10** → no variability
Scores: **5, 10, 15, 20** → more variability

- Variability can be measured with the **range**, the interquartile range and the standard deviation.

- In each case, variability is determined by measuring distance.

Range

- The range is the total distance covered by the distribution, from the highest score to the lowest score (using the upper and lower real limits of the range).

Range=Maximum value - Minimum value

<u>Merits :</u>

a) It is easier to compute.

b) It can be used as a measure of variability where precision is not required.

<u>Demerits :</u>

a) Its value depends on only two scores

b) It is not sensitive to total condition of the distribution.

DEFINE:

Variance

- Variance is the expected value of the squared deviation of a random variable from its mean.

- In short, it is the measurement of the distance of a set of random numbers from their collective average value.

- Variance is used in statistics as a way of better understanding a data set's distribution.

- Variance is calculated by finding the square of the standard deviation of a variable.

$\sigma 2 = \Sigma(X - \mu)^2 / N$

IDEATE:

- In the formula above,

μ - represents the mean of the data points,
x - is the value of an individual data point and
N - is the total number of data points.

- Data scientists often use variance to better understand the distribution of a data set.

- Machine learning uses variance calculations to make generalizations about a data set, aiding in a neural network's understanding of data distribution.

- Variance is often used in conjunction with probability distributions.

TESTING:

Standard Deviation

- Standard deviation is simply the square root of the variance.

- Standard deviation measures the standard distance between a score and the mean.

$$\text{Standard deviation} = \sqrt{\text{Variance}}$$

- The standard deviation is a measure of how the values in data differ from one another or how spread-out data is.

- There are two types of variance and standard deviation in terms of sample and population.

- The standard deviation measures how far apart the data points in observations are from each.

- we can calculate it by subtracting each data point from the mean value and then finding the squared mean of the differenced values; this is called Variance.

- The square root of the variance gives us the standard deviation.

PROTOTYPING:

Example 1: The heights of animals are: 600 mm, 470 mm, 170 mm, 430 mm and 300 mm. Find out the mean, the variance and the standard deviation.

Solution:

Mean = 600+ 470 + 170+ 430 + 300 / 5

=1970 /5= 394

$\sigma 2= \Sigma(X - \mu)^2 / N$

Variance = $(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-394)^2$ /5

Variance = 42436+5776+ 50176 + 1296 +8836 / 5

Variance = 21704

Standard deviation = $\sqrt{Variance} = \sqrt{21704}$

= 142.32 ≈ 142