Class 1

35 35 36 35 36

$\bar{x} = \frac{35+35+36+35+36}{5}$

$\bar{x} = \frac{177}{5} = 35.4$

dataset : series 1

Class 2

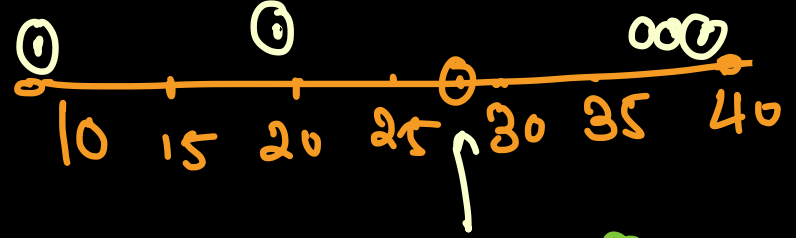
10 20 35 39 40

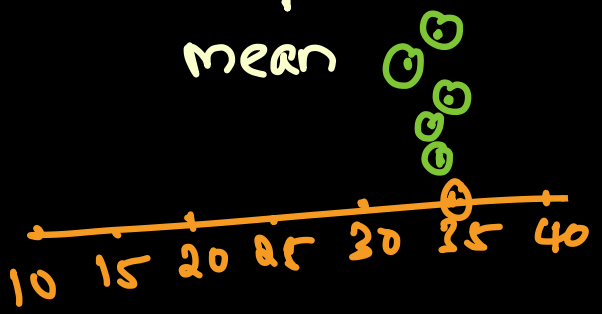
$\bar{x} = 29$

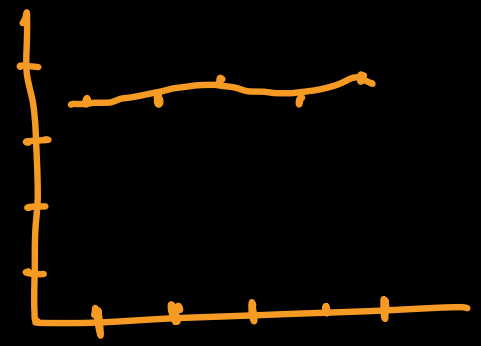
series 2

Measures of Dispersion

Variations → difference between every value & mean





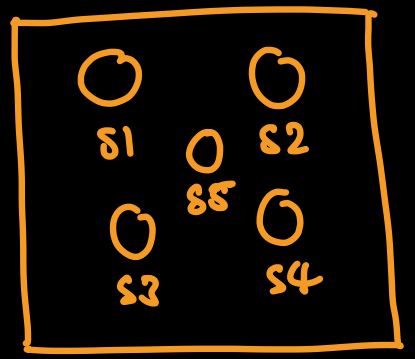


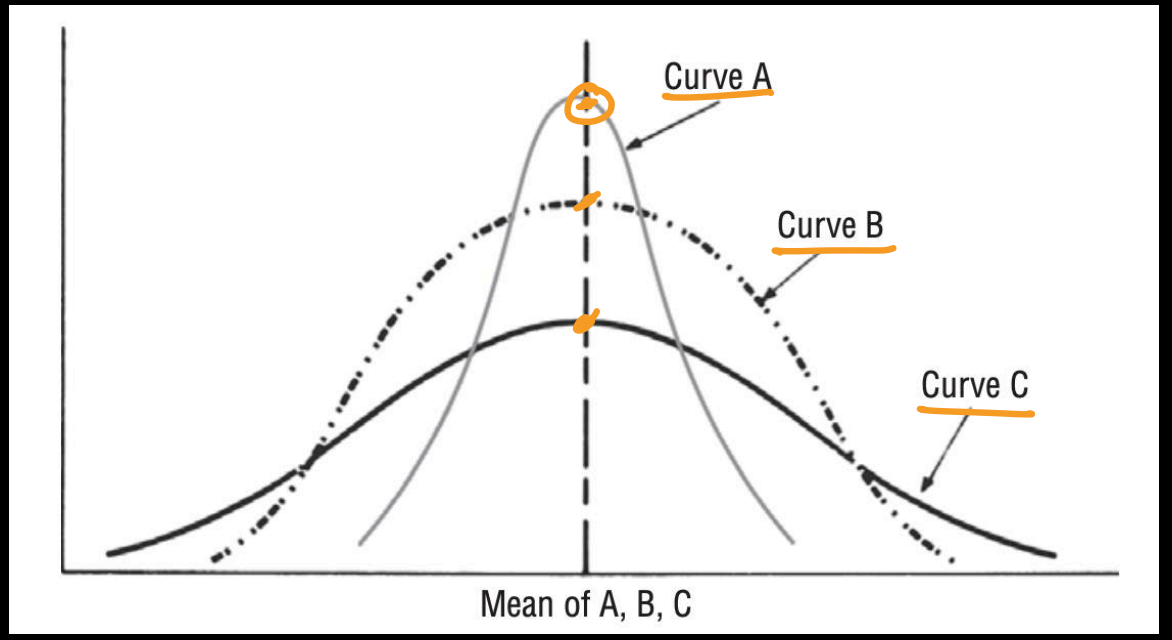
Introduction

- Dispersion is the measure of the variations of the items
- The degree to which numerical data tend to spread about an average value is called as dispersion or variation of the data
- It is also known as scatter, variation or spread
- An average is more meaningful when it is examined in the light of dispersion

choose dataset with minimum variations

sample





Properties of good measure of variation

- It should be simple to understand
- It should be easy to compute
- It should be rigidly defined → formula
- It should be based on each and every item of data
- It should be amenable to further algebraic treatment → numeric
- It should have sampling stability
- It should not be unduly affected by extreme values ↳ outliers

Methods of studying Variation

- The Range
- The Interquartile Range
- Quartile Deviation
- Mean Deviation or Average Deviation
- Standard Deviation ***
- Variance
- Lorenz Curve

Objectives of Dispersion

- **Comparative study**
  - Measures of dispersion give a single value indicating the degree of consistency or uniformity of distribution
  - This single value helps us in making comparisons of various distributions
  - The smaller the magnitude (value) of dispersion, higher is the consistency or uniformity and vice-versa
- **Reliability of an average**
  - A small value of dispersion means low variation between observations and average
