2HSOE52 Introduction to Economics

Chapter 1 _A: Basic Statistical Concept

Measures of Central Tendency, Measures of Variation, Correlation

Samir K Mahajan, Ph.D. , UGC NET

A. MEASURES OF CENTRAL TENDENCY /Averages

- An average i s a single value which is considered as the most representative for a given set of data.
- Summary statistic / a statically constant that represents the center point or typical value of a dataset.
- Measures of central tendency show the tendency of some central value around which data tend to cluster.

A. 1 MEASURES OF CENTRAL TENDENCY : Types

Arithmetic Mean
Weighted Mean
Median
Mode
Geometric Mean
Harmonic Mean

A.1.1 ARITHMETIC MEAN

- **Arithmetic mean (also called mean) is defined as the sum of all the observations divided by the number of observations.**

1. For Ungrouped Data

For ungrouped data, arithmetic mean may be computed by applying any of the following methods:

(1) Direct Method

Mathematically, if $x_1, x_2, \ldots, x_n$ are the n observations then their mean is

$$\overline{X} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If f_i is the frequency of x_i (i=1, 2,..., k), the formula for arithmetic mean would be

$$\overline{X} = \frac{(f_1x_1 + f_2x_2 + \ldots + f_kx_k)}{(f_1 + f_2 + \ldots + f_k)}$$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

A.1.1. ARITHMETIC MEAN

2 For Grouped Data

Direct Method

If f_i is the frequency of x_i ($i = 1, 2, \ldots, k$) where x_i is the mid value of the i^{th} class interval, the formula for arithmetic mean would be

$$\overline{X} = \frac{(f_1x_1 + f_2x_2 + \ldots + f_kx_k)}{(f_1 + f_2 + \ldots + f_k)}$$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N},$$

where, $N = f_1 + f_2 + \ldots + f_k$

A.1.1 Mean: properties

- **Property 1: Sum of deviations of observations from their mean is zero**

$$\sum(x - \text{mean}) = 0$$

- **Property 2: Sum of squares of deviations taken from mean is least in comparison to the same taken from any other average.**

$$\Rightarrow \sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - A)^2$$

where, A is an assumed mean / Median / Mode

Merits : Can be used for further mathematical treatments

Issues: badly affected by extremely small or extremely large values;

A.1.1.1 WEIGHTED MEAN

Weight here refers to the importance of a value in a distribution. The frequency of a number can used as its weight.

∴ If x_i has a weight w_i, then weighted mean is defined as:

$$\overline{X}_W = \frac{\sum_{i=1}^{k} X_i W_i}{\sum_{i=1}^{k} W_i} \qquad \text{for all } i = 1, 2, 3, \ldots, k.$$

A.1.2 MEDIAN

- **Median is that value of the variable which divides the whole distribution into two equal parts.**
- **The data should be arranged in ascending or descending order of magnitude.**

1. Median for Ungrouped Data

Mathematically, if $x_1, x_2, \ldots, x_n$ are the n observations then for obtaining the median first of all we have to arrange these n values either in ascending order or in descending order. When the observations are arranged in ascending or descending order, the middle value gives the median if n is odd. For even number of observations there will be two middle values. So we take the arithmetic mean of these two values.

$$M_d = \left(\frac{n+1}{2}\right)^{th} \text{observation} \; ; \; (\text{when n is odd})$$

$$M_d = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2} \; ; \; (\text{when n is even})$$

A.1.2. Median: merits and issues

- **Merits: not affected by extremely small or extremely large values;**

- **Issues:**
  - **does not utilize all the observations**
  - **not amenable to algebraic treatment**
  - **affected by sampling fluctuations**

A.1.3. MODE

- **Highest frequent observation in a distribution is known as mode**

For Ungrouped Data

Mathematically, if $x_1, x_2, \ldots, x_n$ are the n observations and if some of the observation are repeated in the data, say x_i is repeated highest times then we can say the x_i would be the mode value.

A.1.3. MODE :merits and issues

- **Merits: not affected by extreme values**

- **Issues:**
  - **amenable to algebraic treatment**
  - **greatly affected by sampling fluctuations**

1.6.1 Relationship between Mean, Median and Mode

For a symmetrical distribution the mean, median and mode coincide. But if the distribution is moderately asymmetrical, there is an empirical relationship between them. The relationship is

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Note: Using this formula, we can calculate mean/median/mode if other two of them are known.

