Lesson : 1

INTRODUCTION TO STATISTICS

Author : **Vetter:**

Dr. Pradeep Gupta **Dr. B. S. Bodla**

Plan of the Lesson :

1. What is statistics
2. Definition
3. Definition of Statistics
4. Functions of Statistics
5. Importance of Statistics
6. Statistics and Computers
7. Limitations of Statistics
8. Self-Test Questions
9. Suggested Readings

1. WHAT IS STATISTICS

Life in the modern world is inexuricably bound with the notions of number, counting and measurement. One day try to think of a community that cannot count or take measurements and yet is concerned with such acts as selling and buying, carrying on bank transactions, operating locomotives, cars, ships, aircraft and taking part in government. The overriding importance of numerical data in modern life will then be all too apparent. Statistics is being used both is a singular noun and a plural noun.

Statistics, as a plural noun, is used to mean numerical data which arise from a host of uncontrolled, and mostly unknown, causes acting together. It is in this sense that the term statistics is used when our daily newspapers give vital statistics, crime statistics or soccer statistics of Calcutta, or when the Food Minister in the Lok Sabha quotes statistics of sugar exports or those of food grain production.

Used as singular, statistics is a name for the body of scientific methods which are meant for the collection, classification, tabulation, analysis and interpretation of numerical data. But modern literature on the subject does away with any such distinction.

2. ORIGIN AND GROWTH OF STATISTICS

Statistics is not a new discipline but as old as the human society itself. In the old days statistics was regarded as the 'Science of Statecraft' and was the by-product of the administrative activity of the State. It has been the traditional function of the governments to keep records of population, births, deaths, taxes crop yields and many other types of activities. Counting and measuring these events may generate many kind of numerical data.

The word 'statistics' comes from the Italian word 'statista' (meaning "Statesman") or the German word 'Statistik' each of which means a Political State. It eas first used by Professor Gottfried in 1749 to refer to the subject matter as a whole. The science of statistics is said to have originated from two main sources.

(i) *Government Record* : This is the earliest foundation because all cultures with a recorded history had recorded statistics, and the recording, as far as is known, was done by agents of the government for governmental purpose. Since

statistical data were collected for governmental purpose, statistics was then described as the 'science of kings' or 'the science of statecraft'.

(ii) Mathematics : Statistics is said to be a branch of applied mathematics. The present body of statistical methods, particularly those concerned with drawing inferences about population from a sample is based on the mathematical theory of probability.

The following are the two main factors which are responsible for the *development* or statistics in modern time :

(a) *Increased demand for statistics :* In the present century considerable development has taken place in the field of business and commerce, governmental activities and science. Statistics help in formulating suitable policies, and as such its need is increasingly felt in all these spheres.

(b) *Reduced cost of statistics :* The time and cost of collecting data are very important limiting factors in the use of statistics. However, with the development of electronic machines, such as calculators, computers etc. the cost of analysing data has considerably gone down. This has led to the increasing use of statistics in solving various problems. Moreover, with the development of statistical theroy the cost of collecting and processing data has gone. For example, considerable advance has been made in the sampling techniques which enable us to know the characteristics of the population by studying only a part of it.

3. DEFINITION OF STATISTICS

The purpose of definition is to lay down precisely the meaning, the scope and the limitations of a subject. There are many definitions of the term

'statistics'. A few definitions are analytically examined below:

Webster defined statistics as "the classified facts representing the conditns of the people in a state ....especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement".

Yule and Kendall defined statistics as "By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes".

Croxton and Cowden have given a very simple and concise definition of statistics. In their view "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data".

According to Berenson and Levin, "The science of statistics can be viewed as the application of the scientific method in the analysis of numerical data for the purpose of making rational decisions".

Boddington defines statistics as "the science of estimates and probabilities".

According to Lincon L. Chao, "Modern statistics refers to a body of methods and principles that have been developed to handle the collection, description, summarisation and analysis of numerical data. Its primary objective is to assist the researcher in making decisions or generalizations about the nature and characteristics of all the potential observations under consideration of which the collected data form only a small part".

All the above definitions are less comprehensive than the one given by Prof. Horace who defined statistics as follows :

"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner

for a predetermined purpose and placed in relation to each other".

(i) *Statistics are aggregate of facts :* Single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared. To illustrate, if it is stated that the income or Mr. A is Rs. 1,00,000 per annum, this would not constitute statistics although it is numerical state of fact. Similarly, a single figure relating to production, sale, birth, employment, purchases, accident etc. cannot be regarded statistics although aggregates of such figures would be statistics because of their comparability and relationship as part of common phenomenon.

(ii) *Statistics are affected to a marked extent by multiplicity of causes :* Facts and figures are affected to a considerable extent by a number of forces operating together. For example, statistics of production of rice are affected by the rainfall, quality of soil, seeds, manure, method of cultivation etc.

(iii) *Statistics are numerically expressed :* All statistics are numerical statements of facts i.e. expressed in numbers. Qualitative statements such as 'the population of India is rapidly increasing', or 'the production of wheat is not sufficient' do not constitute statistics. The rason is that such statements are vague and one cannot make at anything from them. On the other hand, the statement 'The estimated population of India at the end of VIIth plan is 803 million' is a statistical statement.

(iv) *Statistics are enumerated or estimated according to a reasonable standard of accuracy :* Facts and figures about any phenomenon can be derived in two ways, viz by actual counting and measurement or by estimate. Estimates cannot be as precise and accurate as actual counts

or measurements. The degree of accuracy desired largely depends or measurements. The degree of accuracy desired largely depends upon the nature and object of the enquiry. For example, in measuring heights of persons even 1/10th of a cm is material whereas in measuring distance between two places, say Madras and Calcutta, even fraction of a kilometre can be ignored. However, it is important that reasonable standards of accuracy should be attained, otherwise numbers may be altogether misleading.

(v) *Statistics are collected in a systematic manner :* Before collecting statistics a suitable plan of data collectiong should be prepared and the work carried out in a systematic manner. Data collected in a haphazard manner would very likely lead to fallacious decisions.

(vi) *Statistics are collected for a pre-determined pupose :* The purpose of collecting data must be decided in advance. The pupose should be specific and will defined. A general statement of purpose is not enough. For example, if the objective is to collect data on prices, it would not serve any useful purpose unless one knows whether he wants to collect data on wholesale or retail prices and what are the relevant commodities in view.

(vii) *Statistics should be placed in relation to each other :* If numerical facts are to be called statistics, they should be comparable. Statistics data are often compared period-wise or region-wise. For example, the per capita income of India at a particular point of time may be compared with that of earlier years or with the per capita income of other countires, say U.S.A., UK, China etc. Valid comparisons can be made only if the

data are homogeneous i.e. relate to the same phenomenon or subject and only likes are compared with likes. It would be meaningless to compare the height of elephants with the height of human beings.

In the absence of the above characteristics, numerical data cannot be called statistics.

4. FUNCTIONS OF STATISTICS

(i) *Statistics bring definiteness and precision in conclusions by expressing them numerically.* It is the quality of definiteness which is responsible for the growing universal applications of statistical methods. The conclusions stated numerically are definite and hence more convincing than conclusions stated qualitatively. This fact can be readily understood by a simple example. In an advertisement, statements expressed numerically have greater attention and more appealing than those expressed in a qualitative manner. The caption 'we have sold more. T.Vs this year', is certainly less attractive than 'Record Sale of 15,000 T.V. in 1998 as compared to 10,000 in 1997'. The latter statement emphasises in a much better manner the growing popularity of the advertise's T.Vs.

(ii) *Statistics make data comprehensible to the human mind by simplifying and summarising it.* Statistics simplifies unwieldy and complex mass of data and presents them in such a manner that they at once become intelligible. The complex data may be reduced to totals, averages, percentage etc. and presented either graphically or diagrammtically. These deives help to understand quickly the significant characterstics of the numerical data, and consequently save from a lot of mental strain. Single figures in the form of averages and percentages can be grasped

more easily than a mass of statistical data comprising thousands of facts. Similarly, diagrams and graphs, because of their greater appeal to the eye and imagination tender valuable assistance in the proper understanding of numerical data. Time and energy of business executives are thus economised, if the statistician supplies them with the results of production, sale and finances in a condensed form.

(iii) *Statistics facilitate comparisons in the data.* Certain facts, by themselves, may be meaningless unless they are capable of being compared with similar facts at other places or at other periods in times. For example, we estimate the national income of India not essentially for the value of that fact itself, but mainly in order that we may compare the income of today with that of the past and thus draw conclusions as to whether the standard of living of the people is on the increase, decrease or is stationary. It is with the help of statistics that the cost accountant is able to compare the actual accomplishment (in terms of cost). Some of the modes of comparison provided by statistics are : Totals, ratios, averages or measure of central tendencies, graphs & diagrams and coefficients. Statistics thus 'serves as a scale in which facts in various combinations are weighed and valued'.

(iv) *Statistics studies and esstablishes among the variables.* Certain statistical measures such as coefficient of correlation, regression etc. establish relationship between different types of data. For example, it is possible to observe the relationship between income and expenditure, export and forex reserves etc.

(v) *Statistics helps in formulating and testing hypothesis.* Statistical

methods are extremely useful in formulating and testing hypothesis and to develop new theroies. For examples the hypothesis that a new drug is effective in checking malaria, will require the use of statistical technique of association of attributes.

(vi) *Statistics helps in prediction.* Almost all our activities are based on estimates about future and the judicious forecasting of future trends is a prerequisite for efficient implementation of policies. The statistical techniques for extrapolation, time series etc. are highly useful for forecasting future events.

(vii) *Statistics helps in the formulation of suitable policies.* Statistics help in formulating policies in social, economic and business fieods. Various government policies in the field of planning taxation, foreign trade, social security etc. are formulated on the basis of analysis of statistical data and the inferences drawn from them. For example, vital statistics comprising birth and morality rates help in assessing future growth in population. This information is necessary for designing any scheme of family planning. Similarly, the rate of dearness allowance to be given to the employeess is calculated with the help of index numbers.

(viii) *Statistics draws inferences for taking decisions.* Statistical tests are devised to help in drawing valid inference in regard to the nature and characteristics of the universe on the basis of the study of the sample. It can also be the other way when the nature of the sample is judged on the basis of the parameters based on the study of the universe. The validity of such inferences depends on the type of statistical methods employed for the purpose.

(ix) *Statistics endeavours to interpret conditions.* Statistics render useful service by enabling the interpretation of condition, by developing possible causes for the results described. For example, if the production manager discovers that a certain machine is turning out some articles which are not of standard specifications, he will be able to find statistically if this conditon is due to some defects in the machine or wheather such a condition is normal.

(x) *Statistics measures uncertainty.* Statistical methods help not only in ascertaining the chance of occurrence of an event but also in finding out the total effect of an uncertain event if the consequences of various occurrences are known. Both objective and subjective probability estimates are employed depending upon the nature of the enquiry.

(xi) *Statistics enlarges individual experience.* A proper function of statistics, indeed is to enlarge individual experience. Many fields of knowledge would have remained closed to mankind, without the efficient and useful techniques of statistical analysis.

5. IMPORTANCE OF STATISTICS

Statistical methods have become useful tools in the world of affairs. Economy and a high degree of flexibility are the important qualities of statistical methods that render them specially useful to businessmen and scientists.

(i) *Statistics and Business :* Statistical information is needed from the time the business is launched till the time of its exit. At the time of the floatation of the concern facts are required for the purpose of drawing up the financial plan of the proposed unit. All the factors that are likely

to affect judgement on these matters are quantitatively weighed and statistically analysed before taking the decisions.

Statistical methods of analysis are helpful in the marketing function of an enterprise though enormous help in market research, advertisement campaigns and in comparing the sales performances. Statistics also directs attention towards the effective use of advertising funds.

Correlation and regression analysis help in the estimation of relationships between dependent and one or more independent varibles e.g. relationships are established between market demand and per capita income, inputs and outputs etc.

They theory and techniques of sampling can be used in connection with various business surveys with a considerable saving in time and money. Likewise these techniques are now being extensively used in checking of accounts.

Statistical quality control is now being used in industry for establishing quality standards for products, for maintaining the equisite quality, and for assuring that the individual lots sold are of a given standard of acceptance.

The use for statistical information in the smooth functioning of an undertaking increases along with its size. The bigger the concern the greater is the need for statistics.

Statistics is thus a useful tool in the hands of the management. But it must be remembered that no volume of statistics can replace the knowledge and experiences of the executives. Statistics supplements their knowledge with more precise facts than were hitherto available.

(ii) *Statistics & Economics :* Statistical data and methods of statistical analysis render valuable assistance in the proper understanding of the

economic problems and the formulation of economic policy. Economic problems almost always involve facts that are capable of being expressed numerically, e.g. volume of trade, output of industires - manufacturing, mining and agricultrue - wages, prices, bank deposits, clearing house returns etc. These numerical magnitudes are the outcome of a multiplicity of causes and are consequently subject to variations from time to time, or between place or among particular cases. Accordingly, the study of economic problem is specially suited to statistical treatment.

The development of economic theroy has also been facilitated by the use of statistics. Statistics is now being used increasingly not only to develop new economic cancepts but also to test the old ones. The increasing importance of statistics in the study of economic problem has resulted in a new branch of study called Econometrics.

(iii) Statistics and Biology : Statistics is being used more and more in biological sciences as an aid to the intelligent planning of experiments, and as a means of assuring the significance of the results of such experiments. Experiments about the growth of animals under different diets and environments, or the crop yields with different seeds, fertilizers and types of soil are frequently designed and analysed according to statistical principles.

(iv) Statistics and physical sciences : Statistics is not much in use in the fields of Astronomy, Geology and Physics. This is due mainly to their relatively high precision of measurements. Statistics has not made any progress in physical sciences beyond the calculation of standard error, and fittings of curves.

6. STATISTICS AND COMPUTERS

The development of statistics has been closely related to the evolution of electronic computing machinery. Statistics is a form of data processing, a way of converting data into information useful for decision making. A huge mass of raw data, of related and unrelated nature, derived from intrnal and external sources of different period of time can be organised and processed into information by computers with accuracy and high speed. The computers can make complex computations, analysis, comparisons and summarisations. Though humans can do the processing, the computer's ability to process huge data is phenomenal, considering its speed, reliability and faithfulness in perfectly following the set of instructions.

The input data in the computer can be processed into a number of different outputs and for a variety of purposes. The system is so organised that managers at different levels and in different activity units are in a position to obtain information in whatever form they want, provided that relevant 'programmes' or instructions have been designed for the purpose. However, the output from a computer is only as good as the data input. 'Garbage In Garbage Out' is an adage familiar to computer users. This warning applies equally to statistical analysis. Statistical decisions based on data are no better than the data used.

As statisticians devise new ways of describing and using data for decisions, computer scientists respond with newer and more efficient ways of performing these operations. Conversely, with the evolution of more powerful computing techniques, people in statistics are encouraged to explore new and more sophisticated methods of statistical analysis.

Statistical Analysis Pacakages

Statistical Analysis Pacakages are preprogrammed with all the specialized formulas and built-in procedures a user may need to carry out a range of statistical studies. Statistical programs can:

* Accept data from other sources.
* Add or remove data items, colums or rows.
* Sort, merge and manipulate facts in numerous ways.
* Perform analysis on single and multiple sets of data.
* Convert numeric data into charts and graphs that people can use to grasp relationships, spot patterns and make more informed decisions.
* Print summary values and analysis results.

For a period of at least twenty years, groups of standardized statistical programs assembled as a collection or "package" have been available from various software developers. Recently, there has been a widespread development of statistical pacakages for use on a microcomputer. Certain packages that were previously available only for mainframe and minicomputers, (such as SAS, SPSS and Minitab) are now available in microcomputer versions, and many new packages (such as STATGRAPHICS, SYSTAT, MYSTAT) have been specifically developed for microcomputer use. The easy and relatively inexpensive access to this type of software has led to its ever-increasing use for business applications.

7. LIMITATIONS OF STATISTICS

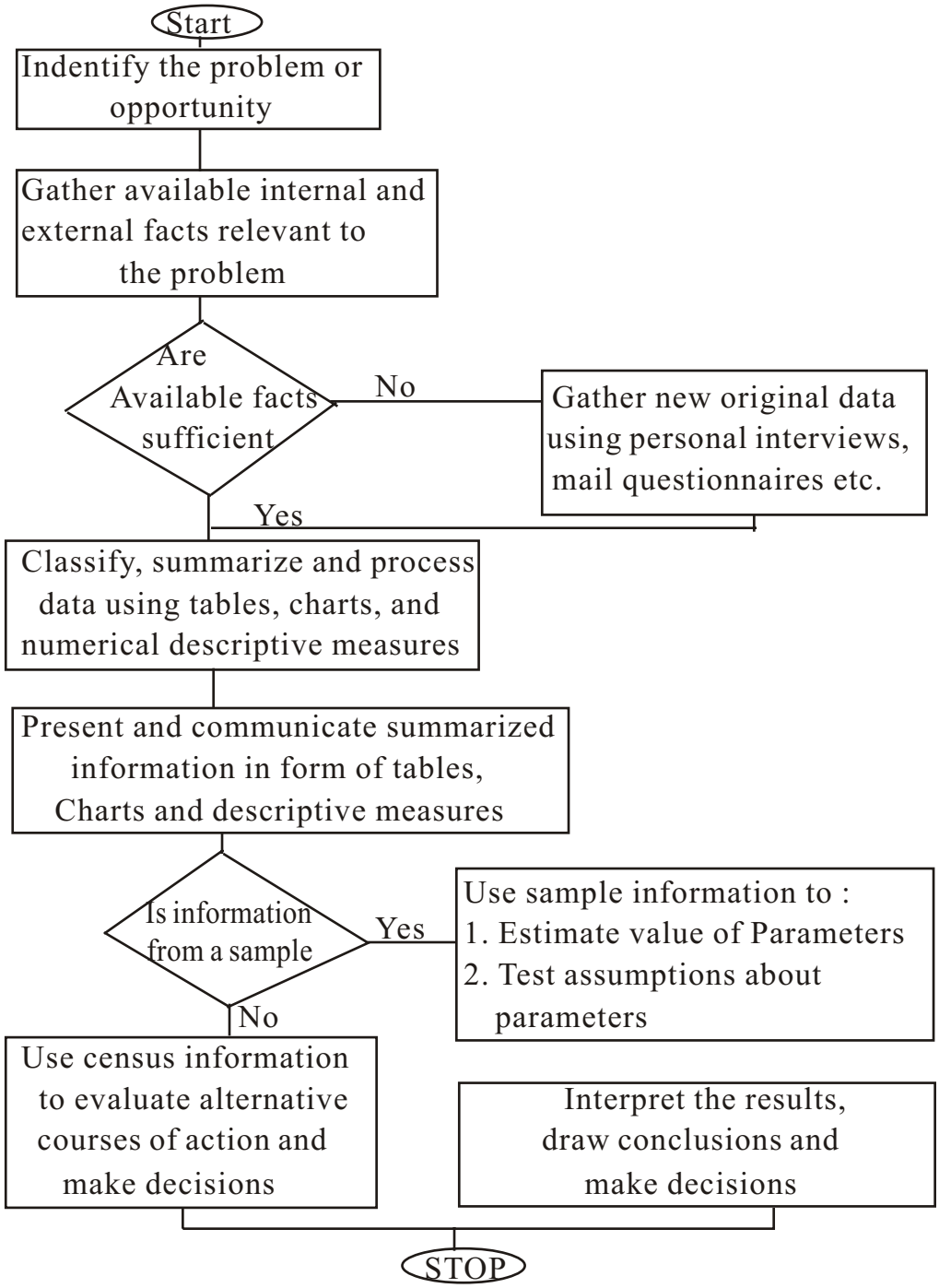
Though the science of statistics has been profitably applied to an increasingly large number of problems, it has its own limitations and is at times misused by interested people which restricts its scope and utility. According to Newsholme, "It (Statistics) must be regarded as an instrument of research of great value, but having severe limitations which are not possible to overcome

and as such they need our careful attention."

The following are some of the imortant limitations of statistics.

(i) Statistics does not study qualitative phenomenon : Statistics

Statistical Problem Solving Methodology



delas with only those subject of inquirty which are capable of being quantitatively measured and numerically expressed. This is an essential conditon for the application of statistical methods. Now all subjects cannot be expressed in numbers. Health, poverty, intelligence (to name only a few) are instances of the objects that defy the measuring rod, and hence are not suitable for statistical analysis. The efforts are being made to accord statistical treatment to subjects of this nature also. Health of the people is judged by a study of the death rate, longevity of life and prevalence of any disease or diseases. Similarly intellignece of the sutdents may be compared on the basis of the marks obtained by them in a class test. But these are only indirect methods of approaching the problem and subsidiary to quite a number of other considerations which cannot be statistically dealth with.

(ii) Statistics does not study individuals : Statistics deals only with agregates of facts and no importance is attached to individual items. Individual items, taken separately, do not constitute statistical data and are meaning less for any statistical inquiry. For example, the individual figures of agriculture production, industrial output or national income of any country for a particular year are meaningless, unless these figures enable comparison with similar figures for other countires and in the same country these are given for a number of years.

(iii) Statistical data is only approximately and not mathematically correct : Greater and greater exphasis is being laid on sampling technique of collecting data. This means that by objseving only a limited number of items we make an estimate of the characteristics of the entire population. This system works well so long as the mathematical accuracy is not essential. But when exactness

is essential statistics will fail to do the job.

(iv) Statistics is only one of the methods of studying a problem : Statistical tools do not provide the best solution under all circumstances. Very often, it is necessary to consider a problem in the light of a country's culture, religion and philosophy, Statistics cannot be of much help in studying such problems. Hence statistical conclusions must be supplemented by other evidences.