  - It means that the average is a good representative of observation and very reliable
  - A higher value of dispersion means greater deviation among the observations. In this case, the average is not a good representative and it cannot be considered reliable
- **Control the variability**
  - Different measures of dispersion provide us data of variability from different angles, and this knowledge can prove helpful in controlling the variation
  - Especially in the financial analysis of business and medicine, these measures of dispersion can prove very useful
- **Basis for further statistical analysis**
  - Measures of dispersion provide the basis for further statistical analysis like computing correlation, regression, test of hypothesis, etc.

Types of Dispersion

- **Absolute measures**
  - Absolute measures of dispersion are expressed in the unit of variable itself, like kilograms, rupees, centimeters, marks etc.
  - E.g.
    - Range, Interquartile Range
    - Quartile Deviation
    - Mean Deviation

- **Relative measures**
  - Relative measures of dispersion are obtained as ratios or percentages of the average
  - These are also known as coefficients of dispersion
  - These are pure numbers or percentages that are totally independent of the units
  - E.g.
    - Coefficient of range
    - Coefficient of mean deviation

Range

Largest value - smallest value

Range

- Simplest method of studying dispersion
- It is the difference between the value of smallest item and the value of largest item of distribution

Range = Largest item – Smallest item

5 10 15 3 7 9 20

Range = Largest - Smallest

Range = 20 - 3 = 17

Coefficient of Range = $\frac{L-S}{L+S}$

$= \frac{17}{23} = 0.78$

Coefficient of Range

- The relative measure corresponding to range is called as coefficient of range is obtained by using following formula

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

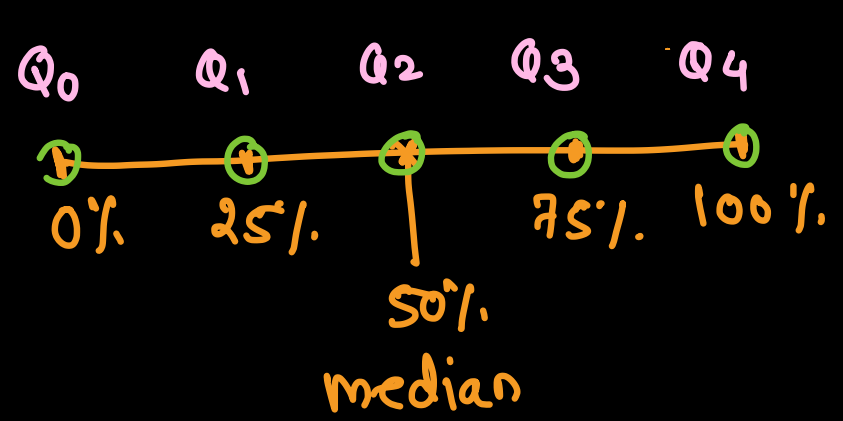
- If the average of two distributions are about the same, a comparison of the range indicates that the distribution with smaller range has less dispersion

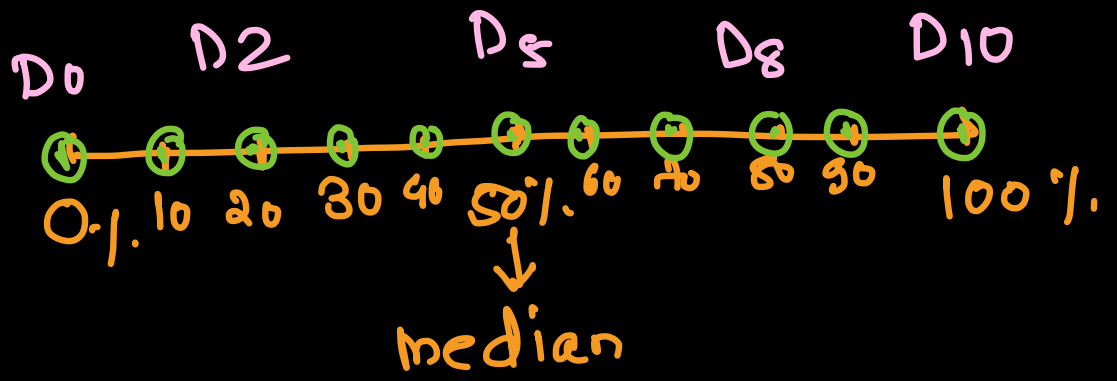
Continuous Series

- There are two methods of determining the range from data grouped into a frequency distribution
- Method 1
  - Find the difference between upper limit of the highest class and lower limit of the lowest class
- Method 2
  - Find the difference between mid point of highest class and mid point of lowest class
- In practice, both the methods are used

Merits and Limitations

- Merits
  - Amongst all the methods, Range is the simplest to understand and easiest to compute
  - It take minimum time to calculate the range value
- Limitations
  - It is not based on each and every item of the distribution
  - It is subject to fluctuations of considerable magnitude from sample to sample
  - Range can not tell us anything about the character of distribution within the two extreme observations
  - Range can not be calculated in open-end distributions