B. MEASURES OF DISPERSION

Dispersion studies scatterdness in data from the central value. Dispersion measures the extent to which values in a distribution differ from the average of the distribution.

- **Measures of dispersion determine the reliability of an average value.**
- **When variation is less, the average closely represents the individual values of the data**
- **when variation is large; the average may not closely represent all the units and be quite unreliable.**

B. MEASURES OF DISPERSION contd.

Many powerful statistical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of variance, techniques of quality control, etc. are based on different measures of dispersion

There are two basic kinds of a measure of dispersion
(i) Absolute measures (used to measure the variability in same unit)

1.Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation and Variance

(ii) Relative measures (used to compare variability of two or more sets of observations/series)

5. Coefficient of Variation

B.1.Range

Range is the simplest measure of dispersion. It is defined as the difference between the maximum value of the variable and the minimum value of the variable in the distribution. Its merit lies in its simplicity. The demerit is that it is a crude measure because it is using only the maximum and the minimum observations of variable. However, it still finds applications in Order Statistics and Statistical Quality Control. It can be defined as

$$R = X_{Max} - X_{Min}$$

where, X_{Max}: Maximum value of variable and

X_{Min}: Minimum value of variable.

B.2. Quartile Deviation

As you have already studied in Unit 1 of this block that Q_1 and Q_3 are the first quartile and the third quartile respectively. $(Q_3 - Q_1)$ gives the inter quartile range. The semi inter quartile range which is also known as Quartile Deviation (QD) is given by

$$\text{Quartile Deviation (QD)} = (Q_3 - Q_1) / 2$$

Relative measure of Q.D. known as Coefficient of Q.D. and is defined as

$$\text{Cofficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Mean Deviation

Mean deviation is defined as average of the sum of the absolute values of deviation from any arbitrary value viz. mean, median, mode, etc. It is often suggested to calculate it from the median because it gives least value when measured from the median.

The deviation of an observation x_i from the assumed mean A is defined as $(x_i - A)$.

Therefore, the mean deviation can be defined as

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$$

The quantity $\sum |x_i - A|$ is minimum when A is median.

For frequency distribution, the formula will be

$$MD = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{x}|}{\sum_{i=1}^{k} f_i}$$

$$MD = \frac{\sum_{i=1}^{k} f_i |x_i - \text{median}|}{\sum_{i=1}^{k} f_i}$$

where, all symbols have usual meanings.

Issue: Negative deviations are straightaway made positive.

B.4 Variance and Standard Deviation

Variance is the average of the square of deviations of the values taken from mean

Variance is defined as

$$\mathrm{Var}(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and for a frequency distribution, the formula is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2$$

where, all symbols have their usual meanings.

B.4 Variance and Standard Deviation contd.

2.6.3 Standard Deviation

Standard deviation (SD) is defined as the positive square root of variance. The formula is

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

and for a frequency distribution the formula is

$$SD = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{\sum_{i=}^{k} f_i}}$$

where, all symbols have usual meanings. SD, MD and variance cannot be negative.

B. 5 Coefficient of variation

Coefficient of Variation (CV) is defined as

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

It is a relative measure of variability. If we are comparing the two data series, the data series having smaller CV will be more consistent. One should be careful in making interpretation with CV. For example, the series 10, 10, 10 has SD zero and hence CV is also zero. The series 50, 50 and 50 also has SD zero and hence CV is zero. But the second series has higher mean. So we shall regard the second series as more consistent than the first.

C. Skewness and Kurtosis

Skewness means lack of symmetry in a distribution

- **In mathematics, a figure is called symmetric it divides the figure into two congruent parts i.e mirror images of each other.**
- **In Statistics, a distribution is called symmetric if mean, median and mode coincide. Otherwise, the distribution becomes asymmetric.**
- **Skewness studies presence of outliers in in one versus the other tail**

C.1.1 Symmetric Distribution : Bell Mean, Median and Mode coincide.

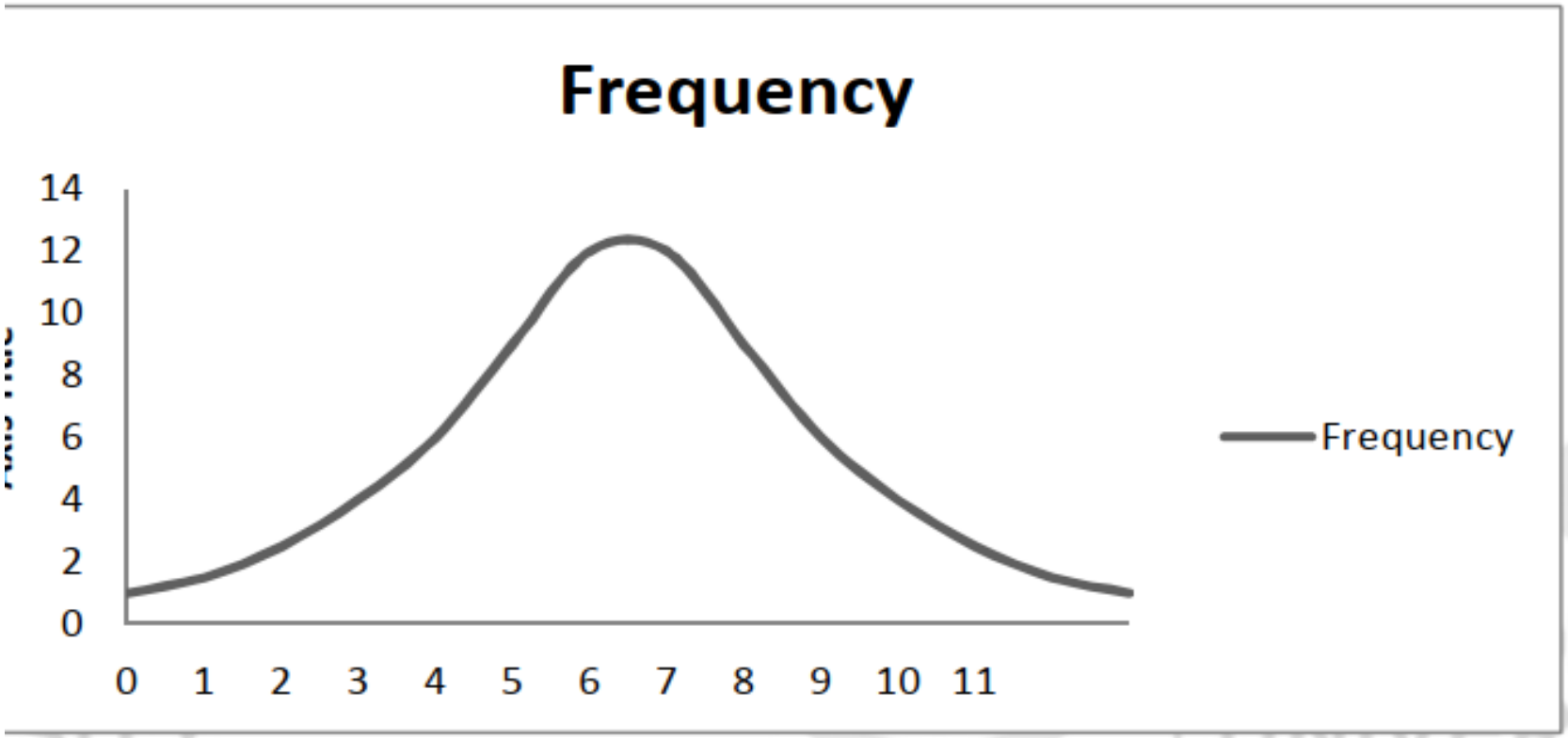


Fig. 4.1: Symmetrical Curve

C.1.2 negatively skewed distribution :mean < median < mode : longer left tail

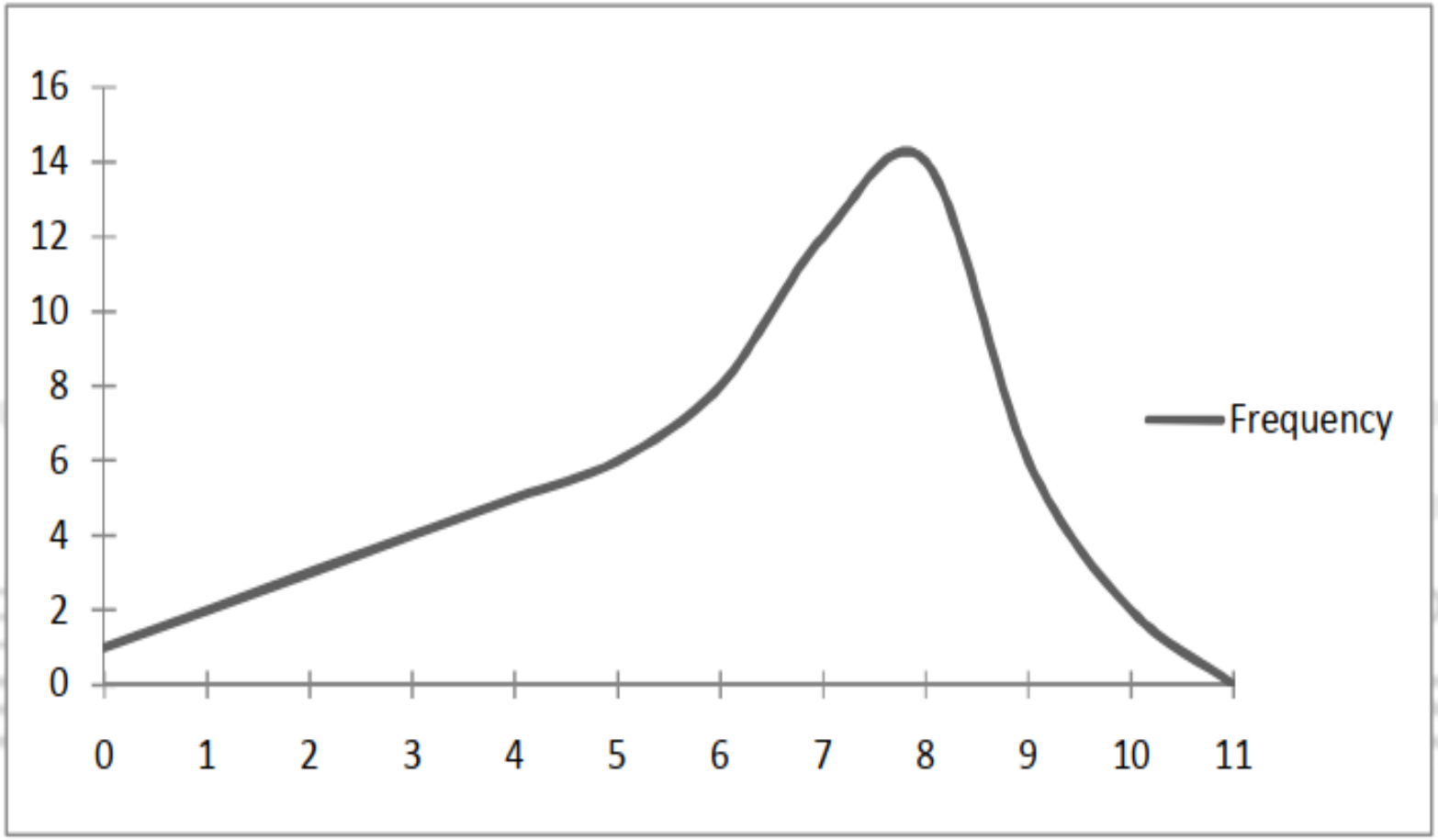


Fig. 4.2: Negative Skewed Curve

C.1.3 Positively skewed distribution : mean > median > mode : longer Right Tail

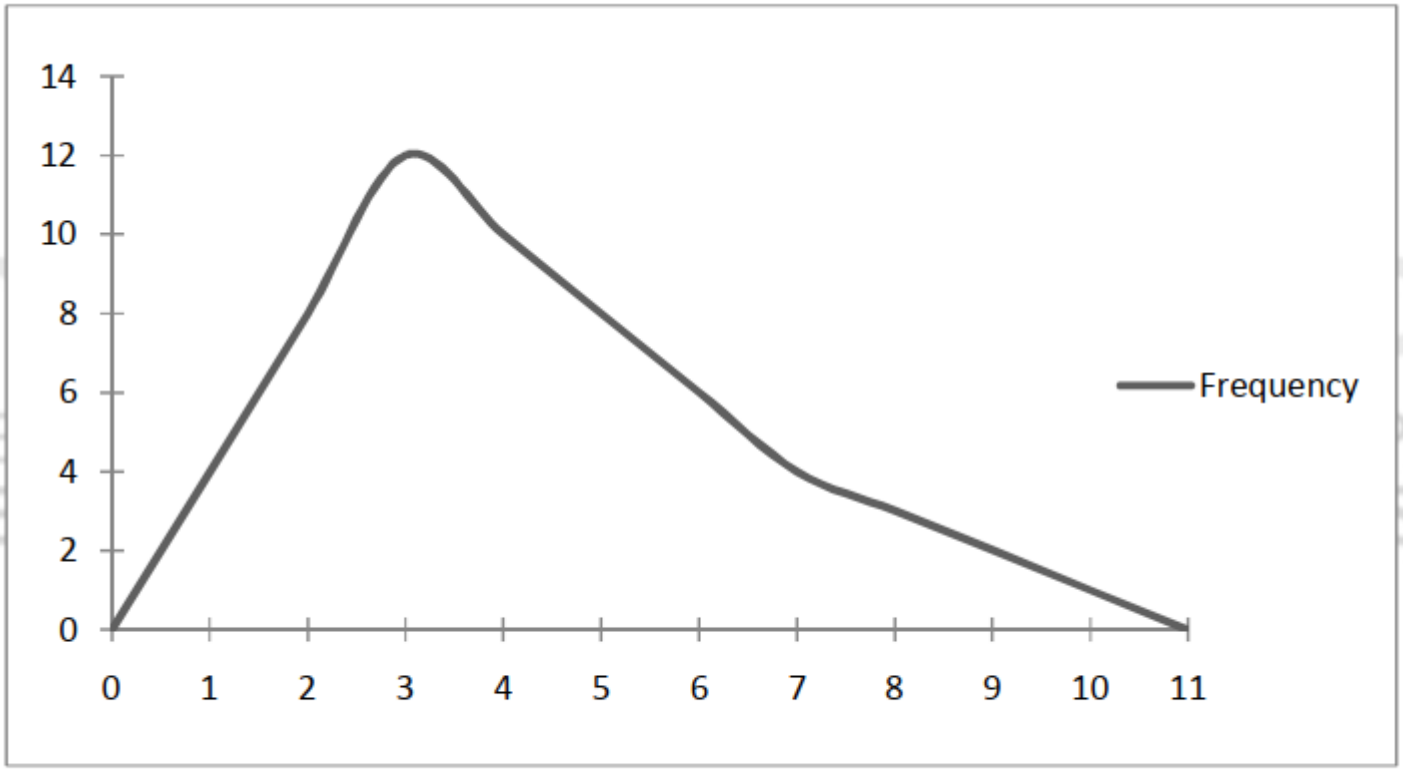
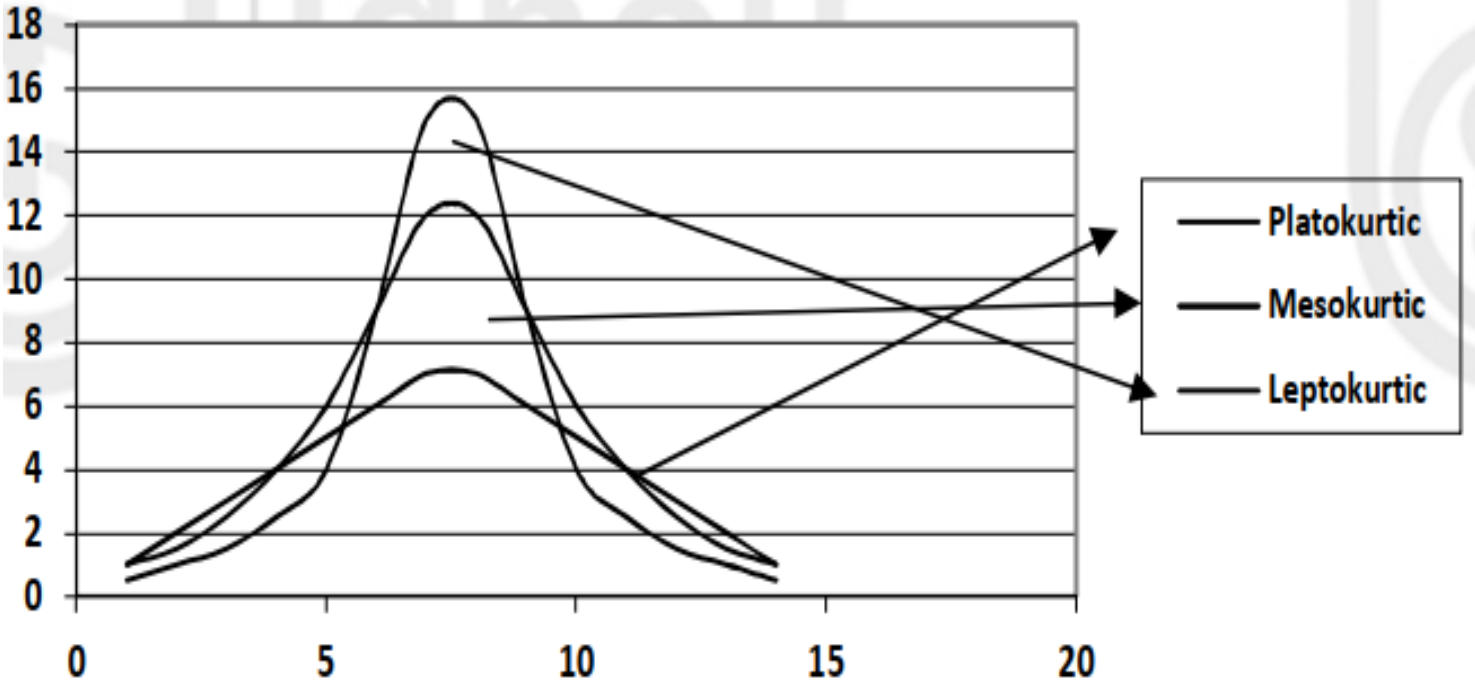


Fig. 4.3: Positive Skewed Curve

C.2 Kurtosis

- Kurtosis gives a measure of flatness /peakness of distribution or tailedness of curve
- Prof. Karl Pearson has called it the “Convexity of a Curve”.
- The degree of kurtosis of a distribution is measured relative to that of a normal curve.

C.2. 1Kurtosis: Types



4: Platykurtic Curve, Mesokurtic Curve and Leptokurtic Curve

D. Distribution with More than one Variable

- **Let us consider situation where observations are available on two or more variables such as:**
  - **Heights and weights of persons of a certain group;**
  - **Sales revenue and advertising expenditure in business;**
  - **Time spent on study and marks obtained by students in exam.**

- **In the case of two random variables, this is called a bivariate distribution.**
- **in multivariate distribution we have more than two random variables**

- **Two study relationship between the variables, following concepts work handy.**

1. **Covariance**
2. **Correlation**
3. **Regression**

D.1. Covariance

- studies how much two random variables vary together – measures variance between two variables

If x and y are two random variables, then covariance between x and y is given by

Where, $\mathrm{Cov}(x, y)$ the covariance between X and Y which is defined as:

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

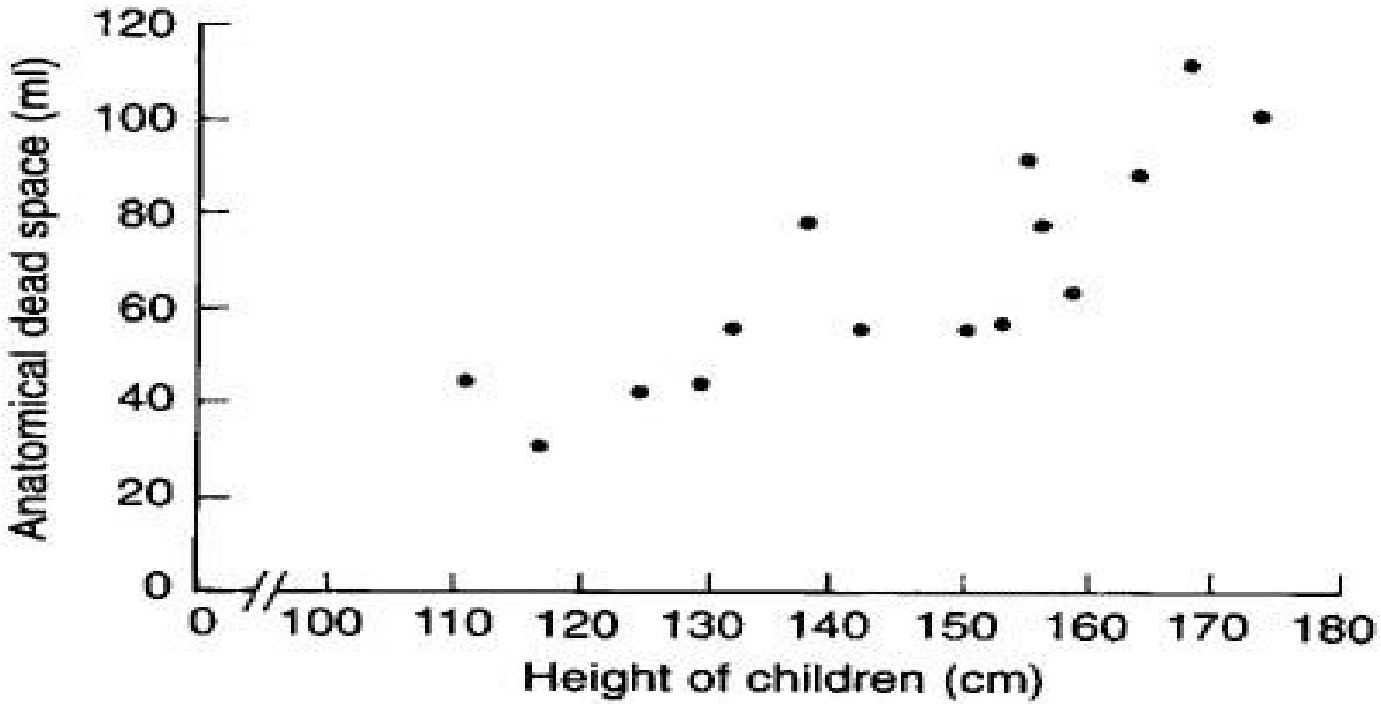
D.1. 1 Covariance: Property

If two variables are independent, then their covariance is zero. This does not always work both ways, that is it does not mean that if the covariance is zero then the variables must be independent.

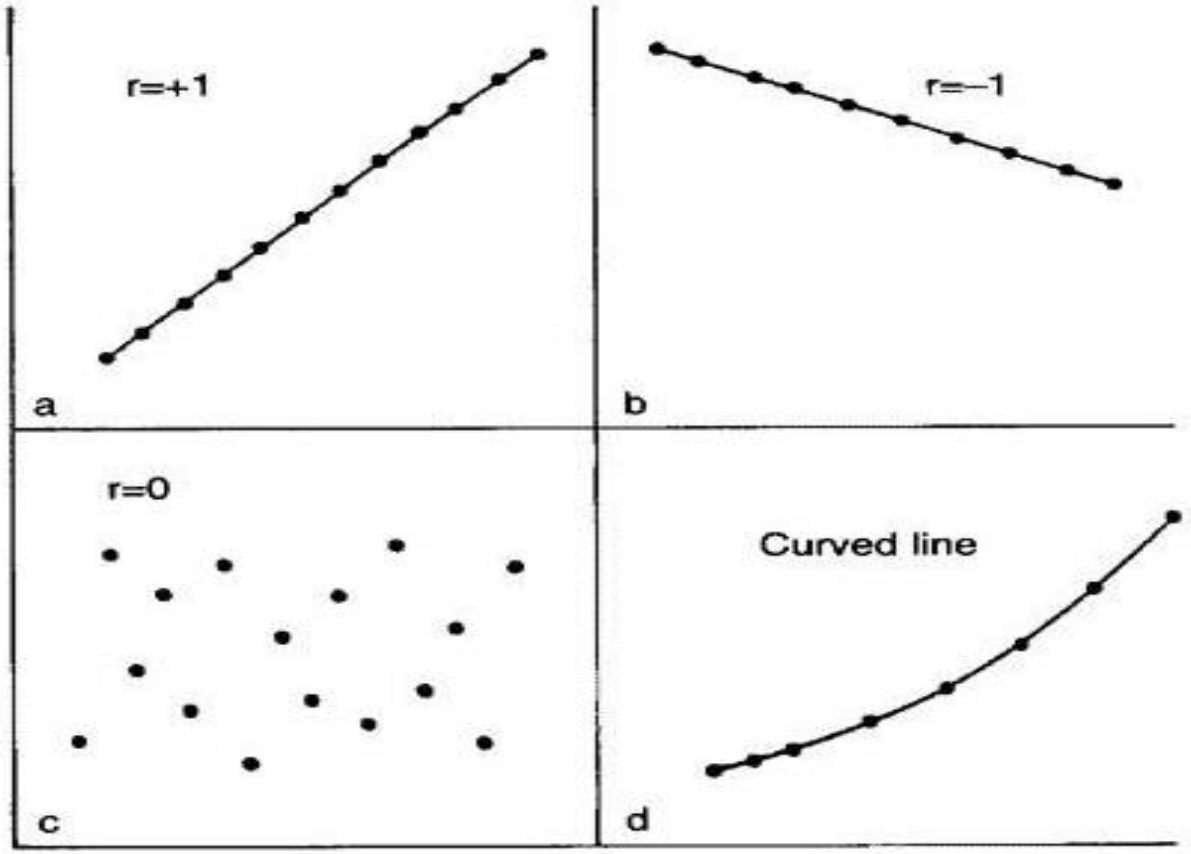
D.2 Correlation

- **Correlation is the scaled measure of covariance between variables**
- **denotes some form of association /linear relationship between variables**
  - **When values /observation of two variables are plotted on a two dimensional graph, correlation studies how closely these values fall on a line.**
- **Thus, Correlation measures the strength as well as direction of the relationship between variables.**
- **Correlation coefficient measures degree of association**
- **Correlation coefficient is unit free and its value lies between -1 and +1.**

D.2 Correlation : Scatter Diagram



D.2 Correlation contd.



D.2 Correlation : Formula .

If X and Y are two random variables then correlation coefficient between X and Y is denoted by r and defined as

$$r = Corr(x, y) = \frac{Cov(x, y)}{\sqrt{V(x)\,V(y)}} \qquad \ldots(1)$$

$Corr(x, y)$ is indication of correlation coefficient between two variables X and Y.

Where, $Cov(x, y)$ the covariance between X and Y which is defined as:

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

and $V(x)$ the variance of X, is defined as:

$$V(x) = \frac{1}{n}\sum^{n}(x_i - \bar{x})^2 ,$$

$V(y)$ the variance of Y is defined by

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where, n is number of paired observations.

D.2 Correlation contd.

$$r = Corr(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (2)$$

Karl Pearson's correlation coefficient r is also called product moment correlation coefficient. Expression in equation (2) can be simplified in various forms. Some of them are

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (3)$$

or

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\left\{\frac{\sum_{i=1}^{n}x_i^2}{n} - \bar{x}^2\right\}\left\{\frac{\sum_{i=1}^{n}y_i^2}{n} - \bar{y}^2\right\}}} \quad \ldots (4)$$

or