(v) Statistics can be misused : The greatest limitation of statistics is that it is liable to be misused. The misuse of statistics may arise because of several reasons. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. Thus the argument that drinking beer is bad for longevity because 99% of the persons who take beer die before the age of 100 years is statistically defective, since we were not told what percentage of persons who do not drink beer die before reaching that age. Statistics are like clay and they can be moulded in any manner so as to establish right or wrong conclusions.

8. SELF-TEST QUESTIONS

Q1. Define statistics. Also discuss the applicaions of statistics in business decision making.

Q2. Discuss the functions and limitations of statistics.

Q3. "Statistical methods are most dangerous tools in the hands of the inexpert" Elucidate.

Q4. "Statistics are numerical statement of facts but all facts numerically stated are not statistics" Comment upon the statement and state briefly which numerical statements of facts are not statistics.

Q5. How the computers can be helpful in making statistical dicision ?

9. SUGGESTED READINGS

(i) Statistical Methods By S.P. Gupta.

(ii) Practical Statistics By Shiv Kumar.

(iii) Statistics for Management By Levin.

(iv) Introduction to Statistical Methods By C.B. Gupta.

(v) Business Statistics By Sancheti & Kapoor.

(vi) Basic Statistics By B.L. Agarwal.

(vii) Elements of Practical Statistics By S.K. Kapur.

❄ ❄ ❄

Lesson : 2

MEASURES OF CENTRAL TENDENCY

Author : **Vetter:**

Dr. Pra deep Gupta **Dr. B.S. Bodla**

Plan of the Lesson :

- Characteristic of a good average.
- Arithmetic Mean.
- Median
- Mode
- Mode
- Geometric Mean
- Harmonic Mean

MEASURES OF CENTRAL TENDENCY

Central thendency or 'average' value is the powerful tool of analysis of data that represents the entire mass of data. The word 'average' is commonly used in day to day conversation. For example, we often talk of average income, average age of employee, average height, etc. An 'average' thus is a single value which is considered as the most representative or typical value for a given set of data. Such a value is neither the smallest nor the largest value, but is a number whose value is somewhere in the middle

of the group. For this reason an average is frequently referred to as a *measure of central tendency of central value.* Measures of central tendency show the tendency of some central value around which the data tends to cluster.

Characteristics of a Good Average :

Since an average is a single value representing a group of values, it is desirable that such a value satisfies the following properties :

(i) **It should be easy to understand :** Since statistical methods are designed to simplify complexity, it is desirable that an average be such that it can be readily understood, otherwise, its use is bound tobe very limited.

(ii) **It should be simple to comput :** Not only an average should be easy to understand but also it should be simple to comput so that it can be used widely. However, though ease of computation is desirable, it should not be sought at the expense of other advantages, i.e. if in the interestof great accuracy, use of more difficult average is desirable one should prefer that.

(iii) **It should be based on all the observations :** The average should depend upon each and every observation so that if any of the observation is dropped average itself is altered.

(iv) **It should be rigidly defined :** An average should be properly defined so that it has one and only one interpretation. It should preferably by defined by an algebraic formula so that if different people compute the average for the same figures they all get the same answer (barring arithmetical mistakes).

(v) It should be capable of further algebraic treatment : We should prefer tohave an average that could be used for further statistical computation. For example, if we are given separately the figures of average income and number of employees of two or more factories, we should be able to compute the combined average.

(vi) It should not be unduly affected by the presence of extreme values : Although each and every observation should influence the value of the average, none of the observation should influence it unduly. If one or two very small or very large observations unduly affect the average i.e. either increase its value or reduce its value, the average cannot be really typical of the entire set of data. In the words, extremes may distort the average and reduce its usefulness.

The following are important measures of central tendency which are generally used in business :

(a) Arithmetic Mean :

The arithmetic mean (usually denoted by the symbol $\bar{x}$ of a set of observations is the value obtained by divideing the sum of all observations in a series by the number of items constituting the series.

Computation of Arithmetic Mean :

1. Un-grouped Data : If $x_1, x_2, ...x_n$ are the n given observations, then their arithmetic mean usuallydenoted by $\overline{X}$ is given by :

$$\overline{X} = \frac{x_1+x_2+.....x_n}{n} \qquad \frac{\Sigma x}{n} \qquad \text{or}$$

The symbol Σ(Greek letter called Sigma) denotes the sum of n items. In normal use only Σx is written in place of $\Sigma x_j (i=1...n)$. However, when the sum is combined to a given range of numbers out of the total, then it becomes necessary to specify.

Problem 1:

The following gives the marks obtained by 10 students at an examination :

Roll Nos.	:	1	2	3	4	5	6	7	8	9	10
Marks obtained	:	43	48	55	57	21	60	37	48	78	59

Calculate the arithmetic mean.

Solution : Computation of Arithmetic Mean

Roll No.	***Marks obtained (x)***
1	*43*
2	*48*
3	*55*
4	*57*
5	*21*
6	*60*
7	*37*
8	*48*
9	*78*
10	*59*
Total	$\Sigma x = 506$

Arithmetic Mean

$$\overline{X} = \frac{\Sigma x}{n}$$

$$= \frac{1}{10} \text{ X } 506$$

$$\overline{X} = 50.6 \text{ Ans.}$$

2. Frequency Distribution : In case of a frequency distribution. The arithmetic mean is given by the following formulae:

$$= \frac{f_1x_1+f_2x_2+.....f_nx_n}{f_1+f_1+.....f_n} = \frac{\Sigma f_ix_i}{\Sigma f_i} = \frac{\Sigma f_ix_i}{N}$$

Where $N=\Sigma f_i$ is the total frequency. The mean value obtained in this manner is sometimes referred as *weighted arithmetic mean,* as distinct from *simple arithmetic mean.*

In case of continuous or grouped frequency distribution, the value of x is taken as the mid-value of the corresponding class.

Problem 2 :

From the following data of marks obtained by 50 students of a class, calculated the arithmetic mean :

Marks	***No. of Sutdents***	***Marks***	***No. of Students***
20	8	50	5
30	12	60	6
40	15	70	4

Let the marks be denoted by X and the number of students by f.

Solution :

Marks(x)	***No. of Students (f)***	***fx***
20	8	160
30	12	360

40	15	600
50	5	250
60	6	360
70	4	280
		2010

$$\bar{X} = \frac{\Sigma fx}{N} = \frac{2010}{50} = 40.2 \text{ marks}$$

Hence the average marks is 40.2.

Problem 3 : Calculate the mean for the following frequency distribution :

Sales (in Rs. lakh) : 0-10 10-20 20-30 30-40 40-50 50-60 60-70

No. of firms : 6 5 8 15 7 6 3

Solution : Computation of Arithmetic Mean

Sales (in Rs. lakh)	***Mid-Value (X)***	***No. of Firms (f)***	***fX***
0-10	5	6	30
10-20	15	5	75
20-30	25	8	200
30-40	35	15	525
40-50	45	7	315
50-60	55	6	330
60-70	65	3	195
		Σf=50	ΣfX=1670

$$A.M. = \frac{\Sigma fx}{N} = \frac{1670}{50} = \text{Rs. } 33.4 \text{ lakhs.}$$

Short-cut Method : When the short-cut method is used arithmetic mean is computed by applying the formula given below :

$$X = A + \frac{\Sigma fd}{N}$$

Where, A = assumed mean and d=deviations from assumed mean (m-A).

Problem 3 will be solved as follows when short-cut method is used :

Mid value (m)	:	5	15	25	35	45	55	65
Deviations (m-A)	:	-30	-20	-10	0	10	20	30
A = 35f	:	6	5	8	15	7	6	3
fd	:	-180	-100	-80	0	70	120	90

Here : A = 35, N = 50, $\Sigma fd = -80$,

$$\bar{x} = 35 + \frac{-80}{50}$$

$$= 33.4 \text{ lakhs.}$$

Step-deviation Method : In the step deviation method the only additional point is that in order to simplify calculations we take a common factor from the data

and multiphy the result by the common factor. The formula is :

$$X = A + \frac{C}{N} \Sigma fd$$

Where A = assumed mean; F = frequency; $D' = \frac{(m - A)}{(C)}$;

C= common factor, N = Total number of observations.

The step deviation method is most commonly used formula as it facilitate calculations.

Problem 4 :

The following table gives the individual output of 180 female workers at a particular plant during a work. Find out the average output per worker.

Output (in units)	500-509	510-519	520-529	530-539
No. of workers	8	18	23	37
Output (in units)	540-549	550-559	560-569	570-579
No. of workers	47	26	16	6

Solution :

Mide-value (m)	**Frequency (*f*)**	$\mathbf{D_1 = \frac{m-534.5}{10}}$	***f*d'**
504.5	8	-3	-24
514.5	18	-2	-36
524.5	23	-1	-23

534.5	37	0	0
544.5	47	1	47
554.5	26	2	52
564.5	16	3	48
574.5	5	4	20
	180		$\Sigma fd = 84$

Average output : $\overline{X} = A + \frac{C}{N} \times \Sigma fd$

$$= 534.5 + \frac{10}{180} \times 84$$

$$= 534.5 + 4.67$$

$$= 539.17 \text{ units}$$

4. Mean of the Combined Series

If a group of n_1 observations has A.M. $\overline{X}_1$ and another group of n_2 observations has A.M. $\overline{X}_2$, then the A.M. ($\overline{X}_{12}$) of the composite group of $n_1 + n_2$ (=n, say) observations is given by

$$\overline{X}_{12} = \frac{n_1 \overline{X}_1 + n_2 X_2.}{n_1 + n_2}$$

where, $\overline{X}_{12}$ = combined mean of the two series or two groups of data. This can be generalised to any number of groups.

Problem 5 : The mean height of 25 male workers in a factory is 61 cms., and the mean height of 25 female workers in the same factory is 58 cms. Find the combined mean height of 50 workers in the

Solution :

$$\overline{X}_{12} = \frac{n_1 \overline{x}_1 + n_2 \overline{x}_2}{n_1 + n_2}$$

$n_1 = 25, \overline{x}_1 = 61, n_2 = 25\ \overline{x}_2 = 58$

$$\overline{x}_{12} = \frac{25x61 + 25x58}{50+50} = \frac{1525 + 1450}{50} = \frac{2975}{50} = 59.6$$

Thus combined mean height of 50 workers is 59.5 cms.

Merits and Limitations of Arithmetic Mean

The arithmetic mean is the most popular average in practice. It is due to the fact that it possess first five out of six characteristics of a goods average (as discussed earlier) and no other average possesses such a large number of characteristics.

However, arithmetic mean is unduly affected by the presence of extreme values. Also in open-end frequency distribution it is difficult to compute mean without making assumption regarding the size of the class-interval of the open-end classes.

Mathematical Properties of Arithmetic Mean

The following area a few important mathematical properties of the arithmetic mean.

1. The sum of the deviations of the items form the arithmetic mean (taking signs

into account) is always zero. i.e. $\Sigma(x-\bar{x}) = 0$.

2. The sum of the squared deviations of the items from arithmatic mean is minimum, that is, less than the sum of the squared deviations of the items from any other value.

3. If we have the arithmetic mean and number of items of two or more than two related groups, we can compute combined avarage of these groups.

(B) MEDIAN : In the words of L.R. Conner : "The median is that value of the variable which divides the data in two equal parts, one part comprising all the values greater and the other, all values less than median." Thus, as against arithmetic mean which is based on all the items of the distribution, the median is only positional average, i.e. the value depends on the position occupied by a value in the frequency distribution.

Computation of Median

1. Ungrouped data : If the number of observation is odd, then the median is the middle value after the observations have been arranged in ascending or descending order of magnitude. In case of even number of observations median is obtained as the arithmetic mean of two middle observations after they are arranged in ascending or descending order of magnitude.

Problem 6: The marks obtained by 12 students out of 50 are : 25, 20, 23, 32, 40, 27, 30, 25, 20, 10, 15, 41

Solution : The values obtained by 12 students arranged in ascending order as : 10, 15, 20, 20, 23, 25, 25, 27, 30, 32, 40, 41

Here the number of items 'N' = 12, which is even

$\therefore$ The two middle items are 6th and 7th items

i.e. $\frac{25+25}{2} = 25$ is the median value.

2. Frequency (Discrete) Distribution :

In case of frequency distribution where the variables take the value $X_1, X_2, ..., X_n$ with respective frequencies $f_1, f_2, ..., f_n$ with $N= \Sigma f$, median is the size of the ½(N+1)th item or observation. In this case the use of comulative frequency (c.f.) distribution facilitates the calculations. The steps involved are :

(i) Prepare the less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Find the c.f. just greater than N/2.

(iv) The corresponding value gives the median.

Problem 7: From the following data find the value of median :

Income (Rs.)	1000	1500	800	2000	2100	1700
No. of Persons	24	26	14	10	5	28

Solution

Income arranged in ascending order	***No. of persons (f)***	***c.f.***
800	14	14
1000	24	38
1500	26	64
1700	28	92
2000	10	102
2100	5	107

Median = Size of (N/2)th item = $\frac{107}{2}$ = 53.5

53.5th item is consisted in the c.f. = 64. The corresponding value to this = 1500. Hence Median = Rs. 1500.

3. Continuous Frequency Distribution : Steps involved for its computation are :

(i) Prepare less than cumulative frequency (c.f.) distribution.

(ii) Find N/2.

(iii) Locate c.f. just greater than N/2.

(iv) The corresponding class contains the median value and is called the median class.

(v) The value of median is now obtained by using the interpolation formula :

$$\text{Median (Md)} = 1 + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where 1 is the lower limit or boundary of the median class;

f is the frequency of the median class;

h is the magnitude or width of class interval;

n = Σf is the total frequency; and

C is the cumulative frequnency of the class preceding the median class.

Problem 8: The annual profits (in Rs. lacs) shown by 60 firms are given below :

Profits :	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65

No. of firms : 4 5 11 6 5 8 9 6 4 2

Calculate the median.

Solution :

Profits	*No.of firms (f)*	*Cumulative frequency (c.f.)*
15-20	4	4
20-25	5	9
25-30	11	20
30-35	6	26
35-40	5	31
40-45	8	39
45-50	9	48
50-55	6	54
55-60	4	58
60-65	2	60

$$\text{Median item} = \frac{1}{2} N = 30$$

The cumulative frequency just greater than 30 is 31 and is corresponding class 35-40 is the median class.

$$\therefore \text{Median} = L + \frac{N/2\text{-c.f.}}{f} \times h$$

$$= 35 + \frac{30-26}{5} \times 5 = 39 \text{ marks.}$$

Merits and Limitations of Median

The median is superior to arithmetic mean in certain aspects. For example, it is specially useful in case of open-ended distribution and also it is not influenced by the presence of extreme values. In fact when extreme values are present in a series, the median is more satisfactory measure of central tendency than the mean.

However, since median is positional average, its value is not determined by each and every observation. Also median is not capable of algebric treatment. For example, median cannot be used for determining the combined median of two or more groups. Furthermore, the median tends to be rather unstable value if the number of observations is small.

(C) MODE : Mode is the value which occurs most frequntly in the set of observations.

Computation of Mode

(i) Ungrouped Data : In case of a frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

Problem 9: Calculate the mode of the following data :

Sr. No.	*Marks obtained*	*Sr. No.*	*Marks obtained*
1	16	6	27
2	27	7	20
3	24	8	18
4	12	9	15
5	27	10	15

Solution. : Calculation of Mode

Size of item (Marks)	*No. of times it occurs*	*Size of item (Marks)*	*No. of times it occurs*
12	1	20	1
15	2	24	1
16	1	27	3
18	1		

Since the item 27 occurs the maximum number of times i.e. 3, hence the modal marks are 27.

(ii) Grouped Data : From the grouped frequency distribution, it is relatively difficult to find the mode accurately. However, if all classes are of equal width, mode is usually calculated by the formula :

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

Where, L = the lower limit or boundary of the modal class;

h = magnitude or width of the modal class'

$\Delta_1 = f_1 - f_0$, $\Delta_2 = f_1 - f_2$;

f_1 = frequency of the modal class;

f_0 = frequency of the class preceding the modal class; and

f_2 = frequency of the class succeeding the modal class.

Mode is generally abbreviated by the symbol M_0.

The above formula takes the following form :

$$\text{Mode (Mo)} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times h = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h \quad \ldots(1)$$

Problem 10 : Calculate mode from the following data :

Marks	*No. of students*	*Marks*	*No. of students*
Above 0	80	Above 60	28
Above 10	77	Above 70	16
Above 30	65	Above 80	10
Above 40	55	Above 90	8
Above 50	43	Above 100	0

Solution :

Since this is cumulative frequency distribution, we are to first convert it into a simple frequency distribution.

Marks	*No. of students*	*Marks*	*No. of students*
0-10	3	50-60	15
10-20	5	60-70	12
20-30	7	70-80	6
30-40	10	80-90	2
40-50	12	90-100	8

By inspection the modal class is 50-60.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$L = 50$, $\Delta_1 = (15-12) = 3$, $\Delta_2 = (15-12) = 3$, $i = 10$

$$M_0 = 50 + \frac{3}{3+3} \times 10 = 50+5 = 55 \text{ Marks.}$$

(iii) Empirical Relation Between Mean (X), Median (Md) and Mode (M_0)

In case of a symmetrical distribution mean, median and mode coincide i.e. Mean = Median = Mode. However, for a moderately asymmetrical (non-symetrical) distribution, mean and mode usually lie on the two ends and median lies in between them and they obey the following important empirical relationship, given by Prof. Karl Pearson.

Mode = 3 Median - 2 Mean ---------- (2).

While applying the formula (1) for calculating mode, it is necessary to see that class intervals are uniform throughout. If they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

A distribution having only one mode is called unimodel. If it contains more than one mode, it is called bimodal or multimodal. In the latter case the values of the mode cannot be determined by formula (1) and hence mode is ill-defined when there is more than one value of mode. Where mode is ill-defined, its value is ascertained by the

formula (2) based upon the relationship between mean, median and mode.

Mode = 3 Median - 2 Mean

Merits and Limitations of Mode

Like Mean, the mode is not affected by extreme values and its value can be obtained in open-end distribution without ascertaining the class limits. Mode can be easily used to describe qualitative phenomenon. For example, when we want to compare the consumer preferences for different types of products, say, soap, toothpastes etc. or different media of advertising, we should compare the modal preferences. In such distributions where there is an outstanding large frequency, mode happens to be meaningful as an average.

However, mode is not rigidly defined measure as there are several formulae for calculating the mode, all of which usually give somewhat different answer. Also the value of mode cannot always be computed, such as, in case of binomial distributions.

D. GEOMETRIC MEAN : The Geometric mean (usually abbreviated as G.M.) of a set of n observations is the nth root of their product.

Computation of Geometric Mean

The Geometric Mean G.M. of n observations X_i, i=1, 2, ..., n is G.M. = $(X_1.X_2. .....X_n)^{1/n}$

The computation is facilitated by the use of logarithms. Taking logarithms of both sides, we get.

$$\log \text{G.M.} = \frac{1}{n} (\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

$$\therefore \text{G.M.} = \text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right) \text{ or antilog}\left(\frac{1}{n}\Sigma\log\right)$$

Problem 11 : From the data given below calculate the G.M.

15, 250, 15.7, 157, 1.57, 105.7, 10.5, 1.06, 25.7, 0.257

Solution :

Value (x)	*Log (x)*
15	1.1761
250	2.3979
15.7	1.1959
157	2.1959
1.57	0.1959
105.7	2.0240
10.5	1.0212
1.06	0.0253
25.7	1.4099
0.257	0.0409
Total	**11.0520**

$$\text{G.M.} = \text{Antilog}\left(\frac{1}{n}\Sigma\log x\right)$$

$$\text{G.M.} = \text{Antilog}\left(\frac{11.0520}{10}\right) = 12.75$$

In case of frequency distribution x_i/f_i (i=1, 2, ..., n)geometric mean, G.M. is given by

$$G.M. = n\sqrt{(x_1.x_1 \text{ ..... } f_1 \text{ times})\,(x_2.x_2 \text{ ..... } f_2 \text{ times}) \text{ ..... } (x_n.x_n \text{ ..... } f_n \text{ times})}$$

Since the product of the values in a frequency distribution is usually very large, formula (3) is not suitable in computing the value of G.M. Taking logarithm of both sides in (3), we get :

$$\log G.M. = \frac{1}{N}\{\log (x_1^{f1}.x_2^{f2} \text{ ..... } x_n^{fn})\}$$

$$= \frac{1}{N}\{f_1 \log x_1 + f_2 \log x_2 + \text{ ..... } + f_n \log x_n)$$

Problem 12 : Calculate Geometric Mean of the following distribution.

X	:	70	100	103	107	149
f	:	10	12	8	5	5

Solution :

X	***f***	***log x***	***f log x***
70	10	1.8451	18.4512
100	12	2.0000	24.0000
103	8	2.0128	16.1024
107	5	2.0294	10.1470
149	5	2.1732	10.8690
			79.5664

$$\text{Log G.M.} = \frac{\Sigma f \log x}{\Sigma f} = \frac{79.5664}{40} = 1.9892$$

$\therefore$ GM = Antilog (1.9892) = 97.54

In the case of grouped frequency distribution, the value of x are the mid-values of the corresponding classes.

Combined Geometric Mean

Just as we have talked of combined arithmetic mean, in a similar manner we can also talk of combined geometric mean. If the Geometric mean of N observations is G.M. and these observations are divided into two sets containing N_1 and second containing N_2 observations having GM_1 and GM_2 as the respective geometric means, then :

$$\log GM = \frac{N_1 \log GM_1 + N_2 \log GM_2}{N_1 + N_2}$$

Merits and Limitations of Geometric Mean

Geometric mean is highly useful in averages, ratios, percentages and in determining rates of increase and decrease. It is also capable of algebraic manipulation. For example, if the geometric mean of two or more series and thier number of observations are known, a combin geometric mean can easily be calculated.