Interquartile Range

median = Q_2, D_5, P_{50}

Quartile = dividing dataset in 4 equal parts

Decile = dividing dataset in 10 equal parts

percentile = dividing dataset in 100 equal parts

Computation of Quartile, Decile and Percentile

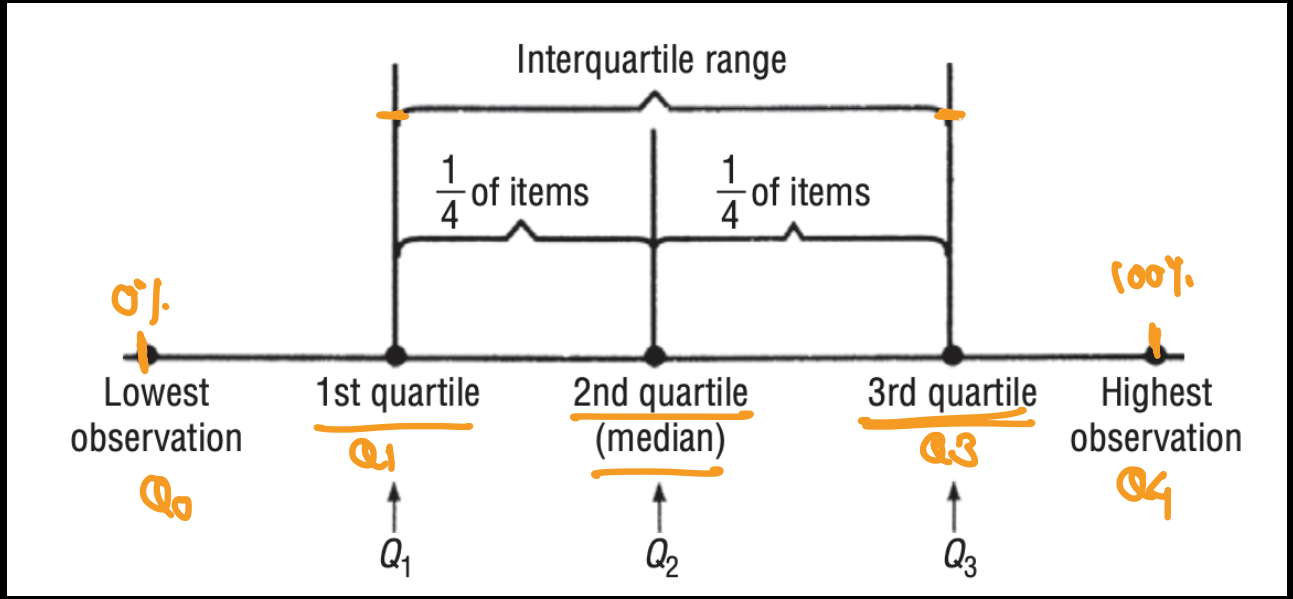
- The procedure for computing quartile, decile and percentile is same as calculating median
- While computing these values in individual and discrete we add 1 to N whereas in continuous series we do not add 1
- Then
  - Q1 = item at ((N + 1) / 4)th item for individual and (N/4)th item for continuous series
  - Q3 = item at (3(N + 1) / 4)th item for individual and (3N/4)th item for continuous series

  - D1 = item at ((N + 1) / 10)th item for individual and (N/10)th item for continuous series
  - D6 = item at (6(N + 1) / 10)th item for individual and (6N/10)th item for continuous series

  - P30 = item at (30(N + 1) / 100)th item for individual and (30N/100)th item for continuous series
  - P90 = item at (90(N + 1) / 100)th item for individual and (90N/100)th item for continuous series

Interquartile Range

- It is also known as Quartile Deviation
- The interquartile range measures approximately how far from the median we must go on either side before we can include one-half the values of the data set
- To compute this range, we divide our data into four parts, each of which contains 25 percent of the items in the distribution
- The quartiles are then the highest values in each of these four parts, and the interquartile range



Interquartile Range

5 7 9 10 15 17 18 15 20

$Q_1 = (N+1)/4 = 10/4 = 2.5^{th} = \frac{(7+9)}{2} = 8$

$Q_2 = 2(N+1)/4 = 2 \times 10/4 = 5^{th} = 15$

$Q_3 = 3(N+1)/4 = 7.5^{th} = \frac{15+19}{2} = \frac{37}{2} = 18.5$

$Q_4 = 20$

$IQR = Q_3 - Q_1 = 18.5 - 8 = 10.5$

Interquartile Range = Q3 – Q1

0% 25% 50% 75% 100%

2 8 9 10 12

Q_0 Q_1 Q_2 Q_3 Q_4

smallest largest

- Very often the IQR is reduced to form of Semi-interquartile range or quartile deviation by dividing by 2

$QD = \frac{Q_3 - Q_1}{2} = \frac{10.5}{2} = 5.25$

$$\text{Quartile Deviation (QD)} = \frac{Q3 - Q1}{2}$$

- QD is an absolute measure, the relative measure of the same is: Coefficient of QD

$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{10.5}{18.5 + 8} = 0.37$

$$\text{Coefficient of QD} = \frac{Q3 - Q1}{Q3 + Q1}$$

Merits and Limitations

- **Merits**
  - In certain respects, it is superior to range as a measure of dispersion
  - It has a special utility in measuring variation in case of open end distributions or one in which the data may be ranked but measured quantitatively
  - It is also useful in erratic or badly skewed distributions, where the other measures of dispersion would be warped by extreme values. The quartile deviation is not affected by the presence of extreme values
- **Limitations**
  - It is not capable of mathematical manipulation
  - Its value is very much affected by sampling fluctuations
  - It is, in fact, not a measure of dispersion as it really does not show the scatter around an average but rather a distance on a scale, i.e., quartile deviation is not itself measured from an average, but it is a positional average

Deviations

Mean Deviation – Individual Series

- Steps
  - Compute the mean of the series
  - The deviations of items from mean ignoring signs and denote these deviations by |D|
  - Obtain the total of these deviations, i.e., $\sum|D|$
  - Divide the total obtained by the number of observations

$$\text{Mean Deviation} = \frac{\sum|D|}{N}$$

$$\text{Coefficient of MD} = \frac{mean\ deviation}{mean}$$

x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$
1	−2	2
2	−1	1
3	0	0
4	1	1
5	2	2
	0	$6 = \Sigma(x-\bar{x})$

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

$$\text{mean deviation} = \frac{\Sigma\lvert x-\bar{x}\rvert}{N}$$

$$= \frac{6}{5} = 1.2$$

mean deviation = 1.2

$$\text{Coefficient of MD} = \frac{\text{MD}}{\text{mean}}$$

$$= \frac{1.2}{3} = 0.4$$

Mean Deviation – Discrete Series

- Steps
  - Calculate the mean of the series
  - Take the deviations of the items from mean ignoring signs and denote them by |D|
  - Multiply these deviations by the respective frequencies and obtain the total
  - Divide the total obtained by the number of observations

$$\text{Mean Deviation} = \frac{\sum f|D|}{N}$$

$$\text{Coefficient of MD} = \frac{mean\ deviation}{mean}$$

Mean Deviation – Continuous Series

- Steps
  - Calculate the mean of the series
  - Take the deviations of the items from mean ignoring signs and denote them by |D|
  - Multiply these deviations by the respective frequencies and obtain the total
  - Divide the total obtained by the number of observations

$$\text{Mean Deviation} = \frac{\sum f|D|}{N}$$

$$\text{Coefficient of MD} = \frac{mean\ deviation}{mean}$$

Merits and Limitations

- **Merits**
  - It is based on each and every item of the data
  - Change in the value of any item would change the mean deviation
  - Mean deviation is less affected by extreme observations
  - Since the deviations are taken from the central, comparison about formation of different distributions can easily be made
- **Limitations**
  - This method may not give very accurate results
  - It is rarely used in the studies

Standard Deviation

- The standard deviation concept was introduced by Karl Pearson in 1823
- It is by far the most important and widely used measure of studying dispersion
- Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion
- Standard deviation is also known as root mean square deviation for the reason that it is the square root of the mean of the squared deviation from the arithmetic mean
- Standard deviation is denoted by the small Greek letter σ (read as sigma)
- The standard deviation measures the absolute dispersion (or variability of distribution; the greater the amount of dispersion or variability), the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean

Standard Deviation

- A small standard deviation means a high degree of uniformity of the observation as well as homogeneity of a series; a large standard deviation means just the opposite
- Thus, if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean
- Hence, standard deviation is extremely useful in judging the representativeness of the mean

Mean Deviation vs Standard Deviation

- Algebraic signs are ignored while calculating mean deviation whereas in the calculation of standard deviation, signs are taken into account
- Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviation of items from arithmetic mean is the least

Standard Deviation – Individual Series

- In case of individual observations, standard deviation may be computed by applying any of the following two methods:
  - By taking deviations of the items from the actual mean
  - By taking deviations of the items from the assumed mean

- Steps (using actual mean)
  - Calculate the actual mean of the series
  - Take the deviations of the items from the mean
  - Square these deviations and obtain the total $\sum(x - \bar{x})^2$
  - Divide total by the total number of observations, i.e., N and extract the square root

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum(x-\bar{x})^2}{N}}$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
	0	$10 = \Sigma (x-\bar{x})^2$

$$\text{Standard deviation}(\sigma) = \sqrt{\frac{\Sigma (x-\bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{10}{5}} = \sqrt{2} = \underline{\underline{1.41}}$$

$$\text{variance} = \sigma^2 = 2$$

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

Standard Deviation – Individual Series

- Steps (using assumed mean)
  - Take the deviations of the items from an assumed mean and denote these deviations by d
  - Take the total of these deviations
  - Square these deviations and obtain the total
  - Apply the formula

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum(x-A)^2}{N} - \left(\frac{\sum(x-A)}{N}\right)^2}$$

Standard Deviation – Discrete Series

- Using Actual Mean

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}}$$

- Using Assumed Mean

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Standard Deviation – Continuous Series

- Using Actual Mean

$$\sigma = \sqrt{\frac{\sum f(m-\bar{x})^2}{N}}$$

- Using Assumed Mean

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Mathematical Properties of SD

- The sum of the squares of the deviations of items in the series from their Arithmetic mean is minimum
- Standard deviation enables us to determine a great deal of accuracy, where the values of frequency distribution are located
- Relation between the deviations
  - Standard Deviation = (3/2) * Quartile Deviation
  - Standard Deviation = (5/4) * Mean Deviation

Applications of Standard Deviation

- Standard deviation is used to measure the variability of values in a data set
- It has a wide range of applications in academia, business, and science, including:
  - **Academic Studies** (coefficient of variation, hypothesis testing, confidence intervals)
  - **Business** (variability of delivery times, inventory, etc.)
  - **Finance** (such as variability of returns in different asset classes)
  - **Forecast Accuracy** (such as weather)
  - **Manufacturing** (quality control, precision machining of parts to ensure proper size)
  - **Medicine** (effectiveness of drugs in pharmaceutical trials)
  - **Polling** (margin of error for opinion polls)
  - **Population Traits** (height, weight, IQ, etc.)

Coefficient of Variation

- The Standard deviation is absolute measure of dispersion
- The relative measure is known as the coefficient of variation
- Developed by Karl Pearson is the most commonly used measure of relative variation
- Used in problems where we want to compare variability of two or more than two series
- That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous
- On the other hand, the series for which coefficient of variation is less is said to be less variable or more consistent, more uniform, more stable or more homogeneous
- Coefficient of variation is denoted by C.V. and is obtained as follows:

$$\text{C.V.} = \frac{\sigma}{\bar{x}} * 100$$

Variance

- The term 'variance' was used to describe the square of the standard deviation by R.A. Fisher in 1913.
- The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in their original series
- Variance is defined as follows:

$$\text{Variance} = \frac{\sum(x-\bar{x})^2}{N}$$

Variance = σ^2

Merits and Limitations

- **Merits**
  - It is possible to find combined standard deviation of two series, which is not possible with any other measure
  - For comparing the variability of two or more distributions, coefficient of variation is considered to be most appropriate and this is based on mean and standard deviation
  - Standard deviation is most prominently used in further statistical work. For example, in computing skewness, correlation, etc., use is made of standard deviation
  - It is keynote in sampling and provides a unit of measurement for the normal distribution
- **Limitations**
  - It gives more weight to extreme items and less to those which are near the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small.