$$r = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left\{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right\}\left\{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right\}}} \quad \ldots (5)$$

or

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i}{\sqrt{\left\{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right\}\left\{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right\}}} \quad \ldots (6)$$

D.2 Correlation contd.

As correlation measures the degree of linear relationship, different values of coefficient of correlation can be interpreted as below:

Value of correlation coefficient	Correlation is
+1	Perfect Positive Correlation
−1	Perfect Negative Correlation
0	There is no Correlation
0 - 0.25	Weak Positive Correlation
0.75 - (+1)	Strong Positive Correlation
−0.25 - 0	Weak Negative Correlation
−0.75 - (−1)	Strong Negative Correlation

D.2 .1 Correlation : PROPERTIES

- **Property 1: Correlation coefficient lies between -1 and +1.**
- **Property 2: If X and Y are two independent variables then correlation coefficient between X and Y is zero, i.e. Corr(x, y) = 0.**
  - **Conversely it is not true. That is, if correlation between two variables is zero, we can not say they are independent always.**
  - **Zero correlation simply says that the is no linear relationship between the variables.**

D.3 Regression

- Regression analysis studies the relationships between a dependent variable and one or more independent variables
- The regression equation studies how much dependent variable changes on an average with any given change in independent variables .
- The direction in which the line slopes depends on whether the correlation is positive or negative.

Regression equation with two variables

$$y = a + b X$$

Reference

1. IGNOU Books

Inputs in these slides are exclusively collected from above sources.