However, compared to arithmetic mean, this average is more difficult to compute and interpret. Also geometric mean cannot be computed when odd number of observations are negative.

E. HARMONIC MEAN : Harmonic mean of a number of observations is the reciprocal of arithmetic mean of reciprocals of the given values.

Computation of Harmonic Mean : If $X_1, X_2, ..., X_n$ are the n observations, their harmonic mean (abbreviated as H) is given by :

$$\text{H.M. } (H) = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Problem 13 : Find the Harmonic mean from the following :

2574, 475, 75, 5, 0.8, 0.08, 0.005, 0.0009

Solution :

X	***1/x***	***X***	***1/X***
2574	0.0004	0.8	1.2500
475	0.0021	0.08	12.5000
75	0.0133	0.005	200.0000
5	0.2000	0.0009	1111.1111
			$\Sigma(1/x)=1325.0769$

$$\text{H.M.} = \frac{n}{\Sigma(1/x)} = \frac{8}{1325.0769} = 0.006$$

In case of frequency distribution, we have

$$\frac{1}{H} = \frac{1}{N}\left[\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots\ldots + \frac{f_n}{x_n}\right] \text{ where } N=\Sigma f$$

Problem 14 : The following table gives weights of 31 persons in a sample enquiry. Calculate mean by using Harmonic mean.

Weight (in 1bs) :	130	135	140	145	146	148	149	150	157
No. of persons :	3	4	6	6	3	5	2	1	1

Solution :

(Weight(x)	*Frequency(f)*	*1/x*	*f(1/x)*
130	3	.00769	.02307
135	4	.00741	.02964
140	6	.00714	.04284
145	6	.00690	.04140
146	3	.00685	.02055
148	5	.00676	.03380
149	2	.00671	.01342
150	1	.00667	.00667
157	1	.00637	.00637
	31		.21776

$$\frac{1}{\text{H.M.}} = \frac{\text{Sf. } 1/x}{N} = \frac{.21776}{31} = .007024$$

$$\text{Or H.M.} = \frac{1}{.007024} = 142.4 \text{ lbs.}$$

The harmonic mean is restricted in its field of application. The harmonic mean is a measure of central tendency for data expressed as rates, for instance - kms. per hour, tonnes per day, kms per litre etc.

Merits and Limitaions of Harmonic Mean

The harmonic mean, like the arithmetic mean and geometric mean is computed from all observations. It is useful in special cases for averaging rates. However, harmonic mean gives largest weight to smallest observations and as such is not a good representation of a statistical series. In dealing with business problems harmonic mean is rarely used.

Do yourself

Q1. What are the measures of central tendency? Why are they called measures of central tendency?

Q2. What are the properties of a good average?

Q3. Give a brief note of the measures of central tendency together with their merits and demerits. Which is the best measured of central tendency and why?

Q4. Following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate median.

Overtime Hours :	10-15	15-20	20-25	25-30	30-35	35-40
No. of Employees :	11	20	35	20	8	6

Q5. The mean monthly salary paid to all employees in a company is Rs. 1600. The mean monthly salaries paid to technical and non-technical employees are Rs. 1800 and Rs. 1200 respectively. Determine the percentage of technical and non-technical employees of the company.

Q6. Calculate the arithmetic mean and the median of the frequency distribution given below. Also calculate the mode using the empirical relation among the three :

Class Limits	*Frequency*	*Class Limits*	*Frequency*
130-134	5	150-154	17
135-139	15	155-159	10
140-144	28	160-164	1
145-149	24		

Q7. In a certain factory a unit of work is completed by A in 4 minutes, by B in 5 minutes, By C in 6 minutes, by D in 10 minutes and by E in 12 minutes.

(a) What is the average number of units of work completed per minute?

(b) At this rate how many units will they complete in a six-hour day.

Q8. Find the average rate of increase in population which in the first decade increased by 20%, in the second decade by 30% and in the third decade by 40%.

Q9. In a class of 50 students, 10 have failed and their average of marks is 2.5. The total marks secured by the entire class were 281. Find the average marks of those who have passed.

SUGGESTED READINGS :

(i) Statistical Method By S.P. Gupta.

(ii) Statistics for Management By Levin.

(iii) Introduction to Statistical Methods By C.B. Gupta.

(iv) Statistics for Business and Economics by R.P. Hooda.

❄ ❄ ❄

Lesson : 3

MEASURES OF DISPERSION

Author : **Vetter:**

Dr. Pradeep Gupta **Dr. B. S. Bodla**

Plan of the Lesson :

1. Introduction
2. Definition
3. Uses of measures of dispersion
4. Properties of a good measure of dispersion
5. Various measures of dispersion
6. Variance
7. Coefficient of Variation
8. Relation between standard deviation, mean deviation and quartile deviation
9. Comparison of the various measures of dispersion
10. Self test questions
11. Suggested readings

1. INTRODUCTION

The value given by a measure of central tendency is considered to be the representative of the whole data. However, it can describe only one of the important characteristics of a series. It does not give the spread or range over which the data are scattered. Measures of dispersion are used to indicate this

spread and the manner in which data are scattered.

2. DEFINITION

Dispersion indicates the measure of the extent to which individual items differ from some central value. It indicates lack of uniformity in the size of items.

Some important definitions of dispersion are given below :

(1) According to Spiegel, "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data."

(2) Simpson and Kafka defines dipersion as "The measurement of the scatternes of the mass of figures in a series about an average is called measure of variation or dispersion."

(3) As defined by Brooks and Dick, "Dispersion or spread is the degree of the scatter or variation of the variable about a central value."

Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the *second order.*

3. USES OF MEASURES OF DISPERSION

Average is a typical value but it alone does not describe the data fully. It does not tell us how the items in a series are scattered around it. To clear this point consider the following three sets of data :

Set A	30	30	30	30	30
Set B	28	29	30	31	32
Set C	3	5	30	37	75

All the three sets. A, B and C have mean 30 and median is also 30. But by inspection it is apparent that the three sets differ remarkably from one another. Thus to have a clear picture of data, one needs to have a measure of dispersion or variability (scatteredness) amongst observations in the set. It is also used for comparing the variability or consistency (uniformity) of two or more series. A higher degree of variation means smaller consistency.

4. PROPERTIES OF A GOOD MEASURE OF DISPERSION

There are various measures of dispersion. The difficulty lies in choosing the best measure as no hard and fast rules have been made to select any one. However, some norms have been set which work as guidelines for choosing a particular measure of dispersion. A measure of dispersion is good or satisfactory if it possesses the following characteristics.

(i) It is easily understandable.

(ii) It utilises all the data.

(iii) It can be calculated with reasonable ease and rapidity.

(iv) It affords a good standard of comparison.

(v) It is suitable for algebraic and arithmatical manipulation.

(vi) It is not affected by sampling variations.

(vii) It is not affected by the extreme values.

5. VARIOUS MEASURES OF DISPERSION

Commonly used measures of dispersion are :

5.1 Range

5.2 Quartile deviation

5.3 Mean deviation

5.4 Standard deviation

5.1 Range

Definition. Range is the difference between the two extreme items, i.e. it is the difference between the maximum value and minimum value in a series.

Range (R) = Largest value (L) minus Smallest value (S)

A relative measure known as *coefficient of range is* given as:

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

Lesser the range or coefficient of range, lower the variability.

Properties.

(a) It is the simplest measure and can easily be understood.

(b) Besides the above merit, it hardly satisfies any property of a good measure of dispersion e.g. it is based on two extreme values only, ignoring the others. It is not liable to further algebraic treatment.

Example 1. The population (in '000) in eighteen Panchayat Samities of a district is as given below :

77, 76, 83, 68, 57, 107, 80, 75, 95, 100, 113, 119, 121, 121, 83, 87, 46, 74

Calculate the range and coefficient of range.

Solution. Largest value (L) = 121

Smallest value (S) = 46

Range (R) = L - S

= 121-46 = 75

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{121-46}{121+46} = \frac{75}{167} = 0.449$$

Range for grouped data. In case of grouped data, the range is the difference between the upper limit of the highest class and the lower limit of the lowest class. No consideration is given to frequencies.

Example 2. Find range of the following distribution.

Class-interval	45-49	50-54	55-59	60-64	65-69
Frequency	37	26	8	5	1

Solution. The series can be written as follows :

Group	Frequency
44.5-49.5	37
49.5-54.5	26
54.5-59.5	8
59.5-64.5	5
64.5-69.5	1

Range = 69.5-44.5 = 25

5.2 Quartile Deviation

Quartile deviation is obtained by dividing the difference between the upper quartile and the lower quartile by 2.

$$\text{Quartile deviation or Q.D.} = \frac{\text{Upper Quartile - Lower Quartile}}{2}$$

$$= \frac{Q_3 - Q_1}{2}$$

The coefficient of quartile deviation is given by the following formula :

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 - Q_1}$$

Coefficient of quartile deviation is a relative measure of dispersion and is used to compare the variability among the middle 50 per cent observations.

Properties.

(i) It is better measure of dispersion than range in the sense that it is based on the middle 50 per cent observations of a series of data rather than only two extreme values of a series.

(ii) It excludes the lowest and the highest 25% values.

(iii) It is not affected by the extreme values.

(iv) It can be calculated for the grouped data with open end intervals.

(v) It is not capable of further algebraic treatment.

(vi) It is not considered a good measure of dispersion as it does not show the scattering of the central value. In fact it is a measure of partitioning of distribution. Hence it is not commonly used.

Example 3. Given the number of families in a locality according to monthly per capita expenditure classes in rupees as:

Class-interval	140-150	150-160	160-170	170-180	180-190	190-200
No. of familes	17	29	42	72	84	107

200-210	210-220	220-230	230-240	240-250
49	34	31	16	12

Calculate Quartile deviation and coefficient of quartile deviation.

Solution.

Monthly per capita expenditure (Rs.)	Number of Families (f)	Cumulative frequency (c.f.)
140-150	17	17
150-160	29	46
160-170	42	88
170-180	72	160
180-190	84	244
190-200	107	351
200-210	49	400
210-220	34	434
220-230	31	465
230-240	16	481
240-250	12	493

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

(i) To calculate Q_1, we have to first find :

$$\frac{N}{4} = \frac{493}{4} = 123.25$$

The number 123.25 is contained in the cummulative frequency 160. Hence the first quartile lies in the class 170-180. By using the formula for Q_1 we get,

$$Q_1 = L + \frac{N/4 - \text{c.f.}}{f} \times i$$

$L = 170, N/4 = 123.25, \text{c.f.} = 88, f=72, i=10$

$$Q_1 = 170 + \frac{123.25 - 88}{72} \times 10 = \text{Rs. } 174.90$$

(ii) To calculate Q_3 we find :

$$\frac{3N}{4} \qquad \frac{3 \times 493}{4} \qquad = 369.75$$

The number 369.75 is contained in the cummulative frequency 400. Hence the class 200-210 is the third quartile class. By using the formula for Q_3 we get :

$$Q_3 = L + \frac{3N/4 - c.f.}{f} \times i$$

L= 200, 3 N/4 = 369.75, c.f. = 351, f=49, i=10

$$Q_3 = 200 + \frac{369.75-351}{49} \times 10 = \text{Rs. } 203.83$$

$$\text{Quartile deviation (Q.D.)} = \frac{203.83-174.90}{2} = \frac{28.93}{2} = 14.465$$

$$\text{Coefficient of Q.D.} = \frac{203.83-174.90}{203.83+174.90} = \frac{28.93}{378.73} = 0.076$$

5.3 Mean deviation

Mean deviation is the mean of deviations of the items from an average (mean, median or mode). As we are concerned with the deviations of the different values from an average and in finding the mean of deviations, we have to find the sum of deviations (whether positive or negative), we take all the deviations as positive. We are concerned with the deviations and not with their

algebraic signs. We ignore negative signs because the algebraic sum of the deviations of individual values from the average is zero.

Calculation of mean deviation (M.D.).

Mean deviation of a set of n observations $x_1, x_2, ..., x_n$ is calculated as follows :

$$M.D. = \frac{1}{n} \sum_{i=1}^{n} |x_i - A|$$

for i= 1,2, ..., n where A is a central value.

Let $|x_i - A| = d_i$

$$\text{Then M.D.} = \frac{1}{n} \sum_{i=1}^{n} |di| \qquad \text{..... (i)}$$

In case data is given in the form of a frequency distribution, the variate values $x_1, x_2, ..., x_n$ occur $f_1, f_2, ..., f_n$ times respectively.

In such series the formula for mean deviation is,

$$M.D. = \frac{1}{N} \sum_{i=1}^{n} f_i |x_i - A| \qquad \text{..... (ii)}$$

Where, $N = \Sigma$ fi for i = 1, 2, ..., n

In case of grouped data, the mid-point of each class interval is treated as x_i and we can use the formula (ii) given above.

Properties.

(i) Mean deviation removes one main objection of the earlier measures, that it involves each value of the set.

(ii) Its main drawback is that algebraic negative signs of the deviations are ignored which is mathematically unsound.

(iii) Mean deviation is minimum when the deviations are taken from median.

(iv) It is not suitable for algebraic treatment.

Example. 4 :

Calculate mean deviation from the mean for the following data:

Size (x) :	2	4	6	8	10	12	14	16
Frequency :	2	2	4	5	3	2	1	1

Solution :

x	f	fx	$\|x-8\|$ $\|D\|$	$f\|d\|$
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8
	N=20	Σ fx=160		Σ f|D|= 56

$$X = \frac{\Sigma fx}{N} = \frac{160}{20} = 8$$

$$M.D. = \frac{\Sigma f|D|}{N} = \frac{56}{20} = 2.8$$

Examples 5. Calculate the mean deviation (using median) from the following data.

Size of items	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution.

Size	Frequency (f)	Cummulative frequency	Deviation from median 9 $\|d\|$	$f\|d\|$
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
			$\Sigma f\|d\|$ =	60

$$\text{Median} = \text{Size of } \frac{48+1}{2} \text{ th item}$$

$$= \text{Size of } 24.5^{th} \text{ item} = 9$$

$$\text{Mean deviation} = \frac{\Sigma f|d|}{N} = \frac{60}{48} = 1.25$$

Example. 6 : Find the median and mean deviation of the following data :

Size	*Frequency*	*Size*	*Frequency*
0-10	7	40-50	16

10-20	12	50-60	14
20-30	18	60-70	8
30-40	25		

Solution : Calculation of Median and Mean Deviation.

Size	*f*	*c.f.*	*m.p.*	$\|m-35.2\|$ $\|D\|$	$f\|D\|$
0-10	7	7	5	30.2	211.4
10-20	12	19	15	20.2	242.4
20-30	18	37	25	10.2	183.6
30-40	25	62	35	0.2	5.0
40-50	16	78	45	9.8	156.8
50-60	14	92	55	19.8	277.2
60-70	8	100	65	29.8	238.4
	N = 100		Σ f D	= 1314.8 $\simeq$	1315

Median = Size of N/2 th item = 100/2 = 50th item.

$\therefore$ Median lies in the class 30-40.

$$\text{Med.} = L + \frac{N/2 - \text{c.f.}}{f} \times i$$

L = 30, N/2 = 50, c.f. = 37, f= 25, i = 10.

$$\text{Med.} = 30 + \frac{50-37}{25} \times 10 = 30 + 5.2 = 35.2$$

$$\text{M.D.} = \frac{\Sigma f D}{N} = \frac{1315}{100} = 13.15$$

Uses of Mean Deviation :

The outstanding advantage of the average deviation is its relative simlicity. It is simple to understand and easy to compute. Any one familiar

with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not familiar with statistics, the average deviation is useful.

5.4 Standard deviation

It is the square root of the quotient obtained by dividing the sum of squares of deviations of items from the Arthmetic mean by the number of observations.

$$\therefore \text{Standard deviation or } \sigma = \sqrt{\frac{\text{Sum of squares of deviation from A.M.}}{\text{Number of observations}}}$$

Standard deviation is an absolute measure of dispersion.

Calculation of standard deviation (σ).

(a) Ungrouped data

First method : S.D. $(\sigma) = \sqrt{\frac{\Sigma d^2}{n}}$

Where d is the deviation of value from the mean.

Second method : In this method, we assume a provisional mean and find the deviations of the values from the provisional mean. The following formula is applied under this method :

$$\text{S.D. or } \sigma = \sqrt{\frac{\Sigma d_x^2}{n} - \left[\frac{\Sigma d_x}{n}\right]^2}$$

Where d is the deviation of values of x observations from the assumed mean. This formula is more useful when values are in decimals and the mean of the series does not happen to be an integer.

In case the frequencies are also given, then standared deviation is calculated by using the formula :

$$\text{S.D. or } \sigma = \sqrt{\frac{\Sigma f d_x^{\,2}}{n} - \left[\frac{\Sigma f d_x}{n}\right]^2} \qquad \text{Where } n = \Sigma f$$

Example 7. Compute the standard deviation by the short method for the following data :

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution. Let us assume that mean is 15.

X	d (x-15)	d^2
11	-4	16
12	-3	9
13	-2	4
14	-1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
	$\Sigma d=11$	$\Sigma d2 = 121$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{121}{11} - \left[\frac{11}{11}\right]^2}$$

$$= \sqrt{11 - 1}$$

$$= \sqrt{10} = 3.16$$

In continuous series, we take the central values of the groups.

Example 8. :

Find the standard deviation of the following distribution :

Age	:	20-25	25-30	30-35	35-40	40-45	45-50
No. of Persons	:	170	110	80	45	40	35

Take assumed average = 32.5

Solution :

Calculation of Standard Deviation.

Age	***m.p. No. of persons (m-32.5)/5***				
	m	***f***	***d***	***fd***	***fd^2***
20-25	22.5	170	-2	-340	680
25-30	27.5	110	-1	-110	110
30-35	32.5	80	0	0	0
35-40	37.5	45	1	45	45
40-45	42.5	40	2	80	160
45-50	47.5	35	3	105	315
		N=480		Σ fd = - 220	Σfd^2 = 1310

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left[\frac{\Sigma fd}{N}\right]^2} \times i$$

$$= \sqrt{\frac{1310}{480} - \left[\frac{-220}{480}\right]^2} \times 5$$

$$= \sqrt{2.729 - .21} \times 5 = 1.587 \times 5 = 7.936$$

Uses of the Standard deviation : As a measure of dispersion, standard deviation is most important. By comparing the standard deviations of two or more series, we can compare the degree of variability or consistency. It is a keystone in sampling and correlation and is also used in the interpretation of normal and skewed curves. It is used to guage the representativeness of the mean also.

Merits of Standard Deviation :

(i) It is suitable for algebraic manipulation.

(ii) It is less erratic.

(iv) Standard deviation is considered to be the best measure of dispersion and is used widely.

Demerits of Standard Deviation :

(i) Its calculation demand greater time and labour.

(ii) If the unit of measurement of variables of two series is not the same, then their variability can not be compared by comparing the values of standard deviation.

(iii) It gives more weight to extreme items and less to those which are nearer the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small. The deviations 2 and 8 are in the ratio of 1 : 4 but their squares i.e. 4 and 64, would be in the ratio of 1 : 16.

Mathematical Properties of Standard Deviation

Standard devation has some very important mathematical properties which considerably enhance its utility in statistical work.

1. *Combined Standard Deviation :* Just as it is possible to compute

combined mean of two or more than two groups, similarly we can also compute combined standard deviation of two or more groups.

2. *Standard deviation of n natural numbers :* The standard deviation of the first in natural numbers can be obtained by the following formula :

$$\sigma = \frac{1}{12} \ (n^2-1)$$

3. The sum of the squares of deviations of items in the series from their arithmetic mean is minimum. This is the reason why standard deviation is always computed from the arithmetic mean.

4. The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located with the help of Teheycheff's theorem, given by mathematician P.L. Tehebycheff (1821-1894). No matter what the shape of the distribution is, at least 75 percent of the values will fall within $\pm$ 2 standard deviation from the mean of the distribution, and at least 89 percent of values will be within + 3 standard deviations from the mean.

For a symmetrical distribution, the following relationships hold good:

Mean $\pm$ 1 σ covers 68.27% of the items.

Mean $\pm$ 2 σ covers 95.45% of the items.

Mean $\pm$ 3 σ covers 99.73% of the items.

6. Variance

The variance is just the square of the standard deviation value :

Variance = $\sigma^2 = (S.D.)^2$

In a frequency distribution where deviations are taken from assumed mean, variance may directly be computed as follows :

$$\text{Variance} = \left\{ \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2 \right. \text{x i}$$

Where $d = \frac{x-A}{i}$ and i = common factor.

Properties :

(i) The main demerit of variance is that its value is the square of the unit of measurement of variate values. For example, the variable x is measured in cms, the unit of variance is cm. Generally, this value is large and makes it difficult to decide about the magnitude of variation.

(ii) The variance gives more weightage to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.

Pooled or combined variance : By the combined variance of two groups, we mean the variance of the observations of the two groups taken together. Let us consider two groups consisting of n_1 and n_2 observations respectively. Suppose the means of the groups are $\bar{x}_1$ and $\bar{x}_2$ and the variances are σ_1^2 and σ_2^2 respectively. We know that the pooled mean of both the groups is,

$$\bar{X}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

The combined variance of the two groups is given by the formula :

$$\sigma_{12}^2 = \frac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

Where, $d_1 = (\bar{x}_1 - \bar{x}_{12})$ and $d_2 = (\bar{x}_2 - \bar{x}_{12})$

The advantage of the formula of combined variance is that once we know the individual mean and variance of each group, we can calculate the variance of

the combined groups without redoing the entire calculations.

Obviously, the combined standard deviation can be found by taking the square root of the combined variance.

Example 9: For a group of 50 male workers, the mean and standard deviation of their weekly wages are Rs. 63 and Rs. 9 respectively. For a group of 40 female workers these are Rs. 54 and Rs. 6, respectively. Find the standard deviation for the combined group of 90 workers.

