BUSINESS STATISTICS

By:

Dr. Sudipta Ghosh

Assistant Professor

Dept. of Commerce (UG & PG)

Prabhat Kumar College, Contai

West Bengal, India

Chapter – 1: STATISTICAL REPRESENTATION OF DATA

DIAGRAMATIC REPRESENTATION OF DATA

Data : A statistician begins the work with the collection of data i.e. numerical facts. The data so collected are called raw materials (or raw data). It is from these raw materials, a statistician analysis after proper classification and tabulation, for the final decision or conclusion. Therefore it is undoubtedly important that the raw data collected should be clear, accurate and reliable.

Statistical Units : The unit of measurement applied to the data in any particular problem is the statistical unit. Physical units of the measurement like quintal, kilogramme, metre, hour and year, etc. do not need any explanation or definition. But in some cases statistician has to give some proper definition regarding the unit.

Types of Methods of Collection of Data :

Statistical data are usually of two types :

(i) Primary, (ii) Secondary

Data which are collected for the first time, for a specific purpose are known as primary data, while those used in an investigation, which have been originally collected by some one else, are known as secondary data.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Primary Method :

The following methods are common in use :

- **Direct Personal Observation : Under this method, the investigator collects the data personally.**
- **Indirect Oral Investigation : In this method data are collected through indirect sources.**
- **Schedules and Questionnaires : A list of questions regarding the enquiry is prepared and printed**
- **Local Reports : Only local agents or correspondents are requested to supply the estimate required.**

Secondary Method :

The main sources from which secondary data are collected are given below–

(i) Official publications by the Central and State Government, District Boards,

(ii) Reports of Committees, Commissions.

(iii)Publications by Research Institutions, Universities,

(iv) Economic and Commercial Journals.

(v) Publications of Trade Associations, Chambers of Commerce, etc.

(vi) Market reports, individual research works of Statisticians.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Editing and Scrutiny :

Secondary data should be used only after careful enquiry and with due criticism. It is advisable not to take them at their face value. Scrutiny is essential because the data might be inaccurate, unsuitable and inadequate. According to Bowley, "It is never safe to take published statistics at their face value without knowing their meanings and limitations …."

Universe or Population :

Statistics is taken in relation to a large data. Single and unconnected data is not statistics. In the field of any statistical enquiry there may be persons, items or any other similar units. The aggregate of all such similar units under consideration is called Universe or Population.

That is, for collecting the data regarding height, weight or age of the male candidates who appeared in the last H.S. Examination, the aggregate of such candidates is universe. Universe may be aggregate of items or any other similar things other than persons. The books in your college library or produced goods in a factory may be taken as Universe.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Sample :

If a part is selected out of the Universe then the selected part (or portion) is sample. It means sample is a part of the Universe.

So, suppose the screws or bulbs produced in a factory are to be tested. The aggregate of all such items is universe but it is not possible to test every item. So in such case, a part of the whole i.e., universe is taken and then tested. Now this part is known as sample.

Classification :

It is the process of arranging data into different classes or group according to resemblance and similarities. An ideal classification should be unambiguous, stable and flexible.

- **Type of Classification :**
- **There are two types of classification depending upon the nature of data.**
- **(i) Classification according to attribute – if the data is of a descriptive nature having several qualifications**
- **i.e. males, female, illiterate, etc.**
- **(ii) Classification according to class-interval if the data are expressed in numerical quantities i.e… ages of**
- **person vary and so do their heights and weights.**

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Classification:

It is the process of arranging data into different classes or group according to resemblance and similarities.

An ideal classification should be unambiguous, stable and flexible.

Type of Classification:

There are two types of classification depending upon the nature of data.

(i) Classification according to attribute – if the data is of a descriptive nature having several qualifications i.e. males, female, illiterate, etc.

(ii) Classification according to class-interval if the data are expressed in numerical quantities i.e., ages of person vary and so do their heights and weights.

Classification according to Attributes :

(i) Simple classification is that when one attribute is present i.e. classification of persons according to sex –males or female.

(ii) Manifold classification is that when more than one attributes are present simultaneously two attributes –deafness and sex. A person may be either deaf or not deaf, further a person may be a male or a female.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

DISCRETE AND CONTINUOUS SERIES :

Statistical series may be either discrete or continuous. A discrete series is formed from items which are exactly measurable, Every unit of data is separate, complete and not capable of divisions. For instance, the number of students obtaining marks exactly 10, 14, 18, 29, can easily be counted. But phenomenon like height or weight cannot be measured exactly or with absolute accuracy. So the number of students (or individuals) having height exactly 5' 2" cannot be counted. Exact height may be either 5'2" by a hundredth part of an inch. In such cases, we are to count the number of students whose heights lie between 5' 0" to 5' 2". Such series are known as 'continuous' series.

TABULATION :

- Tabulation is a systematic and scientific presentation of data in a suitable form for analysis and interpretation.
- After the data have been collected, they are tabulated i.e. put in a tabular form of columns and rows. The
- function of tabulation is to arrange the classified data in an orderly manner suitable for analysis and
- interpretation. Tabulation is the last stage in collection and compilation of data, and is a kind of steppingstone
- to the analysis and interpretation.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Types of Tabulation :

Mainly there are two types of tables – Simple and Complex. Simple tabulation reveals information regarding one or more groups of independent question, while complex table gives information about one or more interrelated questions.

FREQUENCY DISTRIBUTION

Frequency of a value of a variable is the number of times it occurs in a given series of observations. A tally sheet may be used to calculate the frequencies from the raw data (primary data not arranged in the Tabular form). A tally-mark (/) is put against the value when it occurs in the raw data.

Group Frequency Distribution :

When large masses of raw data are to be summarized and the identity of the individual observation or the order in which observations arise are not relevant for the analysis, we distribute the data into classes or categories and determine the number of individuals belonging to each class, called the class-frequency.

A tabular arrangement of raw data by classes where the corresponding class-frequencies are indicated is known as Grouped Frequency distribution.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Class-interval : In the above table, class intervals are 16-20, 21-25 …. etc. In all there are eight class-intervals. If, however, one end of class-interval is not given then it is known as open-end class. For example, less than 10, 10-20, 20-30, 30 and above. The class-interval having zero frequency is know as empty class.

Class frequency : The number of observations (frequency) in a particular class-interval is known as class-frequency. In the table, for the class-interval 26-30, class frequency is 3 and so on. The sum of all frequencies is total frequency. Here in the table total frequency is 50.

Class limits : The two ends of a class-interval are called class-limits.

Class boundaries : The class boundaries may be obtained from the class limits as follows :

Lower class-boundary = lower class limit – ½ d

Upper class-boundary = upper class limit + ½ d

Where d = common difference between upper class of any class-interval with the lower class of the next class-interval. In the table d = 1.

Mid value : (or class mark). It is calculated by adding the two class limits divided by 2.

Width : The width (or size) of a class interval is the difference between the class-boundaries (not class limits)

Width = Upper class boundary – lower class boundary

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Cumulative Frequency distribution :

As the name suggests, in this distribution, the frequencies are cumulated. This is prepared from a grouped frequency distribution showing the class boundaries by adding each frequency to the total of the previous one, or those following it. The former is termed as Cumulative frequency of less than type and the latter, the cumulative frequency of greater than type.

Histogram

This graphical method is most widely used in practice. Histogram is a series of adjacent vertical bars whose height is equal to the frequencies of the respective classes & width is equal to the class interval.

Frequency Polygon

Frequency polygon could be drawn by first drawing histogram & then joining all the midpoints of the tops(upper side) of the adjacent rectangle of the histogram by straight line graphs. The figure so obtained is called a frequency polygon. It should be noted that it is necessary to close the polygon at both ends by extending them to the base line so that it meets the X-axis at the mid points of the two hypothetical classes i.e. the class before the first class & the class after the last class having the zero frequency.

Frequency polygon could alternatively be drawn without first drawing the histogram. This could be done by plotting the frequencies of different classes (along Y-axis) against the mid values of corresponding classes (along X axis). These points are joined by straight line to get a frequency polygon. Here also this polygon would be closed at both ends by extending them to meet the X-axis.

Chapter – 1: STATISTICAL REPRESENTATION OF DATA (Contd.)

Ogive of cumulative frequency polygon: If the cumulative frequencies are plotted against the class boundaries and successive points are joined by straight lines, we get what is known as Ogive (or cumulative frequency polygon). There are two types of Ogive.

(a) Less than type – Cumulative Frequency from below are plotted against the upper class-boundaries.

(b) Greater than type – Cumulative frequencies from above are plotted corresponding lower boundaries. The former is known as less than type, because the ordinate of any point on the curve (obtained) indicates the frequency of all values less than or equal to the corresponding value of the variable reprpresented by the abscissa of the point.

Circular Diagram (or Pie diagram) : It is a pictorial diagram in the form of circles where whole area represents the aggregate and different sectors of the circle, when divided into several parts, represent the different components.

For drawing a circular diagram, different components are first expressed as percentage of the whole. Now since 100% of the centre of a circle is 360 degrees. 1% corresponds to 3.6 degrees. If p be the percentage of a certain component to the aggregate, then (p × 3.6) degrees will be the angle, which the corresponding sector subtends at the centre.

Chapter – 2: MEASURES OF CENTRAL TENDENCY

INTRODUCTION :

A given raw statistical data can be condensed to a large extent by the methods of classification and tabulation. But this is not enough. For interpreting a given data we are to depend on some mathematical measures. Such a type of measure is the measure of Central Tendency.

By the term of 'Central Tendency of a given statistical data' we mean that central value of the data about which the observations are concentrated . A central value which 'enables us to comprehend in a single effort the significance of the whole is known as Statistical Average or simply average.

The three common measures of Central Tendency are :

(i) Mean

(ii) Median

(iii) Mode

The most common and useful measure is the mean.

MEAN :

There are three types of mean :

(i) Arithmetic Mean (A.M.) (ii) Geometric Mean (G. M.) (iii) Harmonic Mean (H.M.) Of these the Arithmetic mean is the most commonly used. In fact, if not specifically mentioned by mean we shall always refer to arithmetic Mean (AM).

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

Shortcut Method (Method of assumed Mean)

In this method, the mid-value of one class interval (preferably corresponding to the maximum frequency lying near the middle of the distribution) is taken as the assumed mean (or the arbitrary origin) A and the deviation from A are calculated.

Calculation of A. M. from grouped frequency distribution with open ends

If in a grouped frequency distribution, the lower limit of the first class or the upper limit of the last class are not known, it is difficult to find the A.M. When the closed classes (other than the first and last class) are of equal widths, we may assume the widths of the open classes equal to the common width of closed class and hence determine the AM. But we can find Median or Mode without assumption.

Finding of missing frequency :

In a frequency distribution if one (or more) frequency be missing (i.e. not known) then we can find the missing frequency provided the average of the distribution is known.

Calculation of A.M. from Cumulative Frequency Distribution

At first we are to change the given cumulative frequency distribution into a general form of frequency distribution, then to apply the usual formula to compute A.M.

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

Advantages of Arithmetic Mean

(i) It is easy to calculate and simple to understand.

(ii)For counting mean, all the data are utilised. It can be determined even when only the number of items and their aggregate are known.

(iii) It is capable of further mathematical treatment.

(iv) It provides a good basis to compare two or more frequency distributions.

(v) Mean does not necessitate the arrangement of data.

Disadvantages of Arithmetic Mean

(i) It may give considerable weight to extreme items. Mean of 2, 6, 301 is 103 and more of the values is adequately represented by the mean 103.

(ii) In some cases, arithmetic mean may give misleading impressions. For example, average number of patients admitted in a hospital is 10.7 per day, Here mean is a useful information but does not represent the actual item.

(iii) It can hardly be located by inspection.

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

GEOMETRIC MEAN (G. M.)

Definition. : The geometric mean (G) of the n positive values x1, x2, x3 ……xn is the nth root of the product of the values.

Advantages Geometric Mean

(i) It is not influenced by the extreme items to the same extent as mean.

(ii) It is rigidly defined and its value is a precise figure.

(iii) It is based on all observations and capable of further algebraic treatment.

(iv) It is useful in calculating index numbers.

Disadvantages of Geometric Mean :

(i) It is neither easy to calculate nor it is simple to understand.

(ii) If any value of a set of observations is zero, the geometric mean would be zero, and it cannot be determined.

(iii) If any value is negative, G. M. becomes imaginary.

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

HARMONIC MEAN (H. M.) :

Definition: The Harmonic Mean (H) for n observations, x1, x2,…….x n is the total number divided by the sum of the reciprocals * of the numbers.

ADVANTANGES OF HARMONIC MEAN :

- **(i) Like A.M. and G. M. it is also based on all observations.**
- **(ii) Capable of further algebraic treatment.**
- **(iii) It is extremely useful while averaging certain types of rates and rations.**

DISADVANTAGES OF HARMONIC MEAN :

- **(i) It is not readily understood nor can it be calculated with ease.**
- **(ii) It is usually a value which may not be a member of the given set of numbers.**
- **(iii) It cannot be calculated when there are both negative and positive values in a series or one of more values in zero.**

Relations among A.M., G.M. and H.M. :

The Arithmetic Mean is never less than the Geometric Mean, again Geometric Mean is never less than the Harmonic Mean. i.e. A.M. ^{3} G. M. ^{3} H. M.

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

MEDIAN :

Definition: If a set of observation are arranged in order of magnitude (ascending or descending), then the middle most or central value gives the median. Median divides the observations into two equal parts, in such a way that the number of observations smaller than median is equal to the number greater than it. It is not affected by extremely large or small observation. Median is, thus an average of position. In certain sense, it is the real measure of central tendency.

So, median is the middlemost value of all the observations when they are arranged in ascending order of magnitudes.

Calculation of Median from Discrete Grouped Distribution

If the class intervals of grouped frequency distribution are in discrete form, at first they are to be converted into class-boundaries and hence to find median by applying usual formula.

Calculation of median from open ends class intervals :

Since the first and last class intervals are not required in computing median, so in case of open end class intervals median is calculated by usual process.

For example, in the above example it the lower-limit of first class interval (i.e.0) and upper limit of last class (i.e. 5) are not given question, there would be no difficulty to compute median.

In case of open end class-intervals, median is preferred than A.M. as average.

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

Advantages of Median :

- **(i) The median, unlike the mean, is unaffected by the extreme values of the variable.**
- **(ii) It is easy to calculate and simple to understand, particularly in a series of individual observations a**
- **discrete series.**
- **(iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.**
- **(iv) It can be located by inspection, after arranging the data in order of magnitude.**
- **(v) Median can be calculated even if the items at the extreme are not known, but if we know the central**
- **items and the total number of items.**
- **(vi) It can be determined graphically.**

Disadvantages of Median :

- **(i) For calculation, it is necessary to arrange the data; other averages do not need any such arrangement.**
- **(ii) It is amenable to algebraic treatment in a limited sense, Median cannot be used to calculate the**
- **combined median of two or more groups, like mean.**
- **(iii) It cannot be computed precisely when it lies between two items.**
- **(iv) Process involved to calculate median in case of continuous series is difficult to follow.**
- **(v) Median is affected more by sampling fluctuations than the mean.**

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

MODE

Definition : Mode is the value of the variate which occurs most frequently. It represents the most frequent value of a series. In other words Mode is the value of the variable which has the highest frequency.

Advantages of mode :

- **(i) It can often be located by inspection.**
- **(ii) It is not affected by extreme values. It is often a really typical value.**
- **(iii) It is simple and precise. It is an actual item of the series except in a continuous series.**
- **(iv) Mode can be determined graphically unlike Mean.**

Disadvantages of mode :

- **(i) It is unsuitable for algebraic treatment.**
- **(ii) When the number of observations is small, the Mode may not exist, while the Mean and Median can be calculated.**
- **(iii) The value of Mode is not based on each and every item of series.**
- **(iv) It does not lead to the aggregate, if the Mode and the total number of items are given.**

Chapter – 2: MEASURES OF CENTRAL TENDENCY (Contd.)

Empirical Relationship among Mean, Median and Mode

A distribution in which the values of Mean, Median and Mode coincide, is known symmetrical and if the above values are not equal, then the distribution is said asymmetrical or skewed. In a moderately skewed distribution, there is a relation amongst Mean, Median and Mode which is as follows :

Mean – Mode = 3 (Mean – Median)

If any two values are known, we can find the other.

QUARTILES:

Quartiles are such values which divide the total number of observations into 4 equal parts. Obviously, there are 3 quartiles—

- **(i) First quartile (or Lower quartile): Q1**
- **(ii) Second quartile, (or Middle quartile) : Q2**
- **(iii) Third quartile (or Upper quartile): Q3**
- **The number of observations smaller than Q1, is the same as the number lying between Q1 and Q2, or between Q2 and Q3, or larger than Q3. For data of continuous type, one-quarter of the observations is smaller than Q1, two-quarters are smaller than Q2, and three quartersare smaller than Q3. This means that Q1, Q2, Q3 are values of the variable corresponding to 'less-than' cumulative frequencies N/4, 2N/4, 3N/4 respectively. Since, 2N/4 = N/2, it is evident that the second quartile Q2 is the same as median.**

Q1 < Q2 < Q3; Q2 = Median.

Chapter – 3: MEASURES OF DISPERSION

DISPERSION

A measure of dispersion is designed to state the extent to which individual observations (or items) vary from their average. Here we shall account only to the amount of variation (or its degree) and not the direction.

Usually, when the deviation of the observations form their average (mean, median or mode) are found out then the average of these deviations is taken to represent a dispersion of a series. This is why measure of dispersion are known as Average of second order.

Measures of dispersion are mainly of two types–

- **(A) Absolute measures are as follows :**
- **(i) Range, (ii) Mean deviation (or Average deviation), (iii) Standard deviation**
- **(B) Among the Relative measures we find the following types :**
- **(i) Coefficient of dispersion. (ii) Coefficient of variation.**

RANGE :

For a set observations, range is the difference between the extremes, i.e.

Range = Maximum value – Minimum value.

Chapter – 3: MEASURES OF DISPERSION (Contd.)

Advantages of Range : **Range is easy to understand and is simple to compute.**

Disadvantages of Range :

- **It is very much affected by the extreme values. It does not depend on all the observations, but only on the**
- **extreme values. Range cannot be computed in case of open-end distribution.**

Uses of Range :

- **It is popularly used in the field of quality control. In stock-market fluctuations range is used.**

Mean Deviation (or Average Deviation) : **Mean deviation of a series is the arithmetic average of the deviations of the various items from the median or mean of that series. Median is preferred since the sum of the deviations from the median is less than from the mean. So the values of mean deviation calculated from median is usually less than that calculated from mean.**

Mode is not considered, as its value is indeterminate.

Mean deviation is known as First Moment of dispersion.

Chapter – 3: MEASURES OF DISPERSION (Contd.)

Steps to find Mean Deviation

- **(1) Find mean or median**
- **(2) Take deviation ignoring ± signs**
- **(3) Get total of deviations**
- **(4) Divide the total by the number of items.**

Advantages of Mean Deviation :

- **(1) It is based on all the observations. Any change in any item would change the value of mean deviation.**
- **(2) It is readily understood. It is the average of the deviation from a measure of central tendency.**
- **(3) Mean Deviation is less affected by the extreme items than the standard deviation.**
- **(4) It is simple to understand and easy to compute.**

Disadvantages of Mean Deviation :

- **(1) Mean deviation ignores the algebraic signs of deviations and as such it is not capable of further algebraic treatment.**
- **(2) It is not an accurate measure, particularly when it is calculated from mode.**
- **(3) It is not popular as standard deviation.**

Uses of Mean Deviation :

- **Because of simplicity in computation, it has drawn the attention of economists and businessmen. It is useful reports meant for public.**

Chapter – 3: MEASURES OF DISPERSION (Contd.)

Standard Deviation :

In calculating mean deviation we ignored the algebraic signs, which is mathematically illogical. This drawback is removed in calculating standard deviation, usually denoted by ‘ s ’ (read as sigma).

Definition : Standard deviation is the square root of the arithmetic average of the squares of all the deviations from the mean. In short, it may be defined as root-mean-square deviation from the mean.

COEFFICIENT OF VARIATION :

It is the ratio of the Standard Deviation to the Mean expressed as percentage. This relative measure was first suggested by Professor Kari Pearson. According to him, coefficient is the percentage variation in the Mean, while Standard Deviation is the total variation in the Mean.

Chapter – 3: MEASURES OF DISPERSION (Contd.)

Advantages of Standard Deviation :

- **1. Standard deviation is based on all the observations and is rigidly defined.**
- **2. It is amenable to algebraic treatment and possesses many mathematical properties.**
- **3. It is less affected by fluctuations of sampling than most other measures of dispersion.**
- **4. For comparing variability of two or more series, coefficient of variation is considered as most appropriate and this is based on standard deviation and mean.**

Disadvantages of Standard Deviation :

- **1. It is not easy to understand and calculate.**
- **2. It gives more weight to the extremes and less to the items nearer to the mean, since the squares of the deviations of bigger sizes would be proportionately greater than that which are comparatively small.**
- **The deviations 2 and 6 are in the ratio of 1 : 3 but their squares 4 and 36 would be in the ratio of 1 : 9.**

Uses of Standard Deviation :

- **It is best measure of dispersion, and should be used wherever possible.**

Chapter – 4: CORRELATION AND REGRESSION

Importance of Correlation

Correlation helps in the following ways

- **1. It helps to predict event and the events in which there is time gap i.e. it helps in planning**
- **2. It helps in controlling events.**

Types of Correlation

- **Correlation can be classified under the following heads-**
- **1. Positive and negative correlation**
- **2. Simple multiple and partial correlation**
- **3. Linear and non-linear correlation**

Positive and Negative Correlation

Two variables are said to be positively correlated when both the variables move in the same direction. The correlation is said to be positive (directly related) when the increase in the value of one variable is accompanied by an increase in the value of the other variable and vice versa.

Two variables are said to be negatively correlated when both the variables move in the opposite direction. The correlation is said to be negative (inversely related) when the increase in the value of one variable is accompanied by a decrease in the value of the other variable and vice versa.

Chapter – 4: CORRELATION AND REGRESSION (Contd.)

Simple, Multiple and Partial Correlation

Correlation is said to be simple when only two variables are studied. In multiple correlation three or more variables are studied simultaneously. In partial correlation though more than two variables are recognized, but only two are considered to be influencing each other; and the effect of other influencing variables are kept constant.

Linear and Non-linear Correlation

If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. The correlation is said to be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other related variable.

Measurement of Correlation

The correlation can be measured by any of the following methods-

- **1. Scatter Diagram**
- **2. Karl Pearson's coefficient of correlation**
- **3. Rank correlation coefficient**

Chapter – 4: CORRELATION AND REGRESSION (Contd.)

Scatter Diagram Method

The scatter diagram represents graphically the relation between two variables Xand Y. For each pair of Xand Y, one dot is put and we get as many points on the graph as the number of observations. Degree of correlation between the variables can be estimated by examining the shape of the plotted dots.

Advantages

- **(1) It is very easy to draw a scatter diagram**
- **(2) It is easily understood and interpreted**
- **(3) Extreme items does not unduly affect the result as such points remain isolated in the diagram**

Disadvantages

- **(1) It does not give precise degree of correlation**
- **(2) It is not amenable to further mathematical treatment**

Chapter – 4: CORRELATION AND REGRESSION (Contd.)

Karl Pearson's Coefficient of Correlation

The measure of degree of relationship between two variables is called the correlation coefficient. It is denoted by symbol r.

This method is most widely used in practice. It is denoted by symbol V. The formula for computing coefficient of correlation can take various alternative forms depending upon the choice of the user.

Rank Correlation

Rank method for the computation of the coefficient of correlation is based on the rank or the order & not the magnitude of the variable. Accordingly it is more suitable when the variables can be arranged for e.g. in case of intelligence or beauty or any other qualitative phenomenon. The ranks may range from 1 to n.

DISTINCTION BETWEEN CORRELATION AND REGRESSION

By correlation we mean the degree of association or relationship between two or more variables. Correlation does not predict anything about the cause & effect relationship. Even a high degree of correlation does not imply necessarily that a cause & effect relationship exists between the two variables.

Whereas in case of regression analysis, there is a functional relationship between Y and X such that for each value of Y there is only one value of X. One of the variables is identified as a dependent variable the other(s) as independent valuable(s). The expression is derived for the purpose of predicting values of a dependent variable on the basis of independent valuable(s).

Chapter – 4: CORRELATION AND REGRESSION (Contd.)

REGRESSION LINES

A regression line is the line which shows the best mean values of one variable correspond-ing to mean values of the other. With two series X and Y, there are two arithmetic regression lines, one showing the best mean values of X corresponding to mean Y's and the other showing the best mean values of Y corresponding to mean X's. In the context of scatter diagram, the regression line is the straight line that best fits the scatter diagram. The most commonly used criteria is that it is the straight line that minimize the sum of the squared deviations between the predicted and observed values of the dependent variable. In the case of two variables X and Y, there will be two regression lines as the regression of X on Y and regression of Y on X.

REGRESSION EQUATIONS

There are different methods of deriving regression equations

- **(1) By taking actual values of X and Y**
- **(2) By taking deviations from actual mean**
- **(3) By taking deviations from assume mean**

Chapter – 4: CORRELATION AND REGRESSION (Contd.)

REGRESSION COEFFICIENTS

The regression coefficient gives the value by which one variable increases for a unit increase in other variable, bXY and bYX are two coefficient of regression.

FEATURES OF REGRESSION COEFFICIENTS

- **(i) Both of regression coefficients should have same sign i e., either positive or negative.**
- **(2) Coefficient of correlation could be found out if regression coefficients are known.**
- **(3) Correlation coefficient would have the same sign as that of regression coefficients. ie., either positive or negative.**
- **(4) Since -1£ r£ 1 this implies both the regression coefficient cannot be greater than one.**

Chapter – 5: INDEX NUMBERS

INDEX NUMBERS

Index number is a statistical device designed to measure changes or differences in magnitudes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession etc.

When the variation in the level of a single item is being studied, the index number is termed . as univariate index. But when the changes in average level of the number of items are being studied then collectively this index number is termed as composite index number. Most index numbers are composite in nature.

USES OF INDEX NUMBERS

1) **Index Numbers are the economic barometers.**
2) **Index number helps in formulation of policy decisions.**
3) **Index numbers reveal trends and tendencies.**
4) **Index numbers help to measures the Purchasing Power of money.**
5) **Consumer price indices are used for deflating.**

Chapter – 5: INDEX NUMBERS (Contd.)

PROBLEMSIN THE CONSTRUCTION OF INDEX NUMBERS

- A clear definition of the purpose for which the index is constructed should be made.
- Selection of number of items.
- Base period.
- Selection of weights.
- Adoption of suitable formula for construction of index number.

METHODS OF CONSTRUCTION OF INDEX NUMBERS

Index numbers may be constructed by any of the following methods—

(1) Unweighted Index :

- **(a) Simple Aggregative Index**
- **(b) Simple Average of Relatives**

(2) Weighted Indices:

- **(a) Weighted Aggregative. Index**
- **(b) Weighted Average of Relatives**

Chapter – 5: INDEX NUMBERS (Contd.)

UNWEIGHTED INDEX : SIMPLE AVERAGE OF PRICE RELATIVE METHOD

Under this method the price of each commodity in the current year is taken as a percentage of the price of corresponding item of the base year and the index is obtained by averaging these percentage figures. Arithmetic mean or geometric mean may be used to average these percentages.

Weighted Aggregate Method

In this method, appropriate weights are assigned to various commodities to reflect their relative importance in group. For the construction of price index number, quantity, weights are used i.e. amount of quantity consumed, purchased or marketed.

Laspeyres' Price Index

In this method the base year quantities are taken as weights.

Paasche's Method

In this method current year quantities are taken as weights.

Dorbish and Bowley's Method

This method is the simple arithmetic mean of the Laspeyres' and Paasche's indices. This index takes into account the influence of quantity weights of both base period and current period.

Chapter – 5: INDEX NUMBERS (Contd.)

Fisher 'Ideal' Method

This method is the geometric mean of Laspeyres' and Paasche's indices.

Advantages

Because of the following advantages this method is seldom referred as ideal method—

- **(1) The formula takes into account both base year and current year quantities as weights, and hence avoids bias associated with the Laspeyres' and Paasche's indices.**
- **(2) The formula is based on geometric mean which is considered to be the best average for constructing index numbers.**
- **(3) This method satisfies unit test, time reversal test and factor reversal test.**

Disadvantages

- **(1) This method is more time consuming than other methods.**
- **(2) It also does not satisfy circular test.**

Marshall-Edgeworth Method

In this method arithmetic mean of base year and current year quantities are taken as weights.

Chapter – 5: INDEX NUMBERS (Contd.)

Kelly's Method

In this method fixed weights are taken as weights. This method is sometimes referred to as aggregative index with fixed weights method. Fixed weights are quantities which may be for some particular period (not necessarily of base year or the current year) and this is kept constant all the time.

Weighted Index : Weighted Average of Relative Method

In this method price of each commodity in the current year is taken as a percentage of the price of corresponding item of the base year. These relatives are multiplied by the given weights and the result is obtained by averaging the resulting figures. Arithmetic mean or geometric mean is used to average these figures.

CONSUMER PRICE INDEX

The consumer price index measures the amount of money which consumer of a particular class have to pay to get a basket of goods & services at a particular point of time in comparison to what they paid for the same in the base period.

Different classes of people consume different types of commodities & even that same type of commodities are not consumed in the same proportion by different classes of people (for e.g. higher class, middle class, lower class). The general indices do not highlight the effects of change in prices of a various commodities consumed by different classes of people on their cost of living.

Chapter – 5: INDEX NUMBERS (Contd.)

Methods of Constructing Consumer Price Index

The consumer price index can be constructed by any of the following two methods :

- **(1) Aggregate Expenditure Method or Aggregative Method**
- **(2) Family Budget Method or the Method of Weighted Relatives**

Uses of Consumer Price Index Number

- **(1) It is used to formulate economic policy and also to measure real earning.**
- **(2) It is used to measure purchasing power of the consumer.**
- **It is used in deflating.**
- **It is used in wage negotiations & wage contracts. It also helps to calculate dearness allowance.**

TEST OF ADEQUACY OF THE INDEX NUMBER FORMULAE

- **(1) Unit Test**
- **(2) Time Reversal Test**
- **(3) Factor Reversal Test**
- **(4) Circular Test**

Chapter – 5: INDEX NUMBERS (Contd.)

CHAIN INDEX NUMBERS

In the fixed base method which is discussed so far the base remains the same & does not change whole throughout the series. But with the passage of time some items may have been included in the series & other ones might have been deleted, & hence it becomes difficult to compare the result of present conditions with those of the old remote period. Hence the fixed base method does not suit when the conditions change. In such a case the changing base period may be more suitable. Under this method the figures for each year are first expressed as a percentage of the preceding year (called link relatives) then they are chained together by successive multiplication to form a chain index.

Chapter – 6: TIME SERIES ANALYSIS

DEFINITION OF TIME SERIES

According to Spiegel, "A time series is a set of observations taken at specified times, usually at equal intervals."

According to Ya-Lun-Chou, "A time series may be defined as a collection of reading belonging to different time period of same economic variable or composite of variables."

COMPONENTS OF TIME SERIES

There are various forces that affect the values of a phenomenon in a time series; these may be broadly divided into the following four categories, commonly known as the components of a time series.

- **(1) Long term movement or Secular Trend**
- **(2) Seasonal variations**
- **(3) Cyclical variations**
- **(4) Random or irregular variations**

Secular Trend or Simple trend - The general tendency of a data to increase or decrease or stagnate over a long period of time is called secular trend or simple trend.

Chapter – 6: TIME SERIES ANALYSIS (Contd.)

Seasonal variations - Over a span of one year, seasonal variation takes place due to the rhythmic forces which operate in a regular and periodic manner. These forces have the same or almost similar pattern year after year.

Cyclical variations - These variations in a time series are due to ups & downs recurring after a period from time to time. Though they are more or less regular, they may not be uniformly periodic. These are oscillatory movements which are present in any business activity and is termed as business cycle.

Random or irregular variations - These fluctuations are a result of unforeseen and unpredictably forces which operate in absolutely random or erratic manner. They do not have any definite pattern and it cannot be predicted in advance. These variations are due to floods, wars, famines, earthquakes, strikes, lockouts, epidemics etc.

MODELS OF TIME SERIES ANALYSIS

The following are the two models which are generally used for decomposition of time series into its four components. The objective is to estimate and separate the four types of variations and to bring out the relative impact of each on the overall behaviour of the time series.

- **(1) Additive model**
- **(2) Multiplicative model**

Chapter – 6: TIME SERIES ANALYSIS (Contd.)

Additive Model - In additive model it is assumed that the four components are independent of one another i.e. the pattern of occurrence and magnitude of movements in any particular component does not affect and are not affected by the other component. Under this assumption the four components are arithmetically additive ie. magnitude of time series is the sum of the separate influences of its four components i.e.

Yt= T + C + S + I

Where: Yt = Time series, T = Trend variation, C = Cyclical variation, S = Seasonal variation, C = Random or irregular variation.

Multiplicative Model - In this model it is assumed that the forces that give rise to four types of variations are interdependent, so that overall pattern of variations in the time series is a combined result of the interaction of all the forces operating on the time series. Accordingly, time series are the product of its four components i.e.

Yt = T x C x S x I

As regards to the choice between the two models, it is generally the multiplication model which is used more frequently. As the forces responsible for one type of variation are also responsible for other type of variations, hence it is multiplication model which is more suited in most business & economic time series data for the purpose of decomposition.

Chapter – 6: TIME SERIES ANALYSIS (Contd.)

MEASUREMENT OF SECULAR TREND

The following are the methods most commonly used for studying & measuring the trend component in a time series—

- **(1) Graphic or a Freehand Curve method**
- **(2) Method of Semi Averages**
- **(3) Method of Moving Averages**
- **(4) Method of Least Squares**

Graphic or Freehand Curve Method

The data of a given time series is plotted on a graph and all the points are joined together with a straight line. This curve would be irregular as it includes short run oscillation. These irregularities are smoothened out by drawing a free hand curve or line along with the curve previously drawn. This curve would eliminate the short run oscillations & would show the long period general tendency of the data. While drawing this curve it should be kept in mind that the curve should be smooth and the number of points above the trend curve should be more or less equal to the number of points below it.

Merits

- **(1) It is very simple and easy to construct.**
- **(2) It does not require any mathematical calculations and hence even a layman can understand it.**

Disadvantages

- **(1) This is a subjective concept. Hence different persons may draw free hand lines at different positions**
- **and with different slopes.**
- **(2) If the length of period for which the curve is drawn is very small, it might give totally erroneous results.**

Chapter – 6: TIME SERIES ANALYSIS (Contd.)

Method of Semi Averages

Under this method the whole time series data is classified into two equal parts and the averages for each half are calculated. If the data is for even number of years, it is easily divided into two. If the data is for odd number of years, then the middle year of the time series is left and the two halves are constituted with the period on each side of the middle year. The arithmetic mean for a half is taken to be representative of the value corresponding to the mid point of the time interval of that half. Thus we get two points. These two points are plotted on a graph and then are joined by straight line which is our required trend line.

Moving Average Method

A moving average is an average (Arithmetic mean) of fixed number of items (known as periods) which moves through a series by dropping the first item of the previously averaged group and adding the next item in each successive average. The value so computed is considered the trend value for the unit of time falling at the centre of the period used in the calculation of the average.

In case the period is odd- If the period of moving average is odd for instance for computing 3 yearly moving average, the value of 1st, 2nd & 3rd years are added up and arithmetic mean is found out and the answer is placed against the 2nd year; then value of 2nd, 3rd & 4th years are added up & arithmetic mean is derived and this average is placed against 3rd year (ie. the middle of 2nd, 3rd & 4th) and so on.

In case of even number of years - If the period of moving average is even for instance for computing 4 yearly moving average, the value of 1st, 2nd, 3rd & 4th years are added up & arithmetic mean is found out. and answer is placed against the middle of 2nd & 3rd year. The second average is placed against middle of 3rd & 4th year. As this would not coincide with a period of a given time series an attempt is made to synchronise them with the original data by taking a two period average of the moving averages and placing them in between the corresponding time periods. This technique is called centering & the corresponding moving averages are called moving average centred.

Chapter – 6: TIME SERIES ANALYSIS (Contd.)

METHOD OF LEAST SQUARES

The method of least squares as studied in regression analysis can be used to find the trend line of best fit to a time series data. The regression trend line (Y) is defined by the following equation—

$Y = a + b X$

where Y = predicted value of the dependent variable

a = Y axis intercept or the height of the line above origin (i.e. when X = 0, Y = a)

b = slope of the regression line (it gives the rate of change in Y for a given change in X) (when b is positive the slope is upwards, when b is negative, the slope is downwards)

X = independent variable (which is time in this case)

To estimate the constants a and b, the following two equations have to be solved simultaneously—

$SY = na + b SX$

$SXY = aSX + bSX2$

Chapter – 7: PROBABILITY

CONCEPT OF PROBABILITY

The concept of probability is difficult to define in precise terms. In ordinary language, the word probable means likely or chance. The probability theory is an important branch of mathematics. Generally the word, probability, is used to denote the happening of a certain event, and the likelihood of the occurrence of that event, based on past experiences. By looking at the clear sky, one will say that there will not be any rain today. On the other hand, by looking at the cloudy sky or overcast sky, one will say that there will be rain today. In the earlier sentence, we aim that there will not be rain and in the latter we expect rain. On the other hand a mathematician says that the probability of rain is 0 in the first case and that the probability of rain is 1 in the second case. In between 0 and 1, there are fractions denoting the chance of the event occurring.

Random Experiment or Trial :

If an experiment or trial can be repeated under the same conditions, any number of times and it is possible to count the total number of outcomes, but individual result i.e. individual outcome is not predictable.

Suppose we toss a coin. It is not possible to predict exactly the outcomes. The outcome may be either head up or tail up. Thus an action or an operation which can produce any result or outcome is called a random experiment or a trial.

Chapter – 7: PROBABILITY (Contd.)

Event: Any possible outcome of a random experiment is called an event. Performing an experiment is called trial and outcomes are termed as events.

An event whose occurrence is inevitable when a certain random experiment is performed, is called a sure event or certain event. At the same time, an event which can never occur when a certain random experiment is performed is called an impossible event. The events may be simple or composite. An event is called simple if it corresponds to a single possible outcome. For example, in rolling a die, the chance of getting 2 is a simple event. Further in tossing a die, chance of getting event numbers (1, 3, 5) are compound event.

Sample space

The set or aggregate of all possible outcomes is known as sample space. For example, when we roll a die, the possible outcomes are 1, 2, 3, 4, 5, and 6 ; one and only one face come upwards. Thus, all the outcomes—

1, 2, 3, 4, 5 and 6 are sample space. And each possible outcome or element in a sample space called sample point.

Mutually exclusive events or cases :

Two events are said to be mutually exclusive if the occurrence of one of them excludes the possibility of the occurrence of the other in a single observation. The occurrence of one event prevents the occurrence of the other event. As such, mutually exclusive events are those events, the occurrence of which prevents the possibility of the other to occur. All simple events are mutually exclusive. Thus, if a coin is tossed, either the head can be up or tail can be up; but both cannot be up at the same time.

Chapter – 7: PROBABILITY (Contd.)

Equally likely events :

The outcomes are said to be equally likely when one does not occur more often than the others. That is, two or more events are said to be equally likely if the chance of their happening is equal. Thus, in a throw of a die the coming up of 1, 2, 3, 4, 5 and 6 is equally likely. For example, head and tail are equally likely events in tossing an unbiased coin.

Exhaustive events

The total number of possible outcomes of a random experiment is called exhaustive events. The group of events is exhaustive, as there is no other possible outcome. Thus tossing a coin, the possible outcome are head or tail ; exhaustive events are two. Similarly throwing a die, the outcomes are 1, 2, 3, 4, 5 and 6. In case of two coins, the possible number of outcomes are 4 i.e. (22), i.e., HH, HT TH and TT. In case of 3 coins, the possible outcomes are 23=8 and so on. Thus, in a throw of n" coin, the exhaustive number of case is 2n.

Independent Events

A set of events is said to be independent, if the occurrence of any one of them does not, in any way, affect the Occurrence of any other in the set. For instance, when we toss a coin twice, the result of the second toss will in no way be affected by the result of the first toss.

Chapter – 7: PROBABILITY (Contd.)

Dependent Events

Two events are said to be dependent, if the occurrence or non-occurrence of one event in any trial affects the probability of the other subsequent trials. If the occurrence of one event affects the happening of the other events, then they are said to be dependent events. For example, the probability of drawing a king from a pack of 52 cards is 4/52, ; the card is not put back ; then the probability of drawing a king again is 3/51. Thus the outcome of the first event affects the outcome of the second event and they are dependent. But if the card is put back, then the probability of drawing a king is 4/52 and is an independent event.

Simple and Compound Events

When a single event take place, the probability of its happening or not happening is known as simple event. When two or more events take place simultaneously, their occurrence is known as compound event (compound probability) ; for instance, throwing a die.

Complementary Events :

The complement of an events, means non-occurrence of A and is denoted by A . A contains those points of the sample space which do not belong to A. For instance let there be two events A and B. A is called the complementary event of B and vice verse, if A and B are mutually exclusive and exhaustive.

Chapter – 7: PROBABILITY (Contd.)

Favourable Cases

The number of outcomes which result in the happening of a desired event are called favourable cases to the event. For example, in drawing a card from a pack of cards, the cases favourable to "getting a diamond" are 13 and to "getting an ace of spade" is only one. Take another example, in a single throw of a dice the number of favourable cases of getting an odd number are three -1,3 and 5.

Classical Approach (Priori Probability)

The classical approach is the oldest method of measuring probabilities and has its origin in gambling games. According to this approach, the probability is the ratio of favourable events to the total number of equally likely events.

Relative Frequency Theory of probability

Classical approach is useful for solving problems involving game of chances—throwing dice, coins, etc. but if applied to other types of problems it does not provide answers. For instance, if a man jumps from a height of 300 feet, the probability of his survival will, not be 50%, since survival and death are not equally alike.

Similarly, the prices of shares of a Joint Stock Company have three alternatives i.e. the prices may remain constant or prices may go up or prices may go down. Thus, the classical approach fails to answer questions of these type.

Chapter – 7: PROBABILITY (Contd.)

Addition Theorem of Probability

The simplest and most important rule used in the calculation is the addition rules, it states, "If two events are mutually exclusive, then the probability of the occurrence of either A or B is the sum of the probabilities of A and B. Thus, P(A or B)=P(A)+P(B)

When events are not mutually exclusive:

The addition theorem studied above is not applicable when the events are not mutually exclusive. In such cases where the events are not mutually exclusive, the probability is :

P(A or B) = P(A) + P(B) - P(A and B)

Multiplication Theorem of Probability

When it is desired to estimate the chances of the happening of successive events, the separate probabilities of these successive events are multiplied. If two events A and B are independent, then the probability that both will occur is equal to the product of the respective probabilities. We find the probability of the happening of two or more events in succession. Symbolically :

P(A and B) = P(A) x P(B)

When events are dependent :

If the events are dependent, the probability is conditional. Two events A and B are dependent ; B occurs only when A is known to have occurred.

Chapter – 7: PROBABILITY (Contd.)

BAYES' THEOREM

This theorem is associated with the name of Reverend Thomas Bayes. It is also known as the inverse probability. Probabilities can be revised when new information pertaining to a random experiment is obtained. One of the important applications of the conditional probability is in the computation of unknown probabilities, on.the basis of the information supplied by the experiment or past records. That is, the applications of the results of probability theory involves estimating unknown probabilities and making decisions on the basis of new sample information. This concept is referred to as Bayes' Theorem.

Chapter – 8: THEORETICAL DISTRIBUTION

THEORETICAL DISTRIBUTION

Broadly speaking, the frequency distributions are of two types : Observed Frequency Distribution and Theoretical Frequency Distribution. The distributions, which are based on actual data or experimentation are called the observed frequency distribution. On the other hand, the distributions based on expectations on the basis of past experience is known as Theoretical Frequency Distribution or Expected Frequency Distribution or Probability Distributions. In short, the observed frequency distribution is based on actual sample studies whereas the theoretical distribution is based on expectations on the basis of previous experience or theoretical considerations.

The following are important distributions:

- **1. Binomial Distribution Discrete Probability Distribution**
- **2. Poisson Distribution Discrete Probability Distribution**
- **3. Normal Distribution Continuous Probability Distribution**

Chapter – 8: THEORETICAL DISTRIBUTION (Contd.)

BINOMIAL DISTRIBUTION

Bernoulli Distribution. He discovered this theory and published it in the year 1700 dealing with dichotomous classification of events one possessing and the other not possessing. The probability of occurrence of an event is *p and its non-occurrence is q. The distribution can be used under the following conditions :*

- **1. The number of trials is finite and fixed.**
- **2. In every triail there are only two possible outcomes success or failure.**
- **3*. The trials are independent. The outcome of one trial does not affect the other trial.***
- **4. *p, the probability of success from trial to trial is fixed and q the probability of failure is equal to 1-p. This* is the same in all the trial.**

PROPERTIES OF BINOMIAL DISTRIBUTION

- **1. Binomial distribution has two parameters – *n and p (or q)***
- **2. Mean = *np***
- **3. Variance = *npq***
- **4. Standard Deviation = Root over of *npq***
- **5. Binomial distribution is symmetrical if p = q = 0.5**

Chapter – 8: THEORETICAL DISTRIBUTION (Contd.)

POISSON DISTRIBUTION

Poisson distribution was derived in 1837 by a French mathematician Simeon *D Poisson (1731-1840). In binomial* distribution, the values of *p and q and n are given. There is a certainty of the total number of events; in* other words, we know the number of times an event does occur and also the times an event does not occur, in binomial distribution. But there are cases where p is very small and *n is very large, then calculation* involved will be long. Such cases will arise in connection with rare events, for example.

- **1. Persons killed in road accidents.**
- **2. The number of defective articles produced by a quality machine,**
- **3. The number of mistakes committed by a good typist, per page.**
- **4. The number of persons dying due to rare disease or snake bite etc.**
- **5. The number of accidental deaths by falling from trees or roofs etc.**

Chapter – 8: THEORETICAL DISTRIBUTION (Contd.)

NORMAL DISTRIBUTION

The Binomial distribution and Poisson distribution discussed above are discrete probability distributions. The normal distribution is highly useful in the field of statistics and is an important continuous probability distribution. The graph of this distribution is called normal curve, a bell-shaped curve extending in both the directions, arriving nearer and nearer to the horizontal axis but never touches it.

The normal distribution was first discovered by the English mathematician De-Moivre (1667-1754) in 1673 to solve the problems in game of chances. Later, it was applied in natural and social science by the French mathematician La Place (1749-1827). Normal distribution is also known as Gaussian distribution (Gaussian Law of Error).

In binomial distribution, which is a discrete distribution as the expression of *N(p + q)n gives the expected* frequencies of 0, 1, 2, *3..N successes. As n gets very large, the problem of computing the frequencies* becomes difficult and tedious. This difficult situation is handled by the application of normal curve. This curve not only eliminates tedious computations but also gives close approximation to binomial distribution.