**SNS COLLEGE OF ENGINEERING**

Kurumbapalayam (po), Coimbatore – 641 107

Accredited by NAAC-UGC with 'A' Grade

Approved by AICTE & Affiliated to Anna University, Chennai

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**19AD504 – DATA VISUALIZATION**

**UNIT – II**

**DATA VISUALIZING**

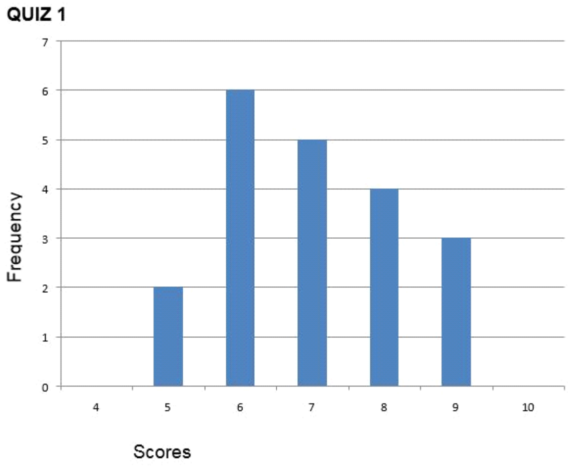
## 2.2 Measure of Variability

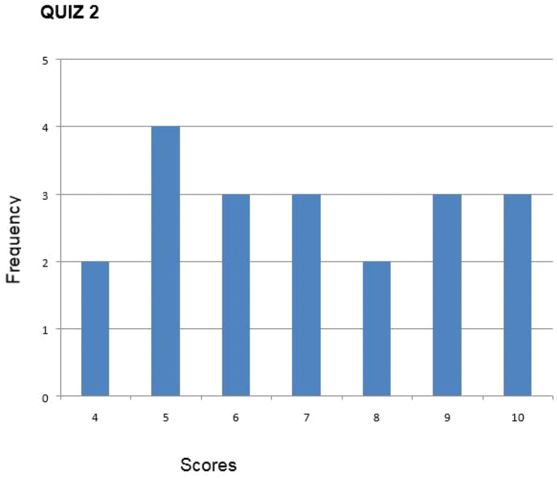
- A breakdown statistic that shows the amount of distribution in a data set is known as a measure of variability. The mean is a common tool used by analysts to describe the centre of a population or a process.
- Although the mean is important, variation frequently generates stronger reactions in people. Values in a dataset are more consistently distributed when a distribution has fewer variability.
- The data points are more varied and extreme values are more likely when the variability is bigger. As a result, awareness of variability improves in understanding the possibility of uncommon events.
- There are four frequently used measures of the variability of a distribution:

- **Range**
- **Interquartile range**
- **Variance**
- **Standard deviation.**

For example, distributions with the same mean can have different amounts of variability or dispersion.

In the  following two histograms, the distribution of scores for Quiz 1 and Quiz 2 are presented. Despite the equal means (the mean score for both quizzes is 7), the scores on Quiz 1 are more packed or clustered around the mean, whilst the scores on Quiz 2 are more spread out. Thus, the differences within the student group were greater on Quiz 2 than on Quiz 1.





## (i) Range :

- The most basic measure of variation is the range, which is the distance from the smallest to the largest value in a distribution.

**Range= Largest value – Smallest Value**

For the distribution of scores of Quiz 1 and Quiz 2, the range is:

Quiz 1. Range = 9-5= 4

Quiz 2. Range = 10-4= 6

which shows (like the histograms above) that Quiz 2 scores have greater spread than Quiz 1 scores.

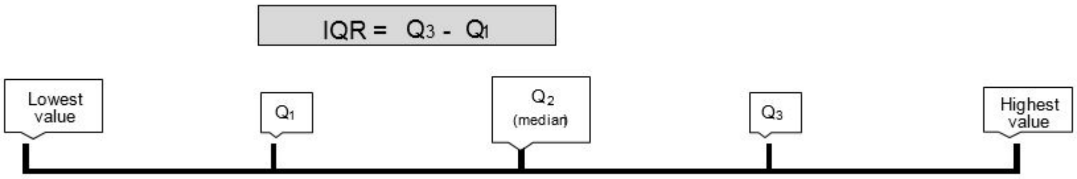
However, the range uses only two values in the data set, and one of these values may be an unusually large or small value.

## (ii) Interquartile range :

- The interquartile range (IQR) is the range of the middle 50% scores in a distribution:

  **IQR= 75th percentile – 25th percentile**

- It is based on dividing a data set into quartiles. Quartiles are the values that divide scores into quarters.
- Q1 is the lower quartile and is the middle number between the smallest number and the median of a data set.
- Q2 is the middle quartile-or median.
- Q3 is the upper quartile and is the middle value between the median set and the highest value of a data set.
- The interquartile range formula is the first quartile subtracted from the third quartile, interquartile range, lowest value, median, and highest value.



**For Quiz 1,** Q3 is 8 and Q1 is 6 .

- These are the scores : 5, 6, 7, 8, 9
- If the median is 7, then Q1 is 6 (middle value between median and lowest value) and Q3 is 8 (middle value between median and highest value).

To calculate the IQR:

**IQR= 8-6= 2**

**For Quiz 2,** Q3 is 9 and Q1 is 5. These are the scores:4, 5, 6, 7, 8, 9, 10

- The median is 7. To find Q1, we'll look at the lower half section of the distribution of scores: 4,5,6.
- Q1 is the median of this section of the distribution : 5
- To find Q2, we'll look at the upper half section of the distribution of scores: 8, 9,10.

- Q3 is the median of this section of the distribution: 9.

To calculate the IQR**,** knowing Q1 and Q3:

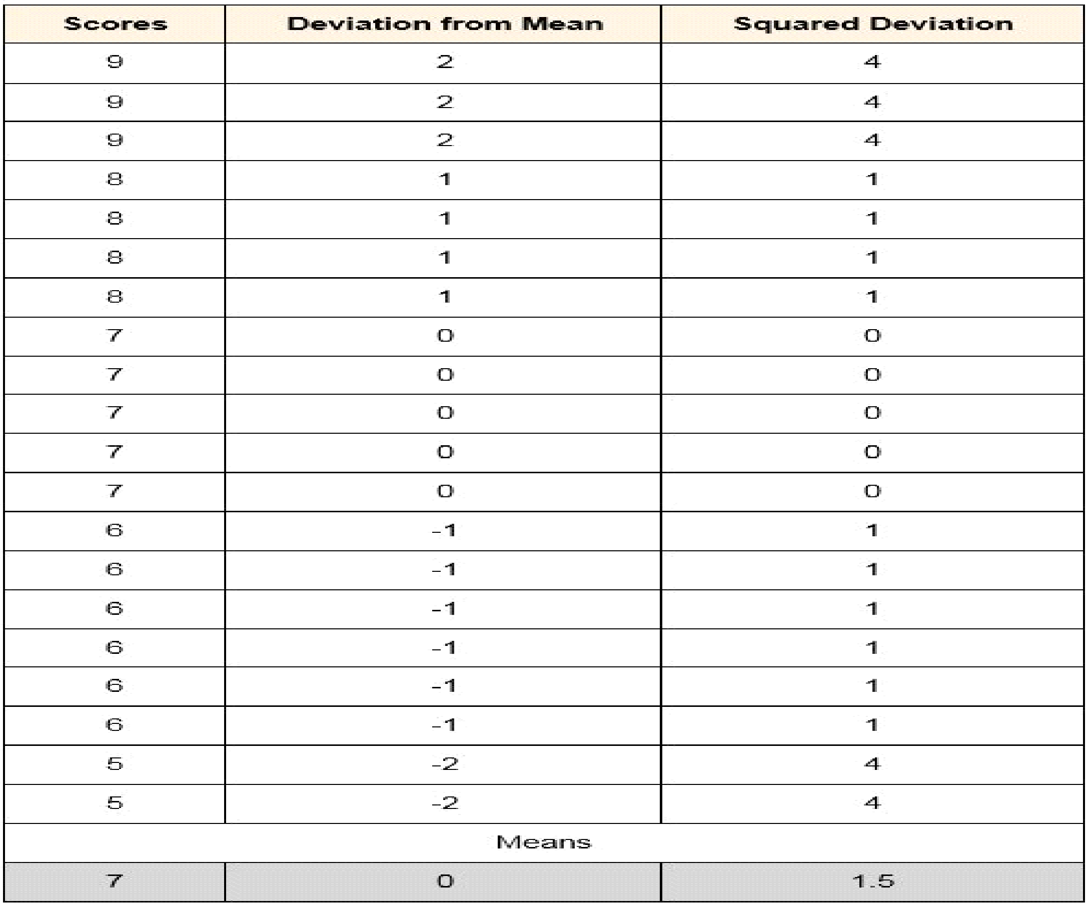
**IQR= 9-5= 4**

## (iii) Variance :

The variance is the average squared difference of the scores from the mean. To compute the variance in a population:

- Calculate the mean
- Subtract the mean from each score to compute the deviation from mean score
- Square each deviation score (multiply each score by itself)
- Add up the squared deviation score to give the sum
- Divide the sum by the number of scores

The table below contains students' scores on a Statistics test. To calculate the variance:



| Scores | Deviation from Mean | Squared Deviation |
|---|---|---|
| 9 | 2 | 4 |
| 9 | 2 | 4 |
| 9 | 2 | 4 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 5 | -2 | 4 |
| 5 | -2 | 4 |
| Means | | |
| 7 | 0 | 1.5 |

**Example of scores to calculate variance**

- The mean is calculated: sum all scores and divide by the number of scores:   140/20= 7
- The deviation from the mean for each score is calculated. For example, for the first score: 9-7= 2- See column Deviation from the mean
- Each deviation from the mean score is squared (multiplied by itself). For the first score: 2x2= 4. See column Squared deviation.
- Finally, the mean of the squared deviations is calculated. The variance is 1.5

This is how the formula to calculate variance in a population looks like:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

Where,

- σ2 is the variance
- µ is the mean of a population
- X are the values or scores
- N is the number of values or scores

➢ If the variance in a sample is used to estimate the variance in a population, it is important to note that samples are consistently less variable than their populations:

➢ The sample variability gives a biased estimate of the population variability.

➢ This bias is in the direction of underestimating the population value.

➢ In order to adjust this consistent underestimation of the population variance, we divide the sum of the squared deviation by N-1 instead of N.

**Formula to calculate variance in a sample is:**

$$s^2 = \frac{\sum(X - M)^2}{N - 1}$$

Where,

- s2 is the variance of the sample
- M is the sample mean
- X are the values or scores
- N is the number of values or scores in the sample

## (iv) Standard deviation :

- The standard deviation is the average amount by which scores differ from the mean. The standard deviation is the square root of the variance, and it is a useful measure of variability when the distribution is normal or approximately normal (see below on the normality of distributions).
- The proportion of the distribution within a given number of standard deviations (or distance) from the mean can be calculated.
- A small standard deviation coefficient indicates a small degree of variability (that is, scores are close together); larger standard deviation coefficients indicate large variability (that is, scores are far apart).

**The formula to calculate the standard deviation is**

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

**Example : how to calculate the standard deviation:**

- In the previous section- Variance- we computed the variance of scores on a Statistics test by calculating the distance from the mean for each score, then squaring each deviation from the mean, and finally calculating the mean of the squared deviations.
- Since we already know the variance, we can use it to calculate the standard deviation. To do so, take the square root of the variance. The square root of 1.5 is 1.22. The standard deviation is 1.22.