Solution :

The data is $n_1 = 50 \quad \bar{x}_1 = 63 \quad \sigma_1 = 9$

$n_1 = 40 \quad \bar{x}_2 = 54 \quad \sigma_2 = 6$

Combined mean $\bar{x}_{12}$ for group of 90 $= \dfrac{n_1 x_1 + n_2 x_2}{n_1 + n_2}$

$= [50 \times 63 + 40 \times 54] / 90$

$= [3150 + 2160]/90 = 5310/90 = 59$

Combined standard deviation $= \dfrac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)}{n_1 + n_2}$

Where, $d_1 = (\bar{x}_1 - \bar{x}_{12})$ and $d_2 = (\bar{x}_2 - \bar{x}_{12})$

$\sigma^2_{12} = [50 (81+16) + 40 (36+25)] / 90$

$= [97 \times 50 + 40 \times 61] / 90 = [4850 + 2440] /90$

$= 7290 / 90 = 81$

$\therefore \sigma_{12} = 9$

Example 10 : The analysis of the results of a budget survey of 150 families gave an average monthly expenditure of Rs. 120 on food items with a standard deviation of Rs. 15. After the analysis was completed it was noted that the figure recorded for one household was wrongly taken as Rs. 15 instead of Rs. 105. Determine the correct value of the average expenditure and its standard

deviation.

Solution.

A.M. ($\bar{x}$) = Rs. 120, No. of items = 150

Total as obtained = 120 x 150 = 18000

Correct total = (Total obtained - item misread) + correct item

= (18000-15) + 105 = 18,090

Correct mean = Correct total/No. of items = 18090/150

= Rs. 120.6

$$(S.D.)^2 = \frac{\Sigma x2}{n} - \left[\frac{\Sigma x}{n}\right]^2$$

$$\text{Before correction } (15)^2 = \frac{\Sigma x^2}{150} - (120)^2$$

$$\text{or } 225 = \frac{\Sigma x2}{150} - 14{,}400 \text{ or } \frac{\Sigma x2}{150} = 14{,}625$$

$$\Sigma x^2 = 14{,}625 \times 150$$

Correct sum of squares = Sum of suqares before correction, minus square of misread item plus square of correct item

= 14625x150 - 15 x 15 + 105 x 105

= 15 x 15 [9750-1 + 7x7] = 225

= [9798]

Correct $(S.D.)^2$ = 225x9798/150 - $(120.6)^2$

$= \quad 1.5 \times 9798 - 14544.36$

$= \quad 152.64$

Correct S.D. $= \quad 12.4$

7. Coefficeint of variation (C.V.)

It two series differ in their units of measurement, their variability cannot be compared by any measure given so far. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in size, the coefficient of variation should be used as a measure of dispersion. It is a unitless measure of dispersion and also takes into account the size of the means of the two series. It is the best measure to compare the variability of two series or set of observations. A series with less coefficient of variation is considered more consistent.

Definition. Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100.

If σ is the standard deviation and 0 is the mean of the set of values, the coefficient of variation is,

$$C.V. = \frac{\sigma}{\overline{x}} \times 100$$

This measure was given by Professor Karl Pearson.

Properties :

(i) It is one of the most widely used measure of dispersion because of its virtues.

(ii) Smaller the value of C.V., more consistent are the data and vice-versa.

Hence a series with smaller C.V. than the C.V. of other series is more consistent, i.e. it possess variability.

Example 11: A time study was conducted in a factory with the help of two samples A and B consisting of 10 workers. The time taken by the workers in each case recorded. From the particulars given below state which of the samples is more variable and which takes less time on an average.

Time taken in minutes.

Sample A	130	125	120	135	140	145	130	145	140	150
Sample B	132	146	137	145	130	125	138	140	143	144

Solution : Let us calculate mean and standard deviation first by rearranging the data in ascending order.

For Sample A			For Sample B		
x	$d = \frac{x - 140}{5}$	d^2	y	d'= (y-140)	d'^2
120	-4	16	125	-15	225
125	-3	9	130	-10	100
130	-2	4	132	-8	64
130	-2	4	137	-3	9
135	-1	1	138	-2	4
140	0	0	140	0	0
140	0	0	143	3	9
145	1	1	144	4	16
145	1	1	145	5	25
150	2	4	146	6	36
	Σd=-8	Σd2=40		Σ d' = -20	Σ d'2 = 488

$0 = 140 + \left(\frac{-8}{10} \times 5\right) = 136$ $\qquad$ $y = 140 + \left(\frac{-20}{10}\right) = 138$

This shows that on an average workers from sample A takes less time.

$$\sigma_x^2 = \left\{\frac{\Sigma d2}{n} - \left(\frac{\Sigma d}{n}\right)^2\right\} \times i^2 \text{ where } i = 5$$

$$= 25 \left\{\frac{40}{10} - \left(\frac{8}{10}\right)^2\right\} = 25 \left\{4 - \frac{16}{25}\right\} = 84$$

$$\therefore CV_x = \frac{\sigma x}{\overline{X}} \times 100 = \frac{\sqrt{84}}{136} \times 100 = 6.75\ \%$$

Similarly,

$$\sigma_y^2 = \left\{\frac{\Sigma d'2}{n} \left(\frac{\Sigma d'}{n}\right)^2\right\}$$

$$= \left\{\frac{488}{10} -- \left(\frac{-20}{10}\right)^2\right\}$$

$$= 48.8 - 4 = 44.8$$

$$\therefore CV_y = \frac{\sigma y}{\overline{y}} \times 100 = \frac{44.8}{138} \times 100 = 4.85\ \%$$

Since $CV_x > CV_y$, the sample 'A' is more variable, though as we have seen the workers of sample A takes less time.

8. RELATION BETWEEN STANDARD DEVIATION, MEAN DEVIATION AND QUARTILE DEVIATION

In any bell-shaped distribution, the S.D. will always be larger than M.D.

and M.D. larger than Q.D. if the distribution approximates the form of normal curve, the M.D. will be 4/5 of S.D. and Q.D. will be about 2/3rd as large as S.D. Usually,

M.D. = 4/5 of S.D.

Q.D. = 2/3 of S.D.

9. COMPARISION OF THE VARIOUS MEASURES OF DISPERSION

Range is not a very satisfactory measure of dispersion because it depends solely on the two extreme values and may be very misleading if there are one or two abnormal items. It is impossible to estimate the range in case where there are open ends series. Therefore, it is an unreliable measure of dispersion. Quartile deviation is most easy to calculate and interpret but it is not amenable to mathematical treatment. Mean deviation is easy to compute but grouped data may be difficult. In almost all other aspects, the advantage rests with the standard deviation. Only the S.D. is suitable for algebraic manipulation. For this reason, it is used in correlation, in sampling and in other aspects of advanced statistics. We can compute the S.D. of the whole group if means and standard deviations of two or more subgroups are known. When it is required to compare two or more than two series or distributions, we compute relative measure of dispersion.

10. SELF TEST QUESTIONS

Q1. What does dispersion indicate about the data ? Why is this of great importance ?

Q2. What are the requirements of a good measure of dispersion ?

Q3. Define and discuss the following terms.

(a) Quartile Deviation

(b) Mean Deviation

(c) Variance

(d) Coefficient of Variation

Q4. Calculate mean deviation from median as well as arithmetic mean.

Class-intervals	2-4	4-6	6-8	8-10
Frequencies	3	4	2	1

Q5. Calculate the standard deviation

Age	50-55	45-50	40-45	35-40	30-35	25-30	20-25
No.	25	30	40	45	80	110	170

Q6. Which of the two students was more consistent ?

x	58	59	60	54	65	66	52	75	69	52
y	84	56	92	65	86	78	44	54	78	68

Q7. Calculate the appropriate measure of dispersion.

Wages in rupees	No. of wage earners
Less than 35	14
35-37	62
38-40	99
41-43	18
Over 43	7

Q8. Fill in the blanks :

(a) Mean Deviation is minimum about.....

(b) Variance is zero when.....

(c) Range is zero when.....

(d) Coefficient of variation is infinity when.....

(e) Coefficient of variation is zero when.....

Q9. In a certain distribution with n = 25 on measurements it was found that $\overline{X}$=56 and σ=2. After these results were computed it was discovered that a mistake had been made in one of the measurements which was recorded as 64. Find the mean and standard deviation if the incorrect value 64 is omitted.

Q10. The following table gives the frequency distribution of marks obtained by students of two classes. Find the arithmetic mean, the standard deviation and coefficient of variation of the marks of two classes. Interpret the results.

Range of Marks	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
Class A	1	10	20	8	6	3	1	0
Class B	5	6	15	10	5	4	2	2

11. SUGGESTED READINGS

(i) Statistical Methods by S.P. Gupta.

(ii) Statistics for Management By Levin.

(iii) Introduction to Statistical Methods By C.B. Gupta.

❄ ❄ ❄

Lesson : 4

MOMENTS, SKEWNESS AND KURTOSIS

Author :
Dr. S. S. Tasak

Vetter:
Dr. B. S. Bodla

A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data."

- R. A. Fisher

4.1 Introduction

We have already introduced two parameters which describe the frequency distribution. These are mean x which locates the distribution, and the standard deviation (σ) which measures the scatter of the items about that mean. These two important parameters go a long way in describing the distribution, but there are many features of it which are not brought out. How symmetrical the distribution is about the mean, or how 'peaky' is the distribution are some other features that specify the distribution.

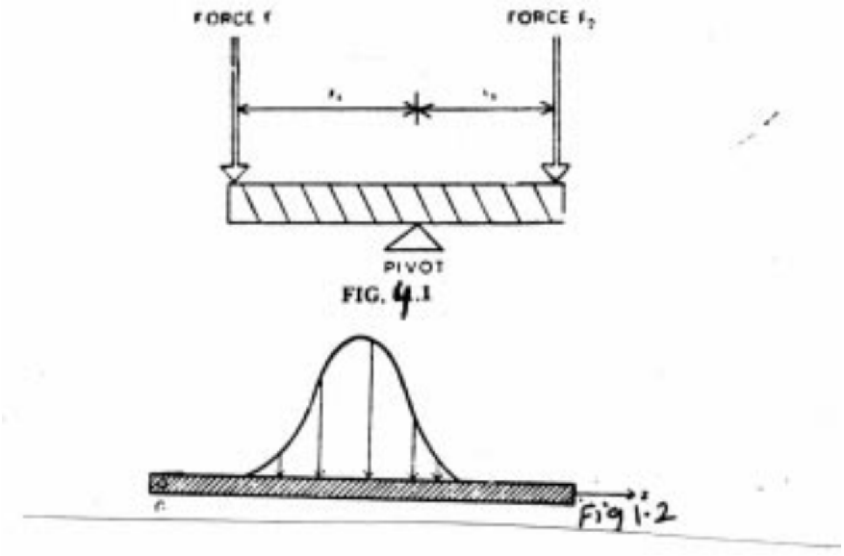
It can be shown that we can define a whole series of measures known as moments which when properly interpreted, give a wealth of information about

the 'shape' of the distribution. It will be seen that the arithmetic mean x and the standard deviation σ are but the first two members of the series. The symmetry and the peakedness of the distribution can also be obtained from the higher members of this series.

Though the concept of moments is a highly mathematical one, an elementary introduction which brings out the physical significance is given here.

4.2 Moments

The term moment is obtained from mechanics where the 'moment of a force' describes the tendency or capacity of a force to turn a pivoted lever (Fig. 4.1). Force F_1 has a counter-clockwise moment F_1-x_1 and force F_2 has a clockwise moment $F_2.x_2$.



For a frequency distribution (Fig. 1.2)we imagine that at a distance x from the

origin 0, a force equal to the frequency (f) associated with x acts and thus the moment about 0 is equal to fx. Taking the contributios of the whol distribution, the moment then is $\sum fx$. This then, is, the moment on the lever produced if the whole distribution was sitting on a lever pivoted at the origin. To correct for the number of items involved (since we are interested in specifying only the 'shape' of the distribution) we divide by $\sum f$, i. e. the total number of items. The 'first' moment about the origin then is nu one (ν_1)prime,

$$\nu_1 = \frac{\sum fx}{\sum f} \qquad (4.1)$$

Just like in mechanics, we can define higher order moments as well by multiplying f by higher powers of x. Thus

$$\nu_2 = \frac{\sum fx^2}{\sum f}$$

$$\nu_3 = \frac{\sum fx^3}{\sum f}, \text{ and so on.}$$

In general

$$\nu_r = \frac{\sum fx^r}{\sum f} \qquad (4.2)$$

Instead of taking the moments about the origin we may also take them about any other point x_0 (equivalent to pivoting the lever bar about that point). Then,

$$\nu_r = \frac{\sum f(x-x_0)^r}{\sum f} \qquad (4.3)$$

Thus ν_r is a special case of ν_r with $x_0 = 0$. The series of moments with $x_0 = x$, i.e., moments about the mean have special significance and are denoted by μ (mu)

$$\nu_r = \frac{\sum f(x-x)^r}{\sum f} \qquad (4.4)$$

4.3 Moments about the mean

The moments of a frequency distribution about the mean x have special significance. We study these in some details here.

The zeroeth moment μ_0 by formula (4.4) is

$$\mu_0 = \frac{\sum f(x-x)^0}{\sum f}$$

But since any number raised to power zero is one, it if clear that

$$\mu_0 = \frac{\sum f}{\sum f} = 1$$

for all distributions.

Similarly,

$$\mu_0 = \frac{\sum f(x-x)}{\sum f}$$

But, by definition of the mean x, the algebraic sum of the deviations about, it, i.e., $\sum f(x-x)$ is zero, so that $\mu_1 = 0$ for all distributions.

Next $\mu_2 = \sum f(x-x)^2/\sum f$ is by definition the variance of the distribution. Thus,

$$\mu_0 = 1$$

$$\mu_1 = 1$$

$$\mu = \sigma^2 \qquad (4.5)$$

for all distributions.

There is yet another point which can be deduced about the moments. We have already seen, while discussing the properties of the arithmetic mean that the sum of deviations below the mean equals the sum of deviations above the mean. This shows that the negative and positive deviations cancel out. This would be so in all distributions whether symmetrical or asymmetrical, when the deviations are raised to the first power.

When the deviations are raised to any even power their signs will all be positive and will no longer cancel out.

When the deviaitons are raised to any odd power (other than 1) and the sum of the negative deviations equals the sum of the positive deviations the distribution is symmetrical. Thus in symmetrical distribution only

$$\mu_3 = 0$$

$$\mu_5 = 0$$

$$\mu_7 = 0 \qquad (4.5)$$

For this reason we can use these moments as measures of asymmetry.

Relation between μ and ν.* If we are given the moments about any arbitary

origin (including 0), then we can compute moments about mean by the following formula :

$$**\mu_1 = \nu_1 - \nu_1 = 0$$

$$\mu_2 = \nu_2 - (\nu_1)^2 = 0$$

$$\mu_3 = \nu_3 - 3\nu_2 x\ \nu_1 + 2(\nu_1)^3$$

$$\mu_4 = \nu_3 - 4\nu_3 x\ \nu_1 + 6\nu_2 x\ (\nu_1)^2 - 3(\nu_1)^4 \qquad (4.6)$$

An important corollary follows from the above, i.e., the mean square deviation about the mean of the observations is less than the mean square deviation about any arbitrary origin. In other words, the mean square deviation or variance (σ^2) about the mean is minimum-smaller than it would be if computed from any other average. So from the equation; since n_2 is positive quantity, being a square μ_2 must be less than n_2.

---

*The method of computing moment about the mean from moments about the arbitrary origin can be easily remembred by understanding the following :

$\mu_1 = (\nu - \nu_1)$ or $(\nu_2 - d)$, Where $d = (X-A) = \frac{\Sigma(fx')}{N} = x\ \nu_1$

$\mu_2 = (\nu-d)^2 = \nu_2 - 2\mu_1 d + d^2$

$= \nu_2 - d^2 = \nu_2 - \nu_1^2$

$\mu_3 = (\nu-d)^2 = \nu_3 - 3\nu_2 d + 3\nu_1 d^2 - d^2$

$= \nu_3 - 3\nu_2\nu_1 + 2\nu_1^2$

$\mu_4 = (\nu-d)^4 = \nu_4 - 4\nu_3 d + 6\nu_2 d^2 - 4\nu_1 d^3 + d^4$

$= \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$

**These measures may be in terms of class interval units or units of one. In the former case we will have to multiply μ_1, μ_2, μ_3 and μ_4 and i, i^2, i^3 and i^4 respectively, where i represents the class interval.

$\mu_2 = \sigma^2$ if both are expressed either in class interval units or in units of one.

4.1 Illustration : Find the first four moments about the mean from the following data.

Class-interval	0-10	10-20	20-30	30-40
Frequency	1	3	4	2

Solution :

Size	*x*	*f*	$x' = \frac{x-25}{10}$	*fx'*	fx'^2	fx'^3	fx'^4
0-10	5	1	-2	-2	4	-8	+16
10-20	15	3	-1	-3	3	-3	+3
20-30	25	4	0	0	0	0	0
30-40	35	2	1	2	2	2	+2
		10		$\sum fx' = -3$	9	-9	21

Moments about Arbitrary Mean

$$\nu_1 = \frac{-3 \times 10}{10} = -3$$

$$\nu_2 = \frac{9}{10} 10 \times {}^2 = 90$$

$$\nu_3 = \frac{-9}{10} \times 10^3 = -900$$

$$\nu_4 = \frac{21}{10} 10 \times {}^4 = 21{,}000$$

Moments about Mean

$\mu_1 = (\nu_1 - \nu_1) = 0$

$\mu_2 = \nu_2 - (\nu_1)^2 = 90\text{-}9 = 81$

$\mu_3 = \nu_3 - 3\nu_2 \times \nu_1 + 2(\nu_1)^3$

$= -900\text{-}270 \times (-3) + 2 \times (-27)$

In other words, we may say that if the extreme variations in a given distribution are towards higher values they give the curve a longer tail to the right and this pulls the median and mean in that direction from the mode. If, however, extreme variations are towards lower values, the longer tail is to the left and the median and mean are pulled to the left of the mode.

It could also be shown that in a symmetrical distribution the lower and upper quartiles are equidistant from the median, so also are corresponding pairs of deciles and percentiles. This means that in a asymmetrical distribution the distance of the upper and lower quartiles from median is unequal.

From the above discussion, we can summarise the tests for the presence of skewness in the following words :

1. When the graph of the distribution does not show a symmetrical curve;
2. When the three measures of central tendency differ from one another;
3. When the sum of the positive deviations from the median are not equal to the negative deviations from the same value;
4. When the distances from the median to the quartiles are unequal; and
5. When corresponding pairs of deciles or percentiles are not equidistant from the median.

4.6 Measures of Skewness

On the basis of the above tests, the following measures of skewness have been developed.

1. Relationship between 3 M's of central tendency-commonly known as the Karl Pearson's measure of skewness.

$= -900 + 810 - 54 = -144$

$\mu_4 = \nu_4 - 4\nu_3 \times \nu_1 + 6\nu_2 \times (\nu_1)^2 - 3(\nu_1)^4$

$= 21,000-4x(-900) \times (-3) + (6x90x9) - 3x81$

$= 21,000-10,800 + 4860-243 = 14,817.$

4.5 Skewness

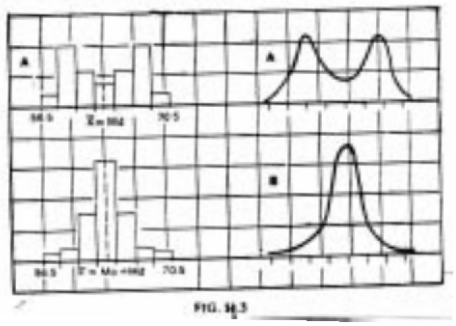
When a frequency distribution is not symmetrical it is said to be asymmetrical or skewed. The nature of symmetry and the various types of asymmetry are illustrated in the example given below.

The following table shows the heights of the students of a college :

Table 4.1

Class	*A*	*B*	*C*	*D*
	f	*f*	*f*	*f*
56.5-58.5	5	3	0	4
58.5-60.5	25	5	4	8
60.5-62.5	15	20	40	20
62.5-64.5	10	44	24	24
64.5-66.5	15	20	20	40
66.5-68.5	25	5	8	4
68.5-70.5	5	3	4	0
N	100	100	100	100
Mean	63.5	63.5	63.5	63.5
Median	63.5	63.5	63	64
Mode	—	63.5	61.9	65.1

The histograms (based on Table 4.1) and the corresponding curves are drawn in Figs. 4.3 and 4.4



A glance at the data of each of the four classes given above makes a very interesting study.

The shape of the curves, histograms and placement of equal items at equal distances on either side of the median clearly show that distributions A and B are symmetrical. If we fold these curves, or histograms on the ordinate at the mean, the two halves of the curve or histograms will coincide. In distribution B all the three measures of central tendency are identical. In A, which is a bimodal distribution mean and median have the same value.

Distributions C and D are asymmetrical. This is evident from the shape of the histograms and curves, and also from the fact that items at equal distances from the median are not equal in number. The three measures of central tendency for each of these distribution are of different sizes.

A point of difference between the asymmetry of distribution C and that of D should be carefully noted. In distribution C, where the mean (63.5) is greater than the Median (63) and the Mode (61.9) the curve is pulled more to the right. In distribution D where Mean (63.5) is lesser than the Median (64) and mode (65.1) the curve is pulled more to the left.

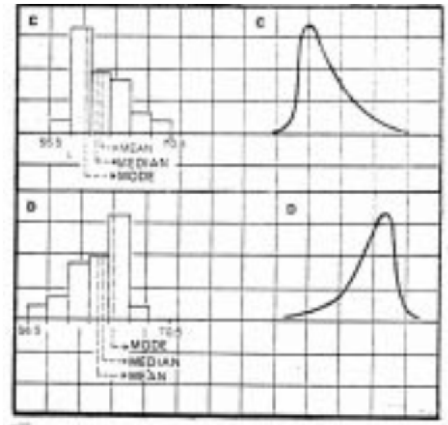


FIG. 6.4

2. Quartile measure of skewness - known as Bowley's measures of skewness.

3. Percentile measure of skewness also called the Kelly's measure of skewness.

4. Measures of skewness based on moments.

All these measures tell us both the direction and the extent of the skewness.

1. Karl Pearson's measure of skewness : It has been shown earlier that in a perfectly symmetrical distribution, the three measures of central tendency, viz, mean, median and mode will coincide. As the distribution departs from symmetry these three values are pulled apart, the difference between the mean and mode being the greatest. Karl Pearson has suggested the use of this difference in measuring skewness. Thus absolute Skewness = Mean - Mode (+) or (-) signs obtained by this formula woud exhibit the direction of the skewness. If it is positive the extreme variation in the given distribution are towards higher values. If it is negative, it shows that extreme variations are towards lower values.

Pearsonian coefficient of skewness : The difference between mean and mode, as explained in the preceding paragraph, is an absolute measure of skewness. An absolute measure cannot be used for making valud comparison between the skewness in two or more distributions for the following reasons: (i) The same size of skewness has different significance in distributions with small variation and in distributions with large variation, in the two series, and (ii) The unit of measurement in the two series may be different.

To make this measure as a suitable device for comparing skewnes, it is

necessary to eliminate from it the disturbing influence of 'variation' and 'units of measurements'. Such elimination is accomplished by dividing the difference between mean and mode by the standard deviation. The resultant ccoefficient is called Pearsonian coefficient of skewness. Thus the formula of Pearsonian coefficient of skewness is

$$\text{Coefficient of Skewness} = \frac{\text{Mean - Mode}}{\text{Standard deviation}} \qquad (4.7)$$

Since, as we have already seen, in moderately skewed distributions

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

We may remove the mode from the formula by substituting the above in the formula for skewness, as follows :

$$\text{Coefficient of Skewness} = \frac{\text{Mean - [Mean - 3(Mean-Median)]}}{\text{Standard deviation}} \qquad (4.7)$$

$$= \frac{\text{Mean - Mean + 3(Mean-Median)}}{\text{Standard deviation}}$$

$$= \frac{3\,(\text{Mean - Median})]}{\sigma} \qquad (4.8)$$

The removal of the mode and substituting median in its place becomes necessary because mode cannot always be easily located and is so much affected by grouping errors that it becomes unreliable.

Illustration 4.2 : Find the skewness from the following data :

Table 4.2

Height (in inches)	*Number of persons*
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Solution : Height is a continuous variable, and hence 58" must be treated as 57.5" - 58.5", 59" as 58.5" —59.5", and so on.

Height (in inches)	*f*	*x'* *(from 6I)*	*fx'*	*fx'²*	*Cumulative frequency*
58	10	-3	-30	90	10
59	18	-2	-36	72	28
59.5"-60-60.5"	30	-1	-30	30	58
60.5"-61-61.5"	42	0	-90	0	100
61.5"-62-62.5"	35	1	35	35	135
62.5" - 63'-63.5"	28	2	56	112	163
63.5"-64-64.5"	16	3	48	144	179
64.5"-65-65.5"	8	4	32	128	187
	187		171	609	
			+75		

Mean = $61 + \frac{75}{187} = 61.4$

Mode = $60.5 + \frac{35}{65} = 61.04$

$\sigma = \sqrt{\frac{609}{187} - (\frac{75}{187})^2}$

$= \sqrt{3.27-.16}$

$= \sqrt{3.11 = 1.76}$

Skewness = 61.4 - 61.04

= 0.36 inches

Coefficient of skewness = $\frac{.36}{1.76} = .205$

Alternatively, we can determine the median

Median = the size of $\frac{187}{2}$ th item

= 93.5th item

$= 60.5 + \frac{1x35.5}{42} = 61.35$ (4.7)

Skewness = 3(61.4-61.35)

= 3(.05)

= .15

Coefficient of Skewness = $\frac{.15}{1.76} = .09$

The two coefficients are differents because of the difficulties associated with determination of mode.

2. Bowley's (Quartile) measures of skewness : In the above two methods of measuring skewness, the whole of the series is taken into consideration. But absolute as well as relative skewness may be secured even for a part of the series. The usual device is to measure the distance between the lower and the upper quartiles. In a symmetrical series the quartiles would be equidistant from the value of the median, i.e.,

$$\text{Median} - Q_1 = Q_3 - \text{Median}.$$

In other words, the value of the median is the mean of Q_1 and Q_3. In a skewed distribution, quartiles would not be equidistant from median unless the entire asymmetry is located at the extremes of the series. Bowley has suggested the following formula for measuring skewness, based on above facts.

$$\text{Absolute SK} = (Q_3 - \text{Med}) - (\text{Med} - Q_1)$$

$$= Q_3 + Q_1 - 2\,\text{Med} \qquad (4.9)$$

If the quartiles are equidistant from the median, i.e. $(Q_3 - \text{Med}) = \text{Med} - Q_1$, then SK = 0. If the distance from the median to Q_1 exceeds that from Q_3 to the median, this will give a negative skewness. If the reverse is the case; it will give a positive skewness.

If the series expressed in different units are to be compared, it is essential to convert the absolute amount into the relative. Using the interquartile range as a denominator we have for the coefficient of skewness the following:

$$\text{Relative SK} = \frac{Q_3 + Q_1 - 2\,\text{Med}}{Q_3 - Q_1} \qquad (4.10)$$

$$= \frac{Q_3 - Med) - (Med - Q_1)}{Q_3 - Med) + (Med\ Q_1)}$$

If in the series the median and lower quartiles coincide, then the SK becomes (+1). If the median and upper quartiles coincide, then the SK becomes (-1).

This measure of skewness is rigidly defined and easily computable. Further, such a measure of skewness has the advantage that it has value limits between (+1) and (-1), with the result that it is sufficiently sensitive for many requirements. The only criticism levelled against such a measure is that it does not take into consideration all the item of these series, i.e., extreme items are neglected.

Illustration 4.3 : Calculate the coefficient of skewness of the data of Table 1.2 based on quartiles.

Solution : With reference to Table 4.2

$$Q_1 = \text{the size of } \frac{N}{4} \text{ th} \quad \left(= \frac{187}{4} = 46.75\text{th}\right) \text{ item}$$

$$= 59.5 + \frac{18.75}{30}$$

$$59.5 + .63 = 60.13$$

$$Q_3 = \text{the size of } \frac{3N}{4} \text{ th} \quad \left(= \frac{3x187}{4} = 140.25\text{th}\right) \text{item}$$

$$= 62.5 + \frac{5.25}{28}$$

$= 62.5 + .19 = 62.69.$

Skewness = 62.69+60.13 - 2(61.35) = .12 (Formula 4.9)

Coefficient of Skewness = $\frac{.12}{2.56}$ =.047 (4.10)

3. Kelly's (Percentile) measure of skewness : To remove the defect of Bowley's measure that it does not take into account all the values, it can be enlarged by taking two deciles (or percentiles), equidistant from the median value. Kelly has suggested the following measure of skewness :

$$SK = P_{50} - \frac{P_{90} - P_{10}}{2}$$

Or $$= D_5 - \frac{D_9 + D_1}{2}$$

Though such a measure has got little practical use, yet theoretically this measure seems very sound.

4. Measure of Skewness Based on Moments

It will be recalled that in symmetrical distribution, all the odd moments about mean, i.e. μ_3, μ_5, μ_7 ... etc., are equal to zero. If the odd moments (other than μ_1) are not equal to zero then it means that the distribution is skewed. But the computation of odd moments alone is not a satisfactory method of measuring skewness. To exhibit the degree of asymmetry we must relate these moments to the standard deviation. Thus, the various moments divided by the proper power of the standard deviation give us another family of useful coefficients

which we denote by the Greek letter α (alpha).

Symoblically,

$$\alpha_1 = \frac{\mu_1}{\sigma} = 0$$

$$\alpha_2 = \frac{\mu_2}{\sigma^2} = 1 \qquad (4.11)$$

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

$$\alpha_4 = \frac{\mu_4}{\sigma^4}$$

$$\alpha_5 = \frac{\mu_5}{\sigma^5} = 0$$

Since μ_1 is always zero for every distriution $\mu_1 = \frac{\sum x}{N} = 0$, , it is useless as a test of skewness. μ_3 is preferable to any higher moment as it is easier to calculate and also because the higher the moment the more will it vary from sample to sample. The positive and negative sign of α_3 will have the same significance as the sign of (mean-mode) has.

Another measure of skewness is obtained by the folloiwng formula :

$$= \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6b_1 - 9)} \qquad (4.12)$$

Where $\beta_1 = \alpha_3^2$ and $\beta_2 = \alpha_4$

This measure will be positive if the mean exceeds, and negative if the mean falls short of the mode.

Illustration 4.4 : Calculate the Karl Pearson's coefficient of skewness from the following data :

Marks	*No. of students*	*Marks*	*No. of students*
above 0	150	above 50	70
above 10	140	above 60	30
above 20	100	above 70	14
above30	80	above 80	0
above 40	80		

Solution

Marks	*f*	*Mid point**	*d' = (X-A)/10*	*fx'*	*fx'²*	*c.f.*
0-10	10	5	-3	-30	90	10
10-20	40	15	-2	-80	160	50
20-30	20	25	-1	-20	20	70
30-40	0	35	0	0	0	70
40-50	10	45	1	10	10	80
50-60	40	55	2	80	160	120
60-70	16	65	3	48	144	136
70-80	14	75	4	56	224	150
	150			+64	808	

Since it is a bi-modal distribute in Karl Pearson's coefficient (Formula 1.8) is appropriate and we need to calculate x, Med and σ.

Mean $x = 35 + \frac{64}{150} \times 10 = 35 + 4.27 = 39.27$

Median = size of 150/2th item

$= 40 + \frac{10 \times 5}{10} = 45$

$$\text{Standard deviation} = i \times \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx'^2}{N}\right)}$$

$$= 10 \times \sqrt{\frac{808}{150} - \left(\frac{64}{150}\right)}$$

$= 10 \times \sqrt{5.387 - .182}$

$= 10 \times 2.28 = 22.8$

$$\text{Skewness} = \frac{3(\text{x-Median})}{\sigma}$$

$$\text{Skewness} = \frac{3(39.27-45)}{22.8}$$

$$= \frac{3(-5.73)}{22.8} = \frac{-17.19}{22.8} = -.75$$

Illustration 4.5 : Find the standard deviation and coefficient of skewness for the following distribution.

Variable	0-	5-	10-	15-	20-	25-	30-	30-40
Frequency	2	5	7	13	21	16	8	3

Solution : As nothing is specified skewness should be computed by Karl Pearson's method.

Variable	*X*	*Frequency*	*(x-A)/c x'*	*fx'*	*fx'²*
0-5	2.5	2	-4	-8	32
5-10	7.5	5	-3	-15	45
10-15	12.5	7	-2	-14	28
15-20	17.5	13	-1	-13	13
20-25	22.5	21	0	0	0
25-30	27.5	16	1	16	16
30-35	32.5	8	2	16	32
35-40	37.5	3	3	9	27
		75		∑fx' = -9	193

Illustration 4.6 : From the following data compute quartile deviation and the coefficient of skewness :

Size	5-7	8-10	11-13	14-16	17-19
Frequency	14	24	38	20	4

Solution :

Size	*Frequency*	*Cumulative Frequency*
4.5-7.5	14	14
7.5-10.5	24	38
10.5-13.5	38	76
13.5-16.5	20	96
16.5-19.5	4	100

$$Q_1 = 7.5 + \frac{3x11}{24} = 8.87$$

$$Q_3 = 10.5 + \frac{3x37}{38} = 10.5 + \frac{111}{38} = 10.5 + 2.92 = 13.42$$

$$\text{Median} = 10.5 + \frac{3x12}{38} = 10.5 + \frac{36}{38} = 10.5 + .947 = 11.447$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{13.42 - 8.87}{2} = \frac{4.55}{2} = 2.275$$

$$\text{Skewness} = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

$$= \frac{13.42 + 8.87 - 22.89}{13.42 - 8.87} = \frac{-.6}{4.55} = -0.13$$

Illustration 4.7 : In a certain distribution the following results were obtained :

$x = 45.00$; Median = 48.00

Coefficient of Skewness = -.4

You are required to estimate the value of standard deviation

Solution :

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$-.4 = \frac{3(45 - 48)}{\sigma}$$

$$-.4\sigma = -9$$

$$\sigma = \frac{9}{.4} = 22.5$$

Illustration 4.8 : You are given the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of the workers and that of the management.

	Before	***After***
No. of workers	2,440	2,359
Mean wages	45.5	47.5
Median wages	49.0	45.0
Standard deviation	12.0	10.0

Solution :

Employment : Since the number of workers employed after the settlment is less than the number of employed before, it has gone against the interest of the workers.

Wages : The total wages paid after the settlement were 2,350 x 47.5 = Rs. 1,11,625; before the settlement the amount disbursed was 2,400 x 45.5 = Rs. 1,09,200.

This means that the workers as a group are better off now than before the settlement, and unless the productivity of workers have gone up this may be

against the interest of management.

Uniformity in the wage structure : The extent of relative uniformity in the wage structure before and after the settlement can be determined by a comparison of the coefficient of variation.

$$\text{Coefficient of variation before} = \frac{12}{45.5} \times 100 = 26.4$$

$$\text{Coefficient of variation after} = \frac{10}{47.5} \times 100 = 21.05$$

This clearly means that there is comparatively lesser disparity in the wages received by the workers. Such a position is good for both the workers and the management.

Pattern of the wage structure : A comparison of the mean with the median leads to the obvious conclusion that before the settlement more than 50 per cent of the workers were getting a wage higher than this mean i.e. (Rs. 45.5). After the settlement the number of workers whose wages were more than Rs. 45.5 became less than 50 per cent. This means that the settlement has not been beneficial to all the workers. It is only 50 pr cent workers who have been benefitted as a result of an increase in the total wages bill.

4.6 Kurtosis

So far we have characterised a frequency distribution by its central tendency, variability, and the extent of asymmetry. There remains one more common type of attribute of frequency distribution viz., its peakedness. Look at the following,

Fig. (1.5) in which are drawn three symmetrical curves A, B and C.

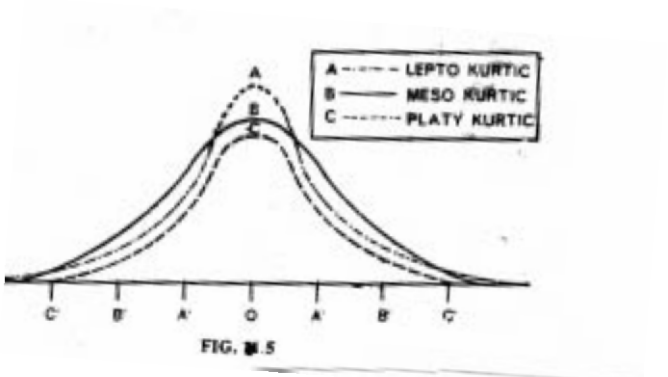


FIG. 4.5

The three curves differ widely in regard to convexity, an attribute to which Karl Pearson referred as 'Kurtosis'. The measure of kurtosis exhibits the extent to which the curve is more peaked or more flat topped than the normal curve. A curve to be called a normal curve must have its convexity as shown in curve B (in addition to two other requisites, viz., (i) Unimodal, (ii) Symmetrical]. When the curve of a distribution is relatively flatter than the normal curve, it is said to have kurtosi. When the curve or polygon is relatively more peaked, it is said to lack kurtosis,

Karl Pearson gave the name "Mesokurtic' to a normal curve or a skewed curve that has the same degree of convexity as the normal curve. In Fig. 4.5 curve B is a mesokurtic curve. If some of the cases about one standard deviation from the mean move in towards the centre and other move out towards the tails, thus making the curve unusually peaked, we say that the result would be a 'Leptokurtic' curve, as curve A. If on the other hand, some of the cases around the mode move out a little towards each half of the curve, thus making the

curve unusually flat topped we say that the result would be a 'Platykurtic' curve, as curve C.

To quote Walker, "The terms platykurtic, leptokurtic, mesokurtic, are not particularly important but they are rather interestng and roll pleasantly under one's tongue." An amusing sentence written by 'Student' (a British Statistician) was quoted by him which reads "the platykurtic curves, like the platypus are squat with short tails, while leptokurtic curves are high with long tails like the Kangaroo which 'leps'."

The kurtosis is measured by

$$b_2 = \alpha_4 = \frac{\mu_4}{\mu_2^2} \quad \text{or} \quad \frac{\mu_4}{\sigma^4}$$

In a normal curve, b_2 will be equal to three. If b_2 is greater than three, the curve is more peaked, if less than three, the curve is flatter at the top than normal.

The above formula may be rewritten as :

$$K = \alpha_4 - 3 = \frac{\mu_4}{\sigma^4} - 3$$

If K is positive, it means that the number of cases near the mean is greater than in normal distribution. If K is negative, the curve is more flat-topped than the corresponding normal curve.

The measures of skewness and kurtosis are also expressed in terms of the Greek letter, gamma (γ).

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \alpha_3$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_3}{\sigma^4} - 3 = \frac{\mu_3}{\mu_2^2} - 3 = \frac{\beta^2 - 3\mu_2^2}{\mu_2^2}$$

γ_1 and γ_2 are the measures of skewness and kurtosis respectively. If γ_1 is more than zero, then the conclusion will be the presence of positive skewness; if γ_1 is less than zero, it will mean negative skew-ness, and in case γ_1 is zero, then there will be absence of skewness.

Similarly, if a curve is Leptokurtic, γ_2 will be positive; if Platykurtic, γ_2 will be negative; and in case Mesokurtic, γ_2 will be exactly zero.

4.7 Sheppard's corretion for grouping errors : Computation of mean, standard deviation, etc., from the grouped data of a frequency distribution are based upon the assumption that all the values in a class interval are concentrated at the centre of that interval. In unimodal (or single peaked) symmetrical distributions this assumption results in a systematic error in the calculation of the even moments. For the odd moments. The sum total of the errors on account of the above presumption is zero in such symmetrical distributions, because of the neutralzing effect of the errors as these appear with positive and negative signs.

Such errors may with advanatage be corrected by the application of the formula propounded by W.F. Sheppard. The correction formulae for 2nd and 4th moments are :

$$\mu_2 \text{ (Corrected)} = \text{Uncorrected } \mu_2 - \frac{i^2}{12}$$

$$\mu_4 \text{ (Corrected)} = \text{Uncorrected } \mu_4 - \tfrac{1}{2} i^2 \mu_2 \text{ (Uncorrected)} + \frac{7}{240} i^4$$

(Where i is the class interval).

The use of this correction be restricted to :

(i) grouped data, i.e., series must not be discrete but continuous;

(ii) symmetrical and moderately skewed distribution, i.e., the distribution tapers off to zero in both directions;

(iii) total frequency is sufficiently large, say 1,000; and

(iv) class interval is more than $\frac{1}{20}$ th of the range.

Illustration 4.9 : From the following frequency distribution calculate the first four moments, β_1, and β_2.

Class	*f*
10-14	1
15-19	4
20-24	8
25-29	19
30-34	35
35-39	20
40-44	7
45-49	1
50-54	5

Solution :

Let the assumed mean A=32

Class	*Mid-point*	*f*	*x'*	*fx'*	fx'^2	fx'^3	fx'^4	*c.f*
10-14	12	1	-4	-4	16	-64	256	1
15-19	17	4	-3	-12	36	-108	324	5
20-24	22	8	-2	-16	32	-64	128	13
25-29	27	19	-1	-19	19	-19	19	32
30-34	32	35	0	-51	0	-255	0	67
35-39	37	20	1	20	20	20	20	87
40-44	42	7	2	14	28	56	112	94
45-49	47	5	3	15	45	135	405	99
50-54	52	1	4	4	16	64	256	100
		100		$\sum fx'$ = +2	$\sum fx'^2$ = +212	$\sum fx'^3$ = +20	$\sum fx'^4$ = +1,520	

$$\nu_1 = \frac{\sum fx' \times (c.i)}{\sum f} = \frac{2x5}{100} = \frac{10}{100} = .1$$

$$\nu_2 = \frac{\sum fx'^2 \times (c.i)^2}{\sum f} = \frac{212x5^2}{100} = \frac{212x25}{100} = 53$$

$$\nu_3 = \frac{\sum fx'^3 \times (c.i)^3}{\sum f} = \frac{20x5^3}{100} = \frac{20x25}{100} = 25$$

$$\nu_4 = \frac{\sum fx'^4 \times (c.i)^4}{\sum f} = \frac{1,520x5^4}{100} = \frac{1,520x625}{100} = 9,500.$$

$$\mu_1 = \nu_1 - \nu_2 = 0$$

$\mu_2 = \nu_2 - \nu_1^2 = 53 - (.1)^2$

$= 53 - .01 = 52.99$

$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3.$

$= 25 - 3x53x(.1) + 2(.1)^3 = 25 - 15.9 + .002 = 9.102.$

$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$

$= 9,500 - 4x25 \times (.1) + 6x53 \times (.1)^2 - 3(.1)^4$

$= 9,500 - 10 + 3.18 - .0003$

$9,493.18.$

Corrected $\mu_2 = \mu_2 - \frac{h^2}{12}$ (where h is the class interval)

$= 52.99 - \frac{5^2}{12} = 52.99 - 2.083 = 50.91.$

Corrected $\mu_3 = 9.1.$

Corrected $\mu_4 = \mu_4 - \frac{h^2\mu_2}{2} + \frac{7h^4}{240}$

$= 9,493.18 - \frac{25}{2} \times 52.99 \quad \frac{7 \times (5)^4}{240}$

$= 9,493.18 - 662.375 + 18.23 = 8,849.03.$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

$$= \frac{(9.1)^2}{(50.91)^3} = 0.000627$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{8,849.03}{(50.91)^2} \qquad = 3.414$$

$$\gamma_1 = \sqrt{\beta_1}$$

$$= \sqrt{.000627} = .025.$$

$$\gamma_2 = \beta_2 - 3$$

$$= 3.414 - 3 = .414.$$

Illustration 4.10 : The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and Kurtosis of the distribution.

Solution : Skewness is tested by μ_3 which should be equal to zero in a symmetrical distribution. Since in the given problem it is 0.7 we can conclude that the distribution is not symmetrical. But to measure the extent and direction of the skewness we make use of constant.

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

$$\alpha_3 = \frac{\mu_3^2}{\mu_2^3}$$

$$\frac{(0.7)^2}{(2.5)^3}$$

$= \sqrt{0.031} = 0.18$

Since $\alpha_3 = 0.18$ we conclude that the distribution is not symmetrical but has a positive skewness = 0.18.

Kurtosis is tested by β_2. In normal case β_2 should be equal to three. If it is greater than three the curve is more peaked, if less than three the curve is more flat-topped.

$$\beta_2 = \frac{\mu_4}{s^4} = \frac{\mu_4}{\mu_2^2}$$

$$\frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

Since $\beta_1 = 3$ we conclude that the curve is mesokurtic.

Self test Questions

Q 1 : Calculate coefficient of skewness from the data as follows :

Life time (Hours)	*No. of Tubes*
300-400	14
400-500	46
500-600	58
600-700	76
700-800	68
800-900	62
900-1000	48
1000-1100	22
1100-1200	6

Q 2 : Karl Pearson's coefficient of skewness of a distribution is +0.32. Its standard deviation is 6.5 and mean is 29.6. Find the mode and median of the distribution.

Q 3. : For a distribution Bowley's coefficient of skewness is -0.36. $Q_1 = 8.6$ and Median = 12.3. What is the quartile coefficient of dispersion ?

Q 4 : (a) The standard deviation of a symmetricl distribution is 3. What must be the value of the 4th moment about the mean in order that the distribution be mesokurtic ?

(b) If the 4 moments of a distribution about the value 5 are equal to -4, 22, -117 and 560 determine the corresponding moments about the mean and about zero.

❄ ❄ ❄

Lesson : 5

CORRELATION ANALYSIS

Author :
Dr. B. S. Bodla

Vetter:
Prof. Chander Shakher

INTRODUCTION

The Statistical techniques like measure of central tendency, dispersion, skewness and kurtosis concern with the analysis of univariate data i.e., the distributions involving only one variable. In practice we come across a large number of problems involving the use of two or more than two variables. In the present lesson we shall discuss method of analysing data on two variables which may be related or associated in some way.

In fact, there may exist a relationship between two or more decisions. For example, the knowledge of the relationship between expenses on advertisement of product and its sales and the relationship between price of a commodity and its supply is worthwhile for the manager of a business concern. Also, a knowledge of the nature and degree of such relationships can be used for predicting with some measure of reliability the value of a variable, which we call the dependent variable, corresponding to some known value (s) of one or two variables, which we call the independent variable (s). The degree of relationship between the variables under consideration is measureed through the correlation analysis. The measures of correlation called the correlation

coefficient summarizes in one figure the direction and degree of correlation. The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

Some Definitions of Correlation

1. "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation".

   Croxton & Cowden

2. "Correlation analysis attempts to determine the "degree of relationship between variables". -Ya Lun Chou

3. "Correlation is an analysis of the covariation between two or more variables". -A. M. Tuttle

4. "Correlation analysis deals with the association between two or more variables". -Simpson and Kafka

Thus, two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable. The problem of analysing the relation between different series can be divided into three steps:

1. Deciding whether a relationship between two or more variables exists and, if it does, measuring the degree of that relationship.

2. Testing whether the arrived value of correlation is significant.

3. Establishing the cause and effect relation, if any.

We shall discuss only the first step in this lesson. It should be noted that detection and analysis of correlation (i.e., covariation) between two statistical variables requires relationship of some sort which associates the observations in pairs, one of each pair being a value of each of the two variables.

It is noteworthy that the computations regarding the degree of closeness is ased on the regression analysis. But the correlation can be arrived at without actually having a regression equation.

Types of Correlation

The relationship found among two or more variables may take any form of the following :

(a) Positive and negative correlation.

(b) Simple, partial and multiple correlation.

(c) Linear and non-linear correlation.

a) Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e., if the increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, correlation is said to be positive. If, on the other hand, the variables are varying in opposite direction, i.e., as one variable is increasing the other is decreasing or vice versa, relationship is said to be negative. The following examples would make a difference between positive and negative correlation :

Positive Correlation :

i)	Family income (X)	:	25	28	30	34	36
	Family expenditure (Y)	:	30	32	36	40	45
ii)	Price (X)	:	70	60	40	20	15
	Supply (Y)	:	110	80	60	20	10

Negative Correlation :

X	:	25	28	30	34	36
Y	:	45	40	35	32	30

b) Simple, Partial and Multiple Correlation

We can make a distinction between simple, partial and multiple correlation on the basis of the number of variables considered for the purpose of computation of relationship between them. In case of simple correlation the relationship between only two variables is examined;whereas when the relationship between three or more variables is studied it is a problem of either partial or multiple correlation. In partial correlation we recognise more than two variables, but consider only two variables, to be influencing each other, when linear effect of other influencing variables on them has been elliminated. On the other hand, in multiple correlation three or more variables are studied simultaneously. For example, when we examine the relationship between the sales of a product and both the amount of advertisement expenditure and the amount of capital invested, it is a problem of multiple correlation. On the other hand, in the sales problem taken above if we limit our analysis of sales and

advertisement to periods when a certain level of capital investment existed, it becomes a problem of partial correlation.

c) Linear and Non-linear Correlation

The relationship between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values. For example, let us consider the following data :

X	:	2	3	4	5	6
Y	:	7	9	11	13	15

Thus for a unit change in the value of X, there is a constant change in the corresoponding values of Y. In mathematical terms, above data can be expressed by the relation.

$$Y=2X+3$$

In general, two variables X and Y are said to be linearly related, if there exists a relationship of the form.

$$Y=a+bX$$

between them. But "Y=a+bX" is the equation of a straight line with slope 'b' and which makes an intercept 'a' on the Y-axis. Therefore, if we plot the values of the two variables as points in the XY-plane, we shall obtain a straight line. Suchj phenomena occur most frequently in physical sciences but in business, economics and social sciences we rarely come across the data which give a straight line graph. The relationship between two variables is said to be Non-

linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but at fluctuating rate. With such data if we prepare a graph we do not get a straight line curve.

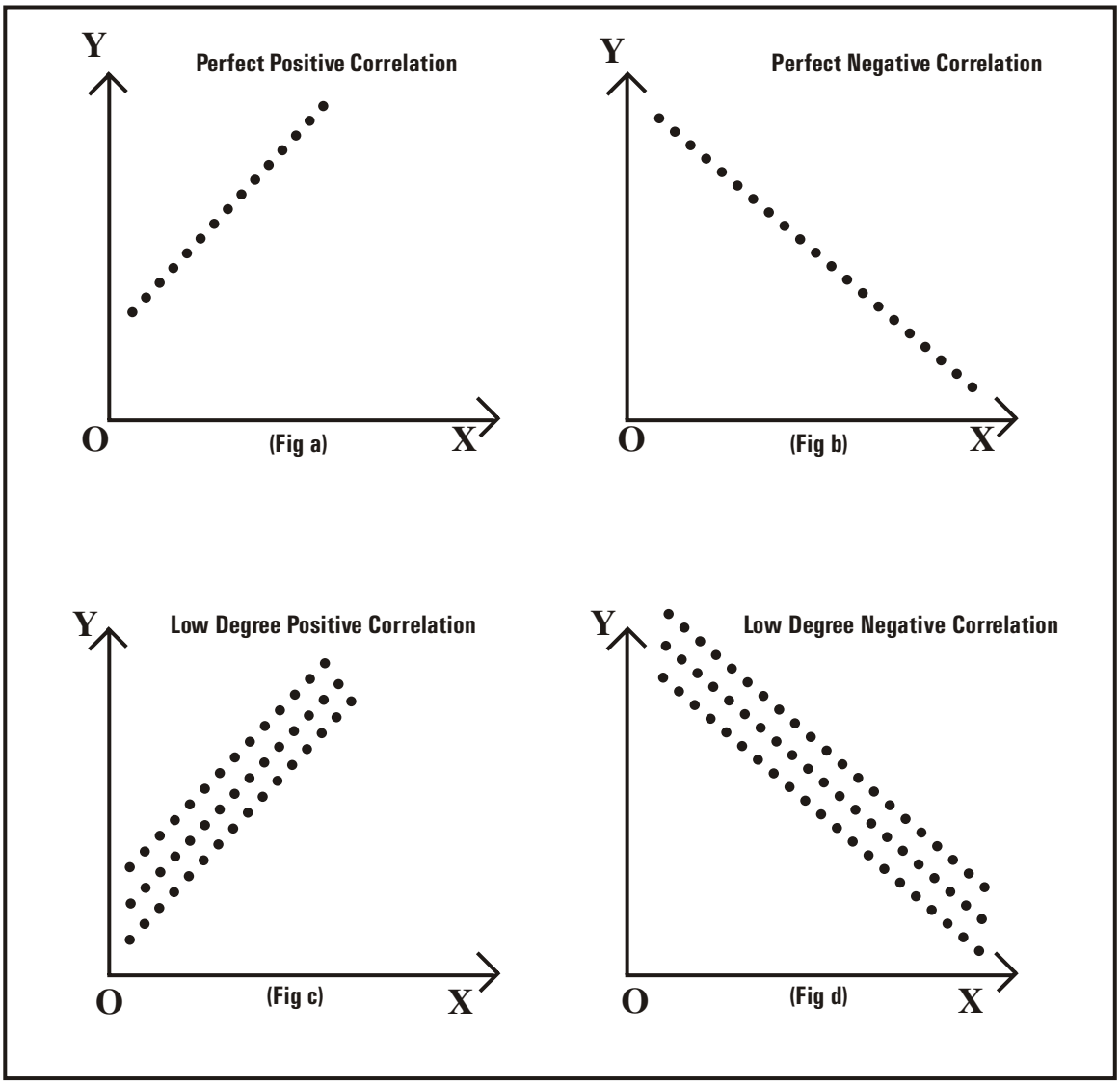
The study of non-linear relationship is quite complicated. Further, the predictive capability of non-linear relationship is not so good, and, hence, our discussion is restricted to linear relationships only. Moreover since linear relationship implies a constant change in the dependent variable with respect to changes in the independent variable (s) it is very useful for predictive purposes.

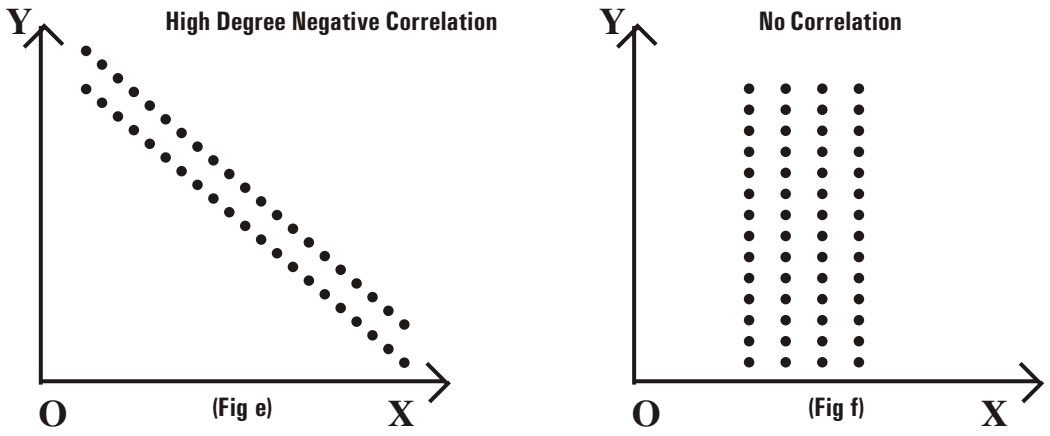
Scatter Diagrams and Fitting a Curve Thereon

The first step in determining whether there is a relationship between two variables is to examine the graph of the observed data. This graph, or chart, is called a scatter diagram. A scatter diagram can give us two types of information. Visually, we can look for patterns that indicate that the variables are related. Then, if the variables are related, we can see what kind of line, or estimating equation, describe this relationship. For developing a scatter diagram, firstly, the given data are plotted on a graph paper in the form of dots, i.e., for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. If all the points lies on a straight line falling from the lower left hand corner to the upper right hand corner, correlation is said to be perfectly positive. If the plotted points fall in a narrow hand there would be a

high degree of correlation between the variables. On the other hand, if the points are widely scattered over the diagram it indicates very little relationship between the variables. If the plotted points lie on a straight line parallel to the X-axis or in haphazard manner, it shows absence of any relationship between the variables.

The following scattered diagrams depict different forms of correlation.



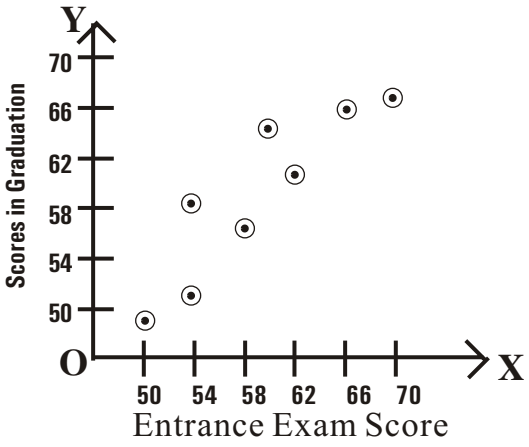


Example 1 : Following are the students scores in BBA entrance examination and their score in graduation :

Student	A	B	C	D	E	F	G	H
Entrance examination scores	51	55	59	62	65	67	50	62
score in Graduation (Y)	52	58	57	63	66	66	52	58

(Note : 100 is Maximum possible score)

Solution : The scatter diagram of the above data is shown below.



Since the points are dense i.e., close to each other, we may expect a high degree of correlation between the series of entrance score and graduation score. As the slope of points is upward from left to right, the correlation is positive.

Fitting A Straight Line on the Scatter Diagram

The method of scatter diagram is readily comprehensible and enables us to form a rough idea of the nature of the relationship between the two variables merely by inspection of the graph. However, the method of scatter diagram only tells us about the nature of the relationship whether it is positive or negative and whether it is high or low.

The scatter diagram enables us to obtain a straight line or an approximate estimating line. The straight line has to be so fitted that in general lies as close as possible to every point on the scatter diagram. This condition requires that the sum of the squares of the vertical deviations of the observed Y values from the straight line should be minimum. Since a straight line so fitted would best approximate all the points on the scatter diagram, it is better known as the best approximating line or the line of best fit. Such a straight line can be fitted by two ways : freehand drawing and the least square method. We shall discuss only freehand drawing method in this lesson. The least square method will be discussed in some other lesson.

Freehand Drawing

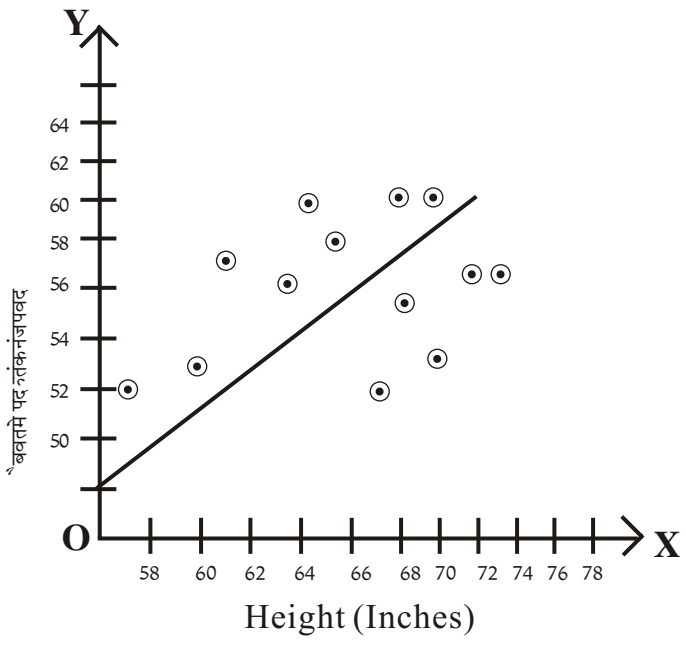
When we view all the points on scatter diagram together, we can visualize the relationship that exists betwen the two variables. As a result, we can draw, or 'fit', straight line through our scatter diaram to represent the relationship. Under

the free hand drawing method, a straight line is drawn through the spread of various points on the scatter diagram by using a transparent ruler such that on the whole it is closest to every point. This method of fitting straight line is particularly useful when approximate prediction estimates of the dependent variable are promptly needed. Although it may turn out to be the nearest to the line of best fit, the major drawback of this method is that the slope of the straight line is inflenced by the element of subjectivity and may vary from person to person. Therefore, the values of the dependent variable estimated on the basis of such a line may not be completely accurate and precise.

Example 2 : Following are the heights and weights of 10 students of BBA class.

Height (in inches)	X :	62	72	68	58	65	70	66	63	60	72
Weight (Kgs)	Y :	50	64	63	50	54	60	61	55	54	64

Draw a scatter diagram and fit a straight line on it by free hand drawing.



Correlation and Causation :

Correlation analysis is used for highlighting the nature of relationship between any two variables. However, it does not tell us anything about cause and effect relationship. Even a greater degree of correlation does not necessarily imply establishing any causal relationship between the two variables. But the existence of causation always implies correlation. The high degree of correlation between the variables may be due to the following reasons :

1. Both the Variables Being Influenced by the Same External Factors

The relationship between the two variables may be due to the effect or inter action of a third variable or a number of variables on each of these two variables. For example, a fairly high degree of correlation may be found between the yield per hectare of two crops, say rice and potato, due to the effect of a number of factors like level of rain fall, fertilizers used, favourable weather conditions, etc. on each of them. But none of the two is the cause of the other.

.2. Mutual Dependence

A high degree of correlation between the two variables may be in existence when both of the variables are mutually influencing each other so that neither can be designated as the cause and the other the effect. Such situations are usually observed in data relating to economic and business situations. For example, such variables as demand and supply, price and production, etc. mutually interact.

3. The correlation may be Due to Pure Chance, Especially in a Small Sample

It may happen that a small randomly selected sample from a bivariate distribution may show a fairly high degree of correlation though, actually, the variables

may not be correlated in the population. Such correlation may be attributed to chance fluctuations. For example, we may observe a high degree of correlation between the size of shoe and the intelligence of a group of persons. Such correlation is called spurious or non-sense correlation.

Karl Pearson's Coefficient of Correlation (Product Moment Formula)

A mathematical method for measuring the intensity of the magnitude of linear relationship between two variable X and Y, was suggested by Karl Pearson (1867-1936). The Pearson Coefficient of Correlation is denoted by the symbol r and is defined as the ratio of the covariance between X and Y. According to Product Moment Formula,

$$\mathbf{r} = \frac{\text{Cov. } (X,Y)}{\sigma X \sigma Y} \qquad \text{.....(5.1)}$$

If (X_1, Y_1), $(X_2, Y_2$ ....., (X_n, Y_n) are N pairs of observation of the variables X and Y in a distribution, then

$$\text{Cov. } (X,Y)\ 1/N\ \Sigma\ (X-\overline{X})\ (Y-\overline{Y})$$
$$\sigma x = \sqrt{1/N\ \Sigma\ (X-\overline{X})^2} \qquad \text{.....(5.2)}$$
$$\sigma y = \sqrt{1/N\ \Sigma\ (Y-\overline{Y})^2}$$

Substituting in (5.1) we get

$$r = \frac{1/N \Sigma (X-\overline{X})\ (Y-\overline{Y})}{\sqrt{1/N\ \Sigma\ (X-\overline{X})^2\ 1/N\ \Sigma\ (Y-\overline{Y})^2}} \qquad \text{.....(5.3)}$$

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2 \; \Sigma(Y-\overline{Y})^2}}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \; \Sigma y^2}} \qquad \text{.....(5.4)}$$

Here, $x = X-\overline{X}$ and $y = Y-\overline{Y}$

Formula (5.4) is quite convenient to apply if the means $\overline{X}$ and $\overline{Y}$ come out to be integers. If X and Y is (are) fractional then the formula (5.4) or (5.3) is quite cumbersome to apply since the computations are quite time consuming and tedious.

Example 3 : Calculate Pearson coefficient of correlation for the following series.

Price (Rs.) :	22	24	26	28	30	32	34	36	38	40
Demand (Tonnes) :	60	58	58	50	48	48	48	42	36	32

Solution : Let price be denoted by X and demand by Y.

Calculation of Pearson Correlation Coefficient

Y	$(X-\overline{X})$		*Y*	$(Y-\overline{Y})$		
	x	x^2		y	y^2	xy
22	-9	81	60	+12	144	-108
24	-7	49	58	+10	100	-70
26	-5	25	58	+10	100	-50
28	-3	9	50	+2	4	-6
30	-1	1	48	0	0	0

32	+1	1	48	0	0	0
34	+3	9	48	0	0	0
36	+5	25	42	-6	36	-30
38	+7	49	36	-12	144	-84
40	+9	81	32	-16	256	-144
$\Sigma X=310$	$\Sigma x=0$	$\Sigma x^2=330$	$\Sigma Y=480$	$\Sigma y=0$	$\Sigma y^2=784$	$\Sigma xy=-492$

$$r = \frac{\Sigma xy \quad .}{\sqrt{\Sigma x^2 \text{ x } \Sigma y^2}}$$

$x = (X - \overline{X}), y = (Y-\overline{Y})$

$$X = \frac{\Sigma x}{N} = \frac{310}{10} = 31, Y = \frac{\Sigma Y}{N} = \frac{480}{10} = 48$$

$\Sigma xy = -492, \Sigma x^2 = 330, \Sigma y^2 = 784$

$$X = \frac{-492}{\sqrt{330 \text{ x } 784}} - = -0.9674$$

Direct Method of Finding out Correlation

Correlation coefficient can also be calculated without taking deviation of items either from actual mean or assumed mean, i.e., actual X and Y values. The formula in such a case is

$$r = \frac{N\Sigma XY(\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2 \, N\Sigma Y^2 - (\Sigma Y)^2}} \quad \text{.....(5.5)}$$

This formulat would give the same answer as we get when deviations of items are taken from actual mean or assumed mean. The following example shall illustrate the point.

Example 4 : Calculate correlation coefficient from the data of illustration 3 by the direct method i.e., without taking the deviations of items from actual or assumed mean.

Solutions : Calculation of correlation coefficient by the direct method

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2 \quad N\Sigma Y^2 - (\Sigma Y)^2}} \qquad \text{.....(5.5)}$$

Calculation of correlation coefficient (Direct Method)

X	X^2	Y	Y^2	XY
9	81	15	225	135
8	64	16	256	128
7	49	14	196	98
6	36	13	169	78
5	25	11	121	55
4	16	12	144	48
3	9	10	100	30
2	4	8	64	16
1	1	9	81	9
$\Sigma X=45$	$\Sigma X^2=285$	$\Sigma Y=108$	$\Sigma Y^2=1,356$	$\Sigma XY=597$

$N=9, \Sigma XY=597, \Sigma X=45, \Sigma Y=108, \Sigma X^2=285, \Sigma Y^2 = 1,356$

$$r = \frac{9 \times 597 - 45 \times 108}{\sqrt{9 \times 285 - (45)^2}\ \sqrt{9 \times 1356 - (108)^2}}.$$

$$r = \frac{5373-4860}{\sqrt{2565 - 2025}\ \sqrt{12204 - 11664}}.$$

$$r = \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540}. = +0.95$$

When actual means of series X and Y are in fractions, the calculation of correlation by the method discussed above would involve too many calculations. In such cases we make use of the assumed mean method for calculating correlation coefficient. Here the following formula is applicable :

$$r = \frac{\Sigma d_x \Sigma d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}} \quad \text{.....(5.6)}$$

Where d_x = X-A, dy = Y-A, A = Assumed mean, N = No. of pairs of variables X and Y.

Example 5 : From the data given below find coefficient of correlation. Assume 8 and 9 as the mean values for X and Y respectively.

X : 6 2 10 4 8

Y : 9 11 5 8 7

Solution : Calculation of Correlation Coeffiecient

X	***Y***	***(X-A)*** ***X-8=d_x***	***(Y-A)*** ***y-9=d_y***	***d_x^2***	***d_y^2***	***$d_x d_y$***
6	9	-2	0	4	0	0
2	11	-6	2	36	4	-12
10	5	2	-4	4	16	-8
4	8	-4	-1	16	1	4
8	7	0	-2	0	4	0
		Σdx =-10	Σd_y =-5	Σd_x^2 =60	Σd_y^2 =25	$\Sigma d_x d_Y$ =-16

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\left[\Sigma d_x^2 - \frac{(\Sigma d_y)^2}{N}\right]\left[\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}\right]}}$$

$$r = \frac{-16 - \frac{(-10) \times (-5)}{5}}{\sqrt{\left[60 - \frac{(-10)^2}{5}\right]\left[25 - \frac{(-5)^2}{5}\right]}}$$

$$= \frac{-16 - 10}{\sqrt{40 \times 20}}$$

$$= \frac{-26}{28.284} = -0.92 \text{ Answer}$$

Correlation of Grouped Data

If in a bivariate distribution the data are fairly large, they may be summarised in the form of a two-way table. The class interval of X are listed in the stubs of the left of the table and those for Y are listed in the captions or column headings. This order can also be reversed.

The formula for computing the correlation coefficient between X and Y for the bivariate frequency tables or grouped data is

$$r = \frac{\Sigma fd_x d_y - \frac{(\Sigma fd_x)(\Sigma fd_y)}{N}}{\sqrt{\left[\Sigma fd_x^2 - \frac{(\Sigma fd_x)^2}{N}\right]\left[\Sigma fd_y^2 - \frac{(\Sigma fd_y)^2}{N}\right]}} \quad \text{.....(5.7)}$$

Steps. (i) Take the step deviations of variable X and denote these deviations by d_x. (ii) Take the step deviations of the variable Y and denote these deviations by d_y. (iii) Multiply dxdy and the respective frequency of each cell and write the figure obtained in the right-hand upper corner of each cell. (iv) Add together all the cornered values as calculated in step (iii) and obtain the total $\Sigma fd_x d_y$. (v) Multiply the frequencies of the variable X by the deviations of X and obtain the total Σfd_x. (vi) Take the squares of the deviations of the variable X and multiply them by the respective frequencies and obtain Σfd_x^2. (vii) Multiply the frequncies of the variable Y by the deviations of Y and obtain the total Σfd_y. (viii) Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and obtain Σfd_y^2. (ix) Substitute the values of $\Sigma fd_x d_y$, Σfd_x, Σfd_x^2, Σfd_y and Σfd_y^2 in the above formula and obtain the value of r.

Example 6 : The following table gives the frequency, according to group relationship between age and intelligence test.

Test Marks	***Age in years***			
	18	***19***	***20***	***21***
200-250	4	4	2	-7
250-300	3	5	4	2

300-350	Σ	2	6	8	5
350-400		1	4	6	10

Solution : Calculation of Coefficient of Correlation

X / *Y*	*Age in Years*								
	dy dx	*18 -1*	*19 0*	*20 +1*	*21 +2*	*f*	*fdy*	*fdy2*	*fdxd*
200-250	-1	4 4	0 4	-2 2	-2 1	11	-11	11	0
250-300	0	0 3	0 5	0 4	0 2	14	0	0	0
300-350	+1	-2 2	0 6	8 8	10 5	21	21	21	10
350-400	+2	2 1	0 4	12 5	40 10	21	42	84	5
	Total	10	19	20	18	67 =12	fd_x =116	Σfd_y^2 =96	$\Sigma fd_x d_y$ =66
	fd_x	10	0	20	36	Σfd_x =46			
	fd_x^2	10	0	20	72	Σfd_x^2 =102			
	$fd_x d_y$	0	0	18	48	$\Sigma fd_y d_x$ =66			

$$r = \frac{\Sigma fd_x d_y - \frac{(\Sigma fd_x)(\Sigma fd_y)}{N}}{\sqrt{\left[\Sigma fd_x^2 - \frac{(\Sigma fd_x)^2}{N}\right]\left[\Sigma fd_y^2 - \frac{(\Sigma fd_y)^2}{N}\right]}}$$

$\Sigma fd_x d_y = 66$, $\Sigma fd_x = 46$, $\Sigma fd_y = 52$, $\Sigma fd_x^2 = 102$, $\Sigma fd_y^2 = 116$, N=67

$$r = \frac{66 - \frac{46 \times 52}{67}}{\sqrt{\left[102 - \frac{(46)^2}{67}\right]\left[116 - \frac{(52)^2}{67}\right]}}$$

$$= \frac{66-35.7}{\sqrt{[102-31.6]\,[116-40.4]}} = \frac{30.3}{\sqrt{(70.4)\,(75.6)}}$$

Coefficient of Determination

Coefficient of corelation between two variable series is a measure of linear relationship between them and indicates the amount of variation of one variable which is associated with or is accounted for by another variable. One very convenient and useful way of interpreting the value of coefficient of correlation is to use the square of coefficient of correlation, which is called coefficient of determination. It gives the percentage variation in the dependent variable that is accounted for by the independent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient of determination is given by the square of the correlation coefficient i.e., r^2. Thus coefficient of determination is defined as follows :

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}. \qquad \text{.....(5.8)}$$

The ratio of unexplained variabce to total variance is frequently called the coefficient of non-determination. It is denoted by k^2 and its square root is called

the coefficient of allienation.

The coefficient of determination is a much useful and better measure for interpreting the value of r. Tuttle has beautifully pointed out that "the coefficient of correlation has been grossly overrated and is used entirely too much. Its square, the coefficient of determination is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlated variables".

For example, if the value of r=0.8, it cannot be concluded that 80% of the variation in the relative series is due to the variation in the subject series (independent variable). But the coefficient of determination in this case is r^2=0.64 which implies that 64% of the variation in the related series has been explained by the subject series and the remaining 36% of the variation is due to other factors.

Rank Correlation Method

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's Coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904 which consists

in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under study.

Suppose we want to find if two characteristics A, say, intelligence and B, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of n individuals in order of merit (ranks) w.r.t. proficiency in the two characteristics. Let the random variables X and Y denote the ranks of the individiuals in the characteristics A and B respectively. If we assume that there is no tie, i.e., if no two individuals get the same rank in a characteristic then, obviously, X and Y as some numerical values ranging from 1 to n.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for that group of individuals.

Spearman's rank correlation coefficient, usually denoted by p (Rho) is given by the formula.

$$p = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad \text{.....(5.9)}$$

where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Computation of Rank Correlation Coefficient

We shall discuss below the method of computing the Spearman's rank correlation coefficient P under the following situations :

(i) When actual ranks are given.

(ii) When rank are not give.

Case (i) When Actual Ranks are given :

In this situation the following steps are involved :

(i) Compute d, the difference of ranks.

(ii) Compute d^2

(iii) Obtain the sum Σd^2

(iv) Use formula (5.9) to get the value of P.

Example 7 : The rank of the same 15 students in two subjects A and B are given below; the two numbers within the brackets denoting the ranks of the same student in A and B respectively. (1,10) (2,7) (3,2) (4,6) (5,4) (6,8) (7,3) (8,1) (9,11) (10,15) (11,9) (12,5) (13,14) (14,12) (15,13).

Solution : Calculation of Spearman's correlation Coefficient

Rank in A *(x)*	***Rank in B*** *(y)*	***d = x-y***	***d2***
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49

9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
		$\Sigma d=0$	$\Sigma d^2=272$

Spearman's rank correlation coefficient P is given by :

$$p = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6x272}{15(225-1)}$$

$$= 1 - \frac{6 \times 272}{15 \times 224} = 1 - \frac{17}{35} = \frac{18}{35} = 0.51$$

Example 8 : Ten competitors in a beauty contest are ranked by three judges in the following order :

1st Judge :	1	6	5	10	3	2	4	9	7	7
2nd Judge :	3	5	8	4	7	10	2	1	6	9
3rd Judge :	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution : Let R_1, R_2 and R_3 denote the ranks given by the first, second and third judges respectively and let P_{ij}, be the rank correlation coefficient between the ranks given by 6th and 7th judges, i ≠ j = 1, 2, 3.

Let dij = R_i - R_j, be the different of ranks of an individual given by the ith and jth judege.

Calculation of Rank Correlation Coefficient

R_1	R_2	R_3	d_{12} $=R_1-R_2$	d_{13} $=R_1-R_2$	d_{23} $=R_1-R_2$	d_{12}^2	d_{13}^2	d_{23}^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
			$\Sigma d_{12}=0$	$\Sigma d_{13}=0$	$\Sigma d_{23}=0$	Σd_{12}^2 =200	Σd_{13}^2 =60	Σd_{23}^2 =214

We have n=10

Spearman's rank correlation coefcients are given by :

$$P_{12} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = \frac{7}{33} = -0.2121$$

$$P_{13} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = \frac{7}{11} = 0.6363$$

$$P_{23} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = \frac{49}{165} = -0.2970$$

Since P_{13} is maximum, the pair of first and thrid judges has the nearest approach to common tastes in beauty.

Remark. Since P_{12} and P_{23} are negative, the pair of judges (1,2) and (2,3) have opposite (divergent) tastes for beauty.

Case (ii) When Ranks are Not Given.

Spearman's rank correlation formula (5.9) can also be used even if we are dealing with variables which are measured quantitatively, i.e., when the actual data but not the rank relating to variables are given. In such a case we shall have to convert the data into ranks. The higher (smallest) observation is given rank 1 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the

variables under consideration.

Example 9 : Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data :

Advertisement cost ('000 Rs.)	39	65	62	90	82	75	25	98	36	78
Sales(lakh)	47	53	58	86	62	68	60	91	51	84

Solution. Let X denote the advertisement cost ('000 Rs.) and Y denote the sales (lakh).

Here n = 10

$$\therefore \quad P = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 30}{10 \times 99}$$

$$P = 1 - \frac{2}{11} = \frac{9}{11} = 0.82$$

Calculation of Rank correlation coefficient

X	***Y***	***Rank of X*** ***(x)***	***Rank of Y*** ***(y)***	***d=x-y***	d^2
39	47	8	10	-2	4
65	53	6	8	-2	4

62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
78	84	4	3	1	1
				$\Sigma d=0$	$\Sigma d^2=30$

Equal Ranks

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a case it is customary to give each individual as average rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$, that is 5.5 while if three are ranked equal at fifth place they are given the rank $\frac{5+6+7}{3} = 6$. In other words, where two or more items are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

Where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustemnt consists of adding ङ्करष्(m^2-m) to the value of ΣD^2, where m

stand for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times the number of such groups. The formula can thus be written :

$$r_s = 1 - \frac{6\{\Sigma D^2 + \frac{1}{12}(m^2-m) + \frac{1}{12}(m^2-m) + ...\}}{N^2-N}$$

Illustration 10. From the following data of the marks obtained by 8 students in the Accountancy and Statistics papers compute Rank Coefficient of Correlation.

Marks in Accountancy	15	20	28	12	40	60	20	80
Marks in Statistics	40	30	50	30	20	10	30	60

Solution :

Computation of Rank Correlation

Marks in Accountancy X	***Rank assigned R_1***	***Marks in Statistics Y***	***Rank assigned R_s***	***(R_1-R_2) D***	***D^2***
15	2	40	6	-4	16.0
20	3.5	30	4	-5	0.25
28	5	50	7	-2	4.00
12	1	30	4	-3	9.00
40	6	20	2	-14	16.00

60	7	10	1	-6	36.00
20	3.5	30	4	-5	0.25
80	8	60	8	0	0.00
					ΣD^2=81.5

$$r_s = 1 - \frac{6\{\Sigma D^2 + \frac{1}{12}(m^2-m) + \frac{1}{12}(m^2-m) + ...\}}{N^2-N}$$

Here $\Sigma D^2 = 81.5$,

The item 20 is repeated 2 times in series X, so m=2. In series Y the item 30 occurs 3 times, so m=3.

Substituting these values in the above formula.

$$r_s = 1 - \frac{6\{81.5 + \frac{1}{12}(2^2-2) + \frac{1}{12}(3^2-3)\}}{8^2-8}$$

$$r_s = 1 - \frac{6\{81.5 + 0.5 + 2)}{504} = 1 - \frac{6x84}{504} = 1 - \frac{504}{504} = 0$$

Merits and Limitations of the Rank Correlation

Merits. 1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answer obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, i.e., all the items are different.

2. Where the data of a qualitative nature like honesty, efficiency, intelligence, etc. this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying this method.

3. This is the only method that can be used where we are given the ranks and not the actual data.

4. Even where actual data are given, rank method can be applied for ascertaining correlation.

Limitations : 1. This method cannot be used for finding out correlation in a grouped frequency distribution.

2. Where the number of items exceed 30, the calculations become quite tedious and required a lot of time. Therefore, this method should not be applied where N exceeds 30 unless we are given the ranks and not the actual values of the variable.

Exercise :

1. "Regression and Correlation are two sides of the same coin". Explain.
2. Write note on the following :
   a) Linear and Non-linear correlation
   b) Correlation and Causation
   c) Coefficient of Determination
   d) Coefficient to Non-determination
   e) Rank correlation
3. Explain clearly the scatter diagram method of measuring correlation.

Do you think it is a perfect method?

4. Why is the regression line known as the line of best fit?

5. What is Spearman's rank correlation coefficient? Discuss its usefulness. When is it preferred to Karl Pearson's coefficeint of correlation?

6. Draw a scatter diagram for the data given below and fit a straight line on it by free hand drawing method :

: X	10	20	30	40	50	60	70	80
: X	32	20	24	36	40	28	38	44

7. Making use of the data given below, calculate the Coefficient of correlation r_{12}

Case :	A	B	C	D	E	F	G	H
X_1 :	10	6	9	10	12	13	11	9
X_2 :	9	4	6	9	11	13	8	4

8. Do you agree with the statement : "r=0.8 implies that 80 per cent of the data are explained".

9. The data relating to import price (Y) and import quantity (x) in respect of a given commodity are as under :

Year	***Import price***	***Quantity importer***
(March ending)	***y***	***x***
1985	2	6
1986	3	5
1987	6	4

1988	5	5
1989	4	7
1990	3	10
1991	5	9

Find the percentage of variation in imported a price that is explained by the variation in the quantity imported.

10. Two ladies were asked to rank 7 different types of lipstics. The ranks given by them are given below :

Lipsticks	A	B	C	D	E	F	G
Veena	2	1	4	3	5	7	6
Neena	1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient

11.A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair have approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The mark obtained by them are tabulated below :

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	1	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

(i) Find the correlation coefficient between the two sets of scores.

(ii) Find the rank order correlations coefficient.

❄ ❄ ❄

Lesson : 6

LINEAR REGRESSION ANALYSIS

Author : **Vetter:**

Dr. B. S. Bodla **Dr. R. K. Mittal**

Meaning of Regression

Regression analysis, in the general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which is extensively used in almost all sciences-natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves, cost functions, production and consumption functions, etc.

The literal or dictionary meaning of the term 'Regression' is 'stepping back or returning to the average value'. The term regression was first used as statistical concept in 1877 by Sir Francis Galton. He made a study that showed that the height of children born to tall parents will tend to move back, or "regress", toward the mean height of the population. He designated the word regression as the name of the general process of predicting one variable (the height of the children) from another (the height of the parents).

Prediction or estimation is one of the major problems in almost all spheres of human activity. Regression analysis is one of the very scientific

techniques for making such predictions. According to M. M. Blair "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data". The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent variable. The independent variable is denoted by X and the dependent variable by Y. The regression analysis confined to the study of only two variables at a time is termed as simple regression.

Importance of Regression Analysis

Every day, managers, irrespective of their area, make business decisions that are based upon predictions of future events. To make these predictions, they rely upon the relationship between what is already known and what is to be estimated. For example, if managers know that advertising and sales are correlated they may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales. If decision makers can determine how the known variable is related to the future event, they can aid the decision-making processs considerably. Regression analysis will show us how to determine both the nature and the strength of a relationship between two variables. We will learn to predict, with some accuracy, the value of unknown variable based on past observations of that variable and others.

Linear and Non-linear Relationships

If the relationship between two (or more) variables is best described by means of a straight line, it is called linear. In case of a linear regression the values of

the dependent variable increase by a constant absolute amount for a unit change in the value of the independent variable. In contrast, if the curve of regression is not a striaight line, the regression is termed as curved or non-linear regression. In this lesson, we shall discuss linear regression.

Difference Between Correlation and Regression Analysis

1. Correlation analysis is concerned with measuring closeness of the relationship between two or more variables. However, the objective of regression analysis is to study the nature of relationship' between the variables so that we may be able to predict the value of one on the basis of another.

2. Correlation need not imply cause and effect relationship between the variables under study. However, in regression analysis one variable is taken as dependent while the other as independent-thus making it possible to study the cause and effect relationship. But, we should always keep it in mind that the statistical evidence can only establish the presence or absence of association between variables whether causation exists or not depends purely or reasoning.

3. In correlation analysis r_{xy} is a measure of direction and degree of linear relationship between two variables X and Y. It is symmetric, i.e., $r_{xy}=r_{yx}$ and it is immaterial which of X and Y is dependent variable and which is independent variable. In regression analysis the regression coefficients b_{xy} and b_{yx} are not symmetric i.e. $b_{xy} \neq b_{yx}$.

4. There may be nonsense correlation between two variables which is due to pure chance and has no practical relevance, e.g., the correlation between the height of BBAstudents of G.J. University and marks obtained by them in annual examination. However, there is nothing like nonsense regression.

5. Whereas correlation coefficient is independent of change of scale and origin, regression coefficients are independent of change of origin but not of scale.

Lines of Regression

When we are concerned with two variables X and Y, we shall have two regression lines as the regression of Y on X and the regression of X on Y. Line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X. Similarly, line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y.

In the scatter diagrams (explained in lesson 5) we have used to this point, the regression lines were put in place by fitting the lines visually among the data points. In this lesson, we shall learn how to calculate the regression line somewhat more precisely, using an equation that relates the two variables mathematically. Here, we are concerned with examining only linear relationships involving two variables.

The equation for a straight line where the dependent variable Y is determined by the independent variable X is :

$$Y = a + b X \quad \ldots\ldots\ldots\ldots (6.1)$$

Y = Dependent Variable

X= Independent Variable

a = Y- intercept

b = Slope of the line

By making use of this equation, we can take a given value of X variable and compute the value of Y. The parameter 'a' determines the level of the fitted line (i.e. the distance of the line directly above or below the orign). In other words, its value is the point at which the regression line crosses the Y-axis- that is, the vertical axis. The b in above equation is the "slope" of the line. It represents how much each unit change of the independent variable X changes the dependent variable Y. Both a and b are numerical constants.

Example 1 : Make use of regression equation : Y=a+bX and determine the value of Y for an X equal to 10, assuming a is 4 and b is 2

Solution :

Regression equation is

Y=a+bX

Where a=4, b=2, and X=10.

Substitute the values of a, b and X in given equation :

Y=4+2(10)

= 4+20

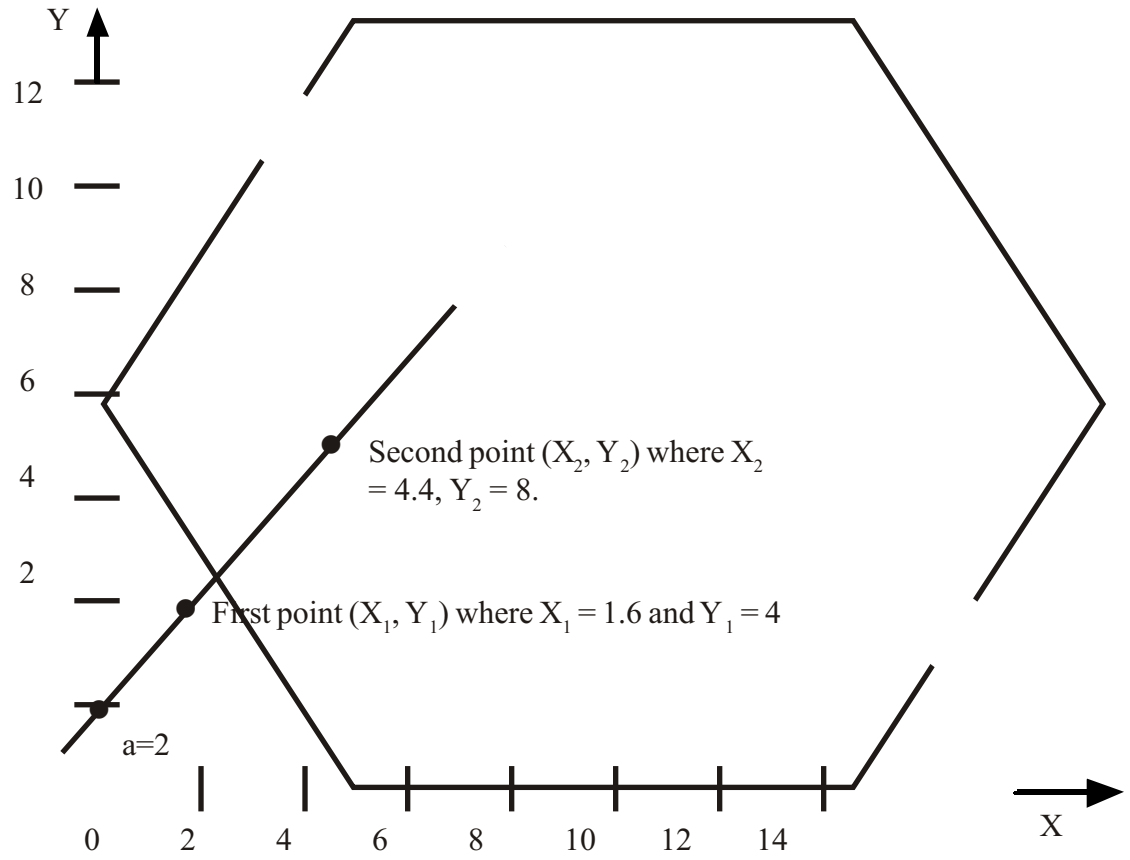
= 24

Finding the Values for a and b Constants

Lte's use the straight line in Figure 1. to understand the process of finding the values for a and b. We can find 'a' visually, by locating the point where the line crosses the Y-axis.

For finding the value of b (i.e., slope of the regression line), first we will

Fig. 1



determine how the depdndent variable, Y, changes as the independent variable, X, changes. For this purpose choose two points on the regression line in Fig. 1. Now, find the values of X and Y (the coordinates) of both points. For the first point on the line, coordinates are (X_1, Y_1) and those of the second point (X_2, Y_2). If we observe the figure under consideration, we will find that the coordinates of first point (X_1, Y_1 = 1.6, 4 and the coordinates of the second point (X_2, Y_2) = (4.4, 4, 8). Now, the value of b can be calculated by using the following equaiton :

$$b = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad \text{..... (6.2)}$$

$$b = \frac{8-4}{4.4 - 1.6} = \frac{4}{2.8}$$

$$b = 1.43$$

Hence, 1.43 is the slope of the line. After having determined the values of a and b, we can describe the line in Fig. 1 by following equation :

$$Y=a + bX$$

$$Y=2 + 1.43\ X \text{ (because, } a=2, b=1.43)$$

Fitting a Straight Line by The Method of Least Squares :

The least square method of fitting a regression line or a line of best fit involves minimising the squares of vertical deviations of each observed Y value from the fitted line. These deviations are shown in Fig. 2 and are given as $Y-Y_c$ where Y is the observed value and Y_c the corresponding individual value of the estimated point (i.e. the point that lies on the estimating line) given by the equation for the estimating line.

$$Y_c = a+bX \qquad \text{..... (6.3)}$$

for the ith value of X. Since our purpose is to fit the estimating line that minimises the sum of the squares of the errors, we call this the least squares method. The task of fitting a straight line reduces itself to only computation

of the value of a and b as it is completely described by these two constants. In the method of least squares the values of a and be are found by solving simultaneously the following two normal equations :

$$\sum Y = aN + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2 \qquad \text{..... (6.4)}$$

Where :

X = values of the independent variable

Y= values of the dependent variable

N = total number of paired observations in the sample

b = slope of the best-fitting estimate line

Table 1

X	Y	XY	X^2	Y^2	Y_c	$(Y-Y_c)^2$
2	1.0	2.2	4	1.2	1.32	0.102
3	2.5	7.5	9	6.3	1.78	0.518
5	2.0	10.0	25	4.0	2.70	0.490
7	4.0	28.0	49	16.0	3.62	0.144
8	4.0	32.0	64	16.0	4.08	0.006
$\sum X=25$	$\sum y=13.5$	$\sum Xy$	$\sum X^2$	$\sum y^2=25$		$\sum Y-Y_c)^2$
X=5	Y=2.7	=79.5	=151	=43.3		=1.260

We have obtained the quantities $\sum X$, $\sum X^2$, $\sum Y$ and $\sum XY$ from the data given in

Table 1. Now, by substituting these quantities in the two normal equations, we have

$$13.5 = 5a + 25\ b$$

$$79.5 = 25a + 151\ b$$

When these two are simultaneously solved for constants a and b, we get

$$a = 0.4,\ b = 0.46$$

We have observed that the simultaneous solution of the two normal equations for a and b may be cumbersome and time consuming. Therefore, the statisticians have developed equations we can use to find directly the slope and the Y-intercept of the regression line. The first formula calculates the slope :

$$b = \frac{\sum XY - N\ X\ Y}{\sum X - NX^2} \qquad \text{..... (6.5)}$$

where N = number of data points (i.e., the number of pairs of values for X and Y variables).

The second formula calculates the Y-intercept of the line whose slope we calculated using equation (6.5).

$$a = Y - bX \qquad \text{..... (6.6)}$$

Where :

a = Y-intercept

b = slope of equation (6.5)

Y = mean of the values of the dependent variable

X = mean of the values of the independent variable

With the information in Table 1, we can now use the equations for the slope and Y-intercept to find the numercial constants for our regression line. The slope is :

$$b = \frac{\sum XY - N\ X\ Y}{\sum X - NX^2}$$

$$= \frac{79.5 - (5)\ (5)\ (2.7)}{151 - (5)\ (25)}$$

$$= \frac{12}{26} = 0.46$$

And the Y-intercept is :

$$a = Y - b\ X$$

$$= 2.7 - (0.46)\ (5)$$

$$= 2.7 - 2.3 = 0.4$$

Now, to get the estimating equation that describes the relationship between the variables X and Y, we can substitute the values of a and b in the general equation for a straight line :

$$Y_c = a + bX$$

$$Y_c = 0.4 + 0.46X \qquad \text{..... (6.7)}$$

Further, to fit the line of best fit on the scatter diagram, only two computed Y_c values can be easily obtained by substituting any two values of X in Eq. (6.7). Then we can plot these two computed Y_c values on the scatter diagram against their corresponding values of X to get two points. When joined by a straight

line, these two points would give us the required line of best fit.

For the data given in Table 1, when X=2 and Y=1, the corresponding Y_c value is

$$Y_c = 0.4 + 0.46(2) = 1.32$$

Similarly, when X = 8 and Y=4, the computed value $Y_c = 4.08$. Then, as shown in Fig. 2, the two points are plotted on the scatter diagram such that $Y_c = 1.32$ agaisnt X=2, and $Y_c = 4.08$ against X=8. These are joined to get the line of best fit which is designated as $Y_c = 0.4 + 0.46$

Example 2 : From the following data obtain the two regression equations :

X	6	2	10	4	8
Y	9	11	5	8	7

Solution :

Obtaining Regression Equations

X	***Y***	***XY***	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum X=30$	$\sum Y=40$	$\sum XY=214$	$\sum X^2=220$	$\sum Y^2=340$

Regression equation of Y on X : $Y_c = a + bX$

To determine the values of a and b the following two normal equations are to be solved.

$$\sum Y = Na + b\sum X \quad \text{..... (i)}$$

$$\sum XY = a\sum X + b\sum X^2 \quad \text{..... (ii)}$$

Substituting the values : $40 = 5a + 30b$

$$214 = 30a+220b$$

Multiplying equation (i) by 6, $240 = 30a+180b$ ..... (iii)

$214 = 30a+220b$ ..... (iv)

Deducting equation (iv) from (iii) $-40b = 26$ or $b=-0.65$

Substituting the value of b in equation (i)

$40=5a+30(-0.65)$ or $5a = 40+19.5=59.5$ or $a=11.9$

Putting the values of a and b in the equation, the regression of Y on X is

$Y_c=11.9-0.65X$

Regression line of X on Y : $X_c = a+bY$ and the two normal equations are :

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y +b\sum Y^2$$

$30 = 5a + 40b$ .....(i)

$214 = 40a + 340b$ ..... (ii)

Multiplying equation (i) by 8 : $240=40a+320b$ ..... (iii)

$214 = 40a+340b$ ..... (iv)

From eqn. (iii) and (iv)

$-20b=26$ or $b = -1.3$

Substituting the value of b in equation (i) ; $30 = 5a + 40 (-1.3)$

$5a = 30 + 52 = 82$ $\therefore a=16.4$

Putting the values of a and b in the equation, the regression line X on Y is

$X = 16.4-1.3Y$

Regression Equations When Deviations Taken from Arithmetic Mean of X and Y

(i) Regression Equation of X on Y

$$X - \overline{X} = \frac{\sigma x}{\sigma y}(Y - \overline{Y}) \qquad \text{..... (6.8)}$$

r is known as the regression coefficient of X on Y and it is symbolized as b_{xy}

It measures the change in X corresponding to a unit change in Y

$$b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{\sum XY}{\sum Y^2}$$

Where :

$x = X - \overline{X}$

$y = Y - \overline{Y}$

(ii) Regressron Equation of Y on X

$$Y - \bar{Y} = r \frac{\sigma y}{\sigma x}(X - \bar{X}) \qquad \text{..... (6.9)}$$

Where r is the regression coefficient of Y on X (i.e. b_{yx}, and

$$r \frac{\sigma y}{\sigma x} = \frac{\sum xy}{\sum x^2}$$

Predicting An Estimate and Its Preciseness

The next process we need to learn in the study of regression analysis is how to use the regression line for predicting the most likely value of the dependent variable corresponding to a given, known, value of the independent variable. This can be easily done by substituting in the regression eq. 6.7) andy known value of X corresponding to which the most likly estimate of Y is to be found. When we assume X = 10, the estimated value of Y (i.e., Y_c) is

$$Y_c = 0.4 + 0.46\ (10) = 5$$

Thus, we see that the regression equations enable us to estimate (predict) the value of the dependent variable for any given value of the independent variable. The estimates so obtained from the regression equations are, however not perfect. The difference between the estimated Y_c values and corresponding observed Y values would depend on the extent of scatter of various points around the line of best fit.

The estimated Y_c values coincide the observed Y values only if all the points

on the scatter diagram fall on a straight line. But such a situation is encountered very rarely. It may be said that the estimated values of one variable based on known values of the other variable are always bound to differ. The greater the difference, the lesser the precision of the estimate and vice-versa. A measure of the precision of the estimates obtained from the regression equations is provided by the Standard Error of the estimate.

The Standard Error of Estimate

To measure the reliability of the estimating equation, statisticians have delivered the "Standard error of estimate". The standard error is similar to the standard deviation. We may say that these both are measures of dispersion. The standard deviation measure the dispersion about an average such as the mean. The standard error of estimate measures the dispersion about an average line, called the regression line. The standard error of estimate of Y on X is symbolised by $\sigma y.\ x$. The following is the equation for calculation of standard error of estimate :

$$s_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N - 2}} \qquad \text{..... (6.10)}$$

Where :

Y = values of the dependent variable

Y_c = estimated values of Y variable

N = number of data points used to fit the regression line.

To calculate Sy.x we refer again to our earlier sample data of Table 2.1. For this purpose, we must first determine the value of $\sum (Y - Y_c)^2$. It has been done

in Table 1. We have found $\sum(Y - Y_c)^2 = 1.26$. Now, use the Eq. (6.10) and find Sy.x.

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N - 2}}$$

$$= \sqrt{\frac{1.26}{5 - 2}} \quad = 0.65 \qquad \text{..... (6.8)}$$

This formula involves tedious calculations because it requires the computation of $(Y-Y_c)$. Fortunately, some of the steps in this task can be eliminated by using the following equation :

$$Sy.x = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N - 2}} \qquad \text{..... (6.11)}$$

Now we can refer to our earlier Table 1 and our previous calculations of a and b in order to calculate Sy.x using Eq. (6.11).

$$Sy.x = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N - 2}}$$

$$Sy.x = \sqrt{\frac{43.3 - (.4)\ (13.5) - 0.46\ )\ (79.5)}{5 - 2}}$$

$$Sy.x = \sqrt{\frac{43.3 - 5.4 - 36.57}{3}} = 0.65$$

This is the same result as the one obtained using Equation (6.10).

Interpreting the Standard Error of Estimate

Similar to standard deviation, the larger the standard error of estimate, the greater the scattering of points around the regression line. In contrast, if Sy.x = 0. it is expected that the estimation equation is a 'perfect' estimator of the dependent variable. Similar to the property of arithmetic mean that the sum of the deviation of $\overline{Y}$ values from Y is equal to zero, the sum of the deviations of different Y values from their corresponding estimated Y_c values is also equal to zero {i.e. $\sum (Y-Y_c) = 0$}.

Standard error of estimate is measured along the Y-axis, rather than perpendicularly from the regression line. Sy.x tells us the amount by which the estimated Y values will, on an average, deviate from the observed Y values. It means Sy.x is an estimate of the average amount of error in the estimated Y_c values. Further, we shall use the standard error of estimate as a tool in the same way that we can use the standard deviation. It implies that we can expect to find 68 per cent of the points within –1 Syx, 95.5 per cent of the points within –2 Sy.x, and 99.7 per cent of the points within – 3 Sy.x.

Moreover, the standard errr of estimate also serves as a measure of the reliability of the estimate since Sy.x measures the closeness of the observed Y values and the estimated Y_c value.

Illustration 5 : From the data given in example 2:

Calculate the standard error of the estimate (S_{yx} and S_{xy})

Solution :

X	Y	Y_c	X_c	$(Y-Y_c)^2$	$(X-X_c)^2$
6	9	8.0	4.7	1.00	1.69
2	11	10.6	2.1	0.16	0.01
10	5	5.4	9.9	0.16	0.01
4	8	9.3	6.0	1.69	4.00
8	7	6.7	7.3	0.09	0.49
$\Sigma X=30$	$\Sigma Y=40$	$\Sigma Y_c=40$	$\Sigma X_c=40$	$\Sigma(Y-Y_c)^2=3.1$	$\Sigma(X-X_c)^2=6.20$

$$Sy.x = \sqrt{\frac{\Sigma (Y - Y_c)^2}{N}} = \sqrt{\frac{3.1}{5}} = \sqrt{0.62} = 0.787$$

$$Sy.x = \sqrt{\frac{\Sigma (X - X_c)^2}{N}} = \sqrt{\frac{6.2}{5}} = \sqrt{1.24} = 1.114$$

Example 6 : For a set of 10 paired observations on X and Y, the coefficient of correlation is 0.856 and the standard deviation of Y is 6.50. Find the standard error of estimate of Y on X.

Solution :

We have the following information :

$r = 0.856$ and $S_y = 6.5$

r^2 = will be 0,7327

$S_{y.x} = S_y \sqrt{1-r^2}$

$= 6.5 \sqrt{1-0.7327} = 6.5\sqrt{0.2673}$

$= (6.5)\ (0.517)$

$= 3.36$

Do yourself

1. What is regression analysis ? How does it differ from correlation ?

2. Discuss the importance of regression analysis in business decisions.

3. Why is the regression line known as the line of best fit ?

4. What is standard error of estimate ? How is it measured ?

5. Give interpretations of standard error of estimate.

6. Find the regression equations from the following data :

Age of Machines (X)	5	3	3	1
Repair Expense (Y)	7	7	6	4

7. Given the following data :

X	5	3	3	1
Y	7	7	6	4

Calculate the standard error of the estimate ($S_{y.x}$) and $S_{x.y}$).

8. The data for 10 years on sales (Y) and advertsement expenditure (X) of a particular product yielded the following summariesed values (Rs. in lac)

$\sum X=15$, $\sum Y=110$, $\sum XY = 400$, $\sum X^2 = 250$, and $\sum T^2 = 3200$.

Find (i) Regression coefficient b of Y on X, and then the Y-intercept.

(ii) Most approximate value of Y for X =5.

(iii) Standard error of estiamte $X_{y.x}$

9. If X=10, Y = 12, $\sum XY = 150$, $\sigma_x = 4.5$, $\sigma_y = 6.0$ and N = 11 find the following :

(i) Regression Coefficient b of Y on X.

(ii) Regression Coefficient b of X on Y.

(iii) Correlation Coefficient between X and Y.

❄ ❄ ❄

Lesson : 7

INDEX NUMBER

Using, Problem in Constructing Index Numbers, and Method of Constructing Index Numbers

Author : **Vetter:**

Dr. S. S. Tasak **Dr. R. K. Mittal**

INTRODUCTION :-

Index numbers are today one of the most widely used statistical indicators Generally used to indicate the state of the economy, index numbers are aptly called 'barometers of economic activity'. Index numbers are used in comparing production, sales or changes in exports or imports over a certain period of time. The role played by index numbers in Indian trade and industry is impossible to ignore. It is a very well known fact that the wage contracts of workers in our country are tied to the cost of living index numbers.

By definition, an index number is a statistical measure designed to show changes in a variable or a group or related variable with respect to time, geographic location or other characteristic such as income, profession etc. Index number is calculated as a ratio of the current value to a base value and expressed as a percentage. It must be clearly understood that the index number for the base year is always 100. An index number is commonly refferred to as an index.

Index Number of wholesale prices

Base year 1981-82 (=100)

1992-93	Primary Articles	Manufactured Product	All Commodities
July	237.6	226.6	226.6
August	240.8	224.7	228.8
September	237.9	227.7	230.7
October	237.1	229.3	232.4
November	235.8	228.7	231.7
December	235.0	228.7	231.4
January	235.0	228.6	231.6
February	234.6	229.9	232.8
March	232.9	230.8	233.1
1993-94			
April	234.2	231.2	234.6
May	234.6	232.2	235.2
June	234.9	233.5	237.7

Source: CMIE,July 1993

Using these data, one can find out that the wholesale price for primary articles (comprising food articles, non-food articles and minerals) in April 1993 was 234.2 percent of theaverage wholesale price for primary articles in 1981-82. Similarly in June 1993, the wolesale price for all commodities was 237.7 times that of the wholesale price prevalent in the yeart 1981-82, on an average, for all commodities.

An index number is an average with a difference. An index number is used for purposes of comparison in cases where the series being compared could be expressed

in different units. i.e., a manufactured products index (a part of the wholesale price index) is constructed using items like Diary products, Sugar, Edible Oils, Tea and Coffee etc. These items naturally are expressed in different units like sugar in kg, milk in litres etc. The index number is obtained as a result of an average of all these items which are expressed in different units. On the other hand, average is a single figure representing a group expressed in the same units.

Index numbers essentially capture the changes in the group of related variables over a period of time. For example, if the index of industrial production is 215.1 in 1992-93 (base year 1980-81) it means that the industrial production in that year was up by 2.15 times as copared to 1980-81. But it does not however mean that the net increaser in the index reflects an equivalent increase in industrial production in all sectors of the industry. Some sectors might have increased their production more than 2.156 times while other sectors may have increased their production only marginally.

Characteristies of Index Number-:

For a proper understanding of the index numbers, one should be clear about its characteristics. The following are the important characteristics of an index number

1. These are expressed in percentage: Index numbers are expressed in terms of percentages so as to show the extent of relative change.

2. These are relative measure: Index numbers are specialised averages used to show the relative change in group of related variables. The group of variables may relate to prices of certain commodity of volume of production of certain items. They compare changes taking place over time or between places. If the wholesale price index for the year 1990 is 140 as compared to 100 in 1988, then we conclude that the general price level has increased by 40% in two years.

3. Index numbers are specialise averages: Simple averages can be use those series which are expressed in the same units. However, index numbers are special type of averages which are used in comparing changes in series expressed in different units. In view of this they are also called specialised averages.

4. Index numbersmeasure changes which are not directly measurable : The index numbers are used for measuring the magnