No. Biostat -8
Date:25.01.2009

MEASURES OF DISPERSION, RELATIVE STANDING AND SHAPE

Dr. Bijaya Bhusan Nanda,
M. Sc (Gold Medalist) Ph. D. (Stat.)
Topper Orissa Statistics & Economics Services, 1988
bijayabnanda@yahoo.com

CONTENTS

➢ What is measures of dispersion?

➢ Why measures of dispersion?

➢ How measures of dispersions are calculated?

  ➢ Range

  ➢ Quartile deviation or semi inter-quartile range,

  ➢ Mean deviation and

  ➢ Standard deviation.

  ➢ Methods for detecting outlier

➢ Measure of Relative Standing

➢ Measure of shape

LEARNING OBJECTIVE

➢ They will be able to:

➢ describe the homogeneity or heterogeneity of the distribution,

➢ understand the reliability of the mean,

➢ compare the distributions as regards the variability.

➢ describe the relative standing of the data and also shape of the distribution.

What is measures of dispersion? (Definition)

- Central tendency measures do not reveal the variability present in the data.

- Dispersion is the scattered ness of the data series around it average.

- Dispersion is the extent to which values in a distribution differ from the average of the distribution.

Why measures of dispersion? (Significance)

- Determine the reliability of an average
- Serve as a basis for the control of the variability
- To compare the variability of two or more series and
- Facilitate the use of other statistical measures.

Dispersion Example

- Number of minutes 20 clients waited to see a consulting doctor

  Consultant Doctor
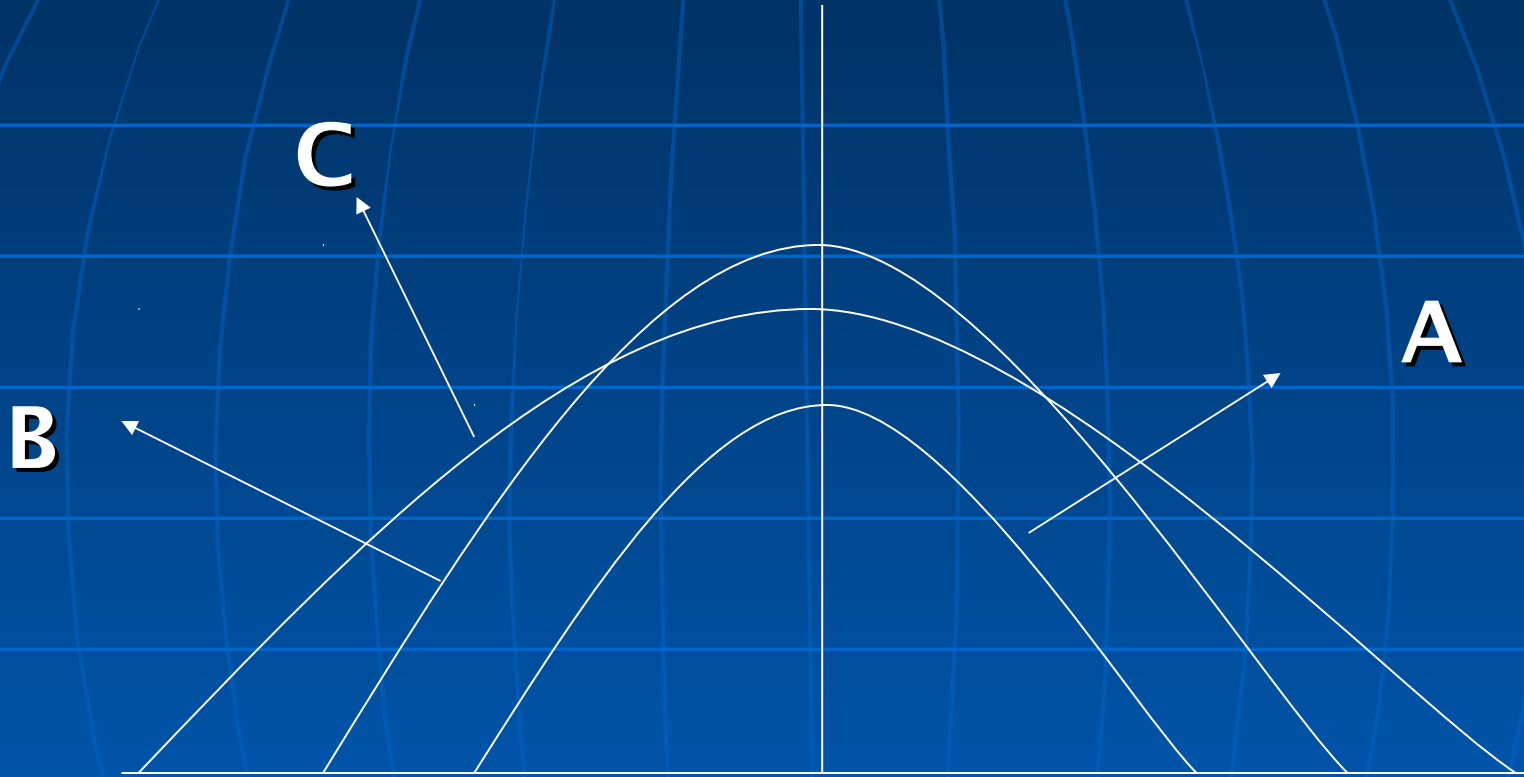
X		Y__	
05	15	15	16
12	03	12	18
04	19	15	14
37	11	13	17
06	34	11	15

- X:Mean Time – 14.6 minutes

- Y:Mean waiting time 14.6 minutes

- What is the difference in the two series?

X: High variability, Less consistency.
Y: Low variability, More Consistency

Frequency curve of distribution of three sets of data

Characteristics of an Ideal Measure of Dispersion

1. It should be rigidly defined.

2. It should be easy to understand and easy to calculate.

3. It should be based on all the observations of the data.

4. It should be easily subjected to further mathematical

5. It should be least affected by the sampling fluctuation .

6. It should not be unduly affected by the extreme values.

How dispersions are measured?

➢ **Measure of dispersion:**

➢ **Absolute: Measure the dispersion in the original unit of the data.**

➢ **Variability in 2 or more distr^n can be compared provided they are given in the same unit and have the same average.**

➢ **Relative: Measure of dispersion is free from unit of measurement of data.**

➢ **It is the ratio of a measaure of absolute dispersion to the average, from which absolute deviations are measured.**

➢ **It is called as co-efficient of dispersion.**

➢ The following measures of dispersion are used to study the variation:
  ➢ The range
  ➢ The inter quartile range and quartile deviation
  ➢ The mean deviation or average deviation
  ➢ The standard deviation

Range:

The difference between the values of the two extreme items of a series.

Example:

Age of a sample of 10 subjects from a population of 169subjects are:

$$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9 \quad X_{10}$$

42 28 28 61 31 23 50 34 32 37

The youngest subject in the sample is 23years old and the oldest is 61 years, The range: $R = X_L - X_s$

$$= 61\text{-}23 = 38$$

Co-efficient of Range:

$R = (X_L - X_S) / (X_L + X_S)$

$= (61 - 23) / (61 + 23) = 38 / 84 = 0.452$

Characteristics of Range

- Simplest and most crude measure of dispersion
- It is not based on all the observations.
- Unduly affected by the extreme values and fluctuations of sampling.
- The range may increase with the size of the set of observations though it can decrease
- Gives an idea of the variability very quickly

Percentiles, Quartiles (Measure of Relative Standing) and Interquartile Range

- Descriptive measures that locate the relative position of an observation in relation to the other observations are called measures of relative standing.

- They are quartiles, deciles and percentiles

- The quartiles & the median divide the array into four equal parts, deciles into ten equal groups, and percentiles into one hundred equal groups.

- Given a set of n observations X_1, X_2, …. X_n, the pth percentile 'P' is the value of X such that 'p' per cent of the observations are less than and 100 –p per cent of the observations are greater than P.

- 25th percentile = 1st Quartile i.e., Q_1

- 50th percentile = 2nd Quartile i.e., Q_2

- 75th percentile = 3rd Quartile i.e., Q_3
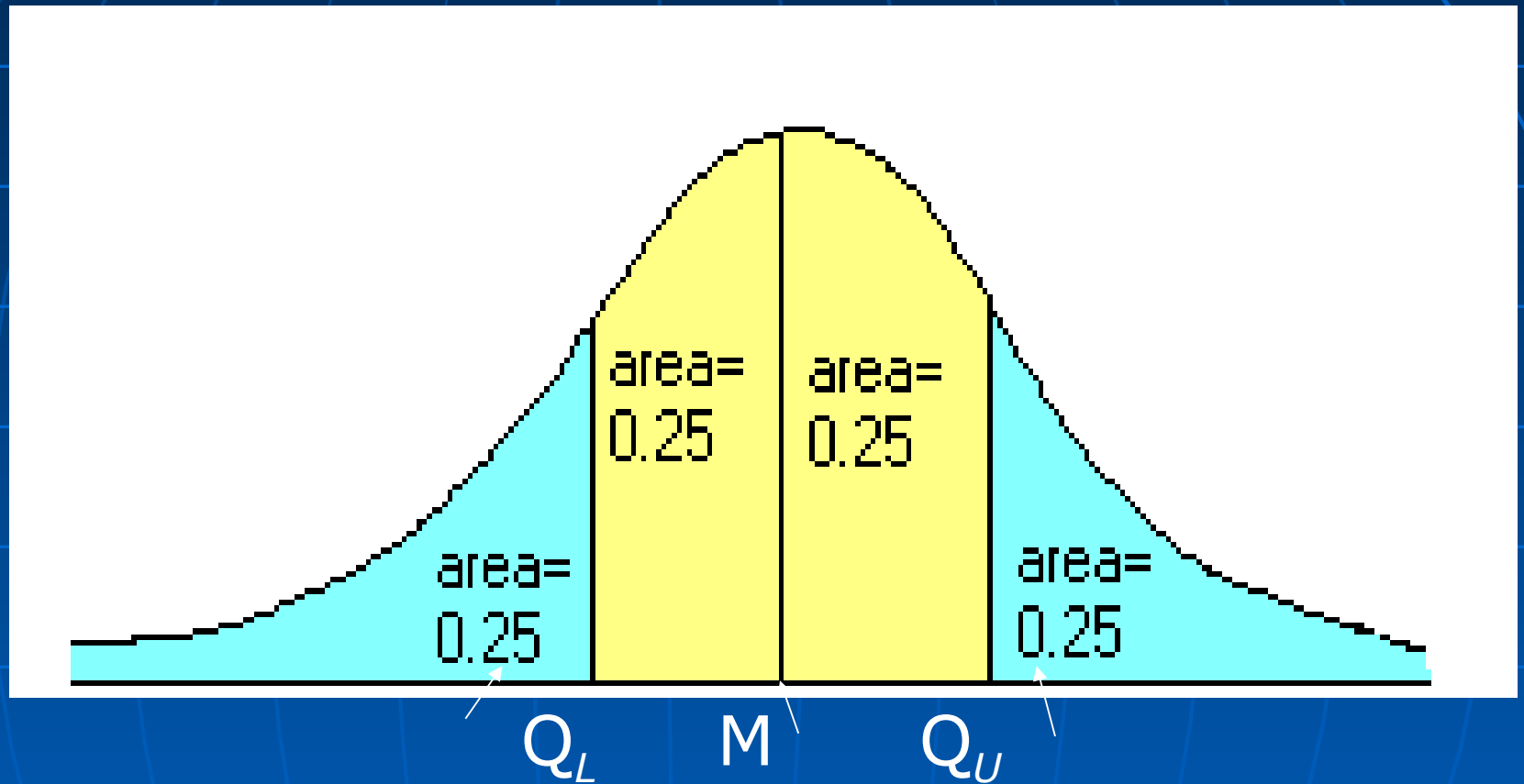
Figure 8.1 *Locating of lower, mid and upper quartiles*

$$Q_1 = \frac{n+1}{4} \text{ th ordered observation}$$

$$Q_2 = \frac{2(n+1)}{4} \text{ th ordered observation}$$

$$Q_3 = \frac{3(n+1)}{4} \text{ th ordered observation}$$

Interquartile Range (IQR): The difference between the 3rd and 1st quartile.

IQR = Q_3 – Q_1

Semi Interquartile Range:= (Q_3 – Q_1)/ 2

Coefficient of quartile deviation:
(Q_3 – Q_1)/(Q_3 + Q_1)

Interquartile Range

Merits:

➢ It is superior to range as a measure of dispersion.

➢ A special utility in measuring variation in case of open end distribution or one which the data may be ranked but measured quantitatively.

➢ Useful in erratic or badly skewed distribution.

➢ The Quartile deviation is not affected by the presence of extreme values.

Limitations:

➢ As the value of quartile deviation dose not depend upon every item of the series it can't be regarded as a good method of measuring dispersion.

➢ It is not capable of mathematical manipulation.

➢ Its value is very much affected by sampling fluctuation.

➢ Another measure of relative standing is the **z-score for an observation (or standard score).**

➢ It describes how far individual item in a distribution departs from the mean of the distribution.

➢ Standard score gives us the number of standard deviations, a particular observation lies below or above the mean.

Standard score (or z -score) is defined as follows:

For a population z-score $= \dfrac{X - \mu}{\sigma}$

where X = the observation from the population

μ the population mean, σ = the population s.d

For a sample z-score $= \dfrac{X - \overline{X}}{s}$

where X = the observation from the sample

$\overline{X}$ the sample mean, s = the sample s.d

Mean Absolute Deviation (MAD) or Mean Deviation (MD)

- **The average of difference of the values of items from some average of the series (ignoring negative sign), i.e. the arithmetic mean of the absolute differences of the values from their average .**

Note:

1. MD is based on all values and hence cannot be calculated for open-ended distributions.

2. It uses average but ignores signs and hence appears unmethodical.

3. MD is calculated from mean as well as from median for both ungrouped data using direct method and for continuous distribution using assumed mean method and short-cut-method.

4. The average used is either the arithmetic mean or median

Computation of Mean absolute Deviation

For individual series: $X_1, X_2, \ldots\ldots X_n$

$$M.A.D = \frac{\sum |X_i - \overline{X}|}{n}$$

For discrete series: $X_1, X_2, \ldots\ldots X_n$ & with corresponding frequency $f_1, f_2, \ldots\ldots f_n$

$$M.A.D = \frac{\sum f_i |X_i - \overline{X}|}{\Sigma f_i}$$

$\overline{X}$: Mean of the data series.

Computation of Mean absolute Deviation:

For continuous grouped data: $m_1, m_2, \ldots m_n$ are the class mid points with corresponding class frequency $f_1, f_2, \ldots f_n$

$$\text{M.A.D} = \frac{\sum f_i |m_i - \bar{X}|}{\Sigma f_i}$$

$\bar{X}$: **Mean of the data series.**

Coeff. Of MAD: = (MAD /Average)

The average from which the Deviations are calculated. It is a relative measure of dispersion and is comparable to similar measure of other series.

Example:

Find MAD of Confinement after delivery in the following series.

Days of Confinement (X)	No. of patients (f)	Total days of confinement of each group Xf	Absolute Deviation from mean $\|X - \overline{X}\|$	$f_i\|X_i - \overline{X}\|$
6	5	30	1.61	8.05
7	4	28	0.61	2.44
8	4	32	1.61	6.44
9	3	27	2.61	7.83
10	2	20	3.61	7.22
Total	**18**	**137**		**31.98**

$\overline{X}$ = Mean days of confinement = 137 / 18 = 7.61

MAD=31.98 / 18=1.78, Coeff.of MAD= 1.78/7.61=0.233

Find the MAD of weight and coefficient of MAD of 470 infants born in a hospital in one year from following table.

Weight in Kg	2.0-2.4	2.5-2.9	3.0-3.4	3.5-3.9	4.0-4.4	4.5+
No. of infant	17	97	187	135	28	6

Merits and Limitations of MAD

- **Simple to understand and easy to compute.**
- **Based on all observations.**
- **MAD is less affected by the extreme items than the Standard deviation.**
- **Greatest draw back is that the algebraic signs are ignored.**
- **Not amenable to further mathematical treatment.**
- **MAD gives us best result when deviation is taken from median. But median is not satisfactory for large variability in the data. If MAD is computed from mode, the value of the mode can not be determined always.**

Standard Deviation (σ)

It is the positive square root of the average of squares of deviations of the observations from the mean. This is also called root mean squared deviation (σ) .

For individual series: x_1, x_2, ... x_n

$$\sigma = \sqrt{\frac{\Sigma (x_i - \overline{x})^2}{n}} \qquad \sigma = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

For discrete series: X_1, X_2, ... X_n & with corresponding frequency f_1, f_2, ... f_n

$$\sigma = \sqrt{\frac{\Sigma f_i (x_i - \overline{x})^2}{\Sigma f_i}} \qquad \sigma = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2}$$

Standard Deviation (σ) Contd.

For continuous grouped series with class midpoints : $m_1, m_2, \ldots\ldots m_n$ & with corresponding frequency $f_1, f_2, \ldots\ldots f_n$

$$\sigma = \sqrt{\frac{\sum f_i(m_i - \overline{x})^2}{\sum f_i}}$$

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

Variance: It is the square of the s.d

Coefficient of Variation (CV): Corresponding Relative measure of dispersion.

$$CV = \frac{\sigma}{\overline{X}} \times 100$$

Characteristics of Standard Deviation:

- SD is very satisfactory and most widely used measure of dispersion

- Amenable for mathematical manipulation

- It is independent of origin, but not of scale

- If SD is small, there is a high probability for getting a value close to the mean and if it is large, the value is father away from the mean

- Does not ignore the algebraic signs and it is less affected by fluctuations of sampling

- SD can be calculated by :
  - Direct method
  - Assumed mean method.
  - Step deviation method.

- It is the average of the distances of the observed values from the mean value for a set of data
- Basic rule --More spread will yield a larger SD

Uses of the standard deviation

- The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located in relation to the mean.

- **Chebyshev's Theorem**
  - For any data set with the mean 'μ' and the standard deviation 's' at least **75%** of the values will fall within the 2σ interval and at least **89%** of the values will fall within the 3σ interval of the mean

TABLE: Calculation of the standard deviation (σ)

Weights of 265 male students at the university of Washington

Class-Interval (Weight)	f	d	fd	fd²
90-99	1	-5	-5	25
100-109	1	-4	-4	16
110-119	9	-3	-27	81
120-129	30	-2	-60	120
130-139	42	-1	-42	42
140-149	66	0	0	0
150-159	47	1	47	47
160-169	39	2	78	156
170-179	15	3	45	135
180-189	11	4	44	176
190-199	1	5	5	25
200-209	3	6	18	108
n =265		$\Sigma f\mathrm{d}= 99$		$\Sigma f\mathrm{d}^2 = 931$

$$\sigma = \sqrt{\frac{(\Sigma f\mathrm{d}^2)}{n} - \frac{(\Sigma f\mathrm{d})^2}{n^2}} \times (i)$$

$$= \sqrt{\frac{931}{265} - \frac{(99)^2}{265}} \times (10)$$

$$= \sqrt{(3.5132 - 0.1396)} \;(\times 10)$$

$$= (1.8367)\;(10)$$

$$= 18.37 \text{ or } 18.4$$

$d = (X_i - A)/i \qquad n = \Sigma f_i$

$A = 144.5, i = 10$

- Means, standard deviation, and coefficients of variation of the age distributions of four groups of mothers who gave birth to one or more children in the city of minneapol in: 1931 to 1935. Interprete the data

CLASSIFICATION	$\overline{X}$	σ	CV
Resident married	28.2	6.0	21.3
Non-resident married	29.5	6.0	20.3
Resident unmarried	23.4	5.8	24.8
Non-resident unmarried	21.7	3.7	17.1

Example: Suppose that each day laboratory technician *A* completes 40 analyses with a standard deviation of 5. Technician *B* completes 160 analyses per day with a standard deviation of 15. Which employee shows less variability?

Uses of Standard deviation

- **Uses of the standard deviation**
  - The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located in relation to the mean. We can do this according to a theorem devised by the Russian mathematician P.L. Chebyshev (1821-1894).

Measure of Shape

The fourth important numerical characteristic of a data set is its shape: Skewness and kurtosis.

▪ Skewness

- Skewness characterizes the degree of asymmetry of a distribution around its mean. For a sample data, the skewness is defined by the formula:

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where n = the number of observations in the sample, x_i = i^{th} observation in the sample, s= standard deviation of the sample, $\bar{x}$ = sample mean
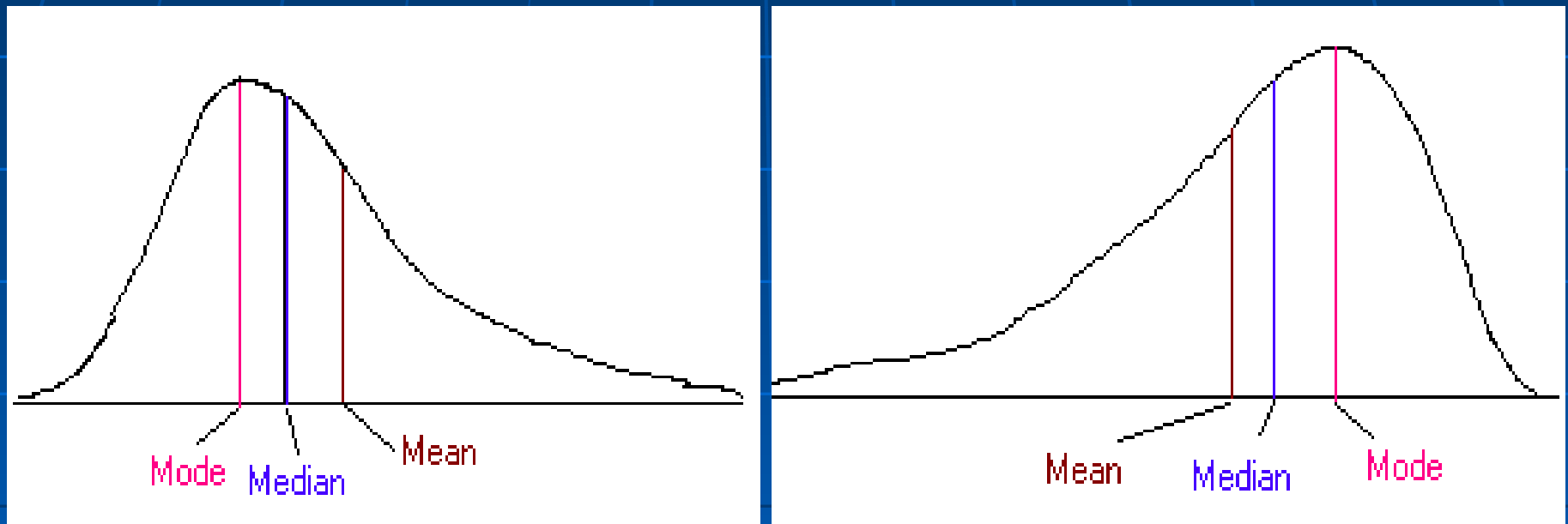
Measure of Shape

Figure 8.2 +ve or Right-skewed distribution

Kurtosis:

Kurtosis characterizes the relative peakedness or flatness of a distribution compared with the bell-shaped distribution (normal distribution).

Kurtosis of a sample data set is calculated by the formula:

$$Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Positive kurtosis indicates a relatively peaked distribution.
Negative kurtosis indicates a relatively flat distribution.

The distributions with positive and negative kurtosis are depicted in Figure 8.4 , where the distribution with null kurtosis is normal distribution.
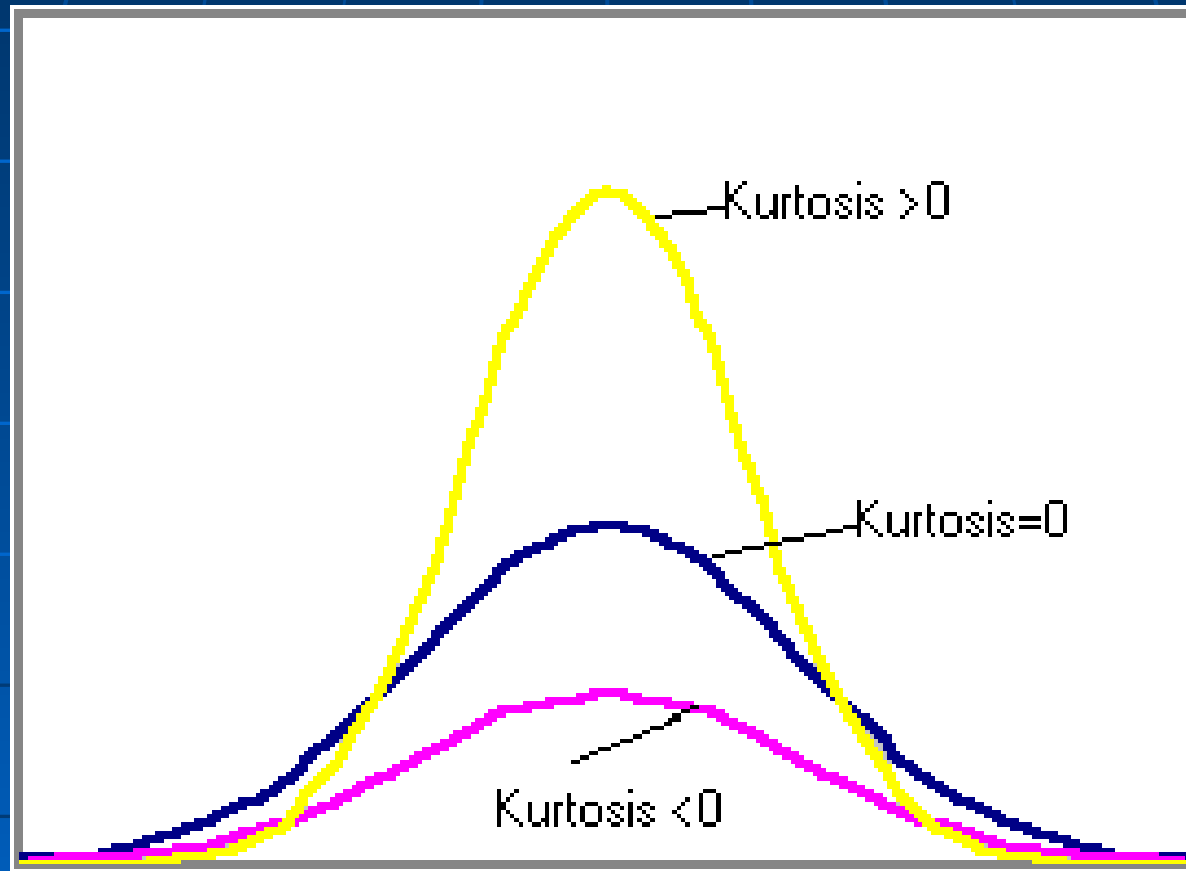
REFERENCE

1. Mathematical Statistics- S.P Gupta

2. Statistics for management- Richard I. Levin, David S. Rubin

3. Biostatistics A foundation for Analysis in the Health Sciences.

THANK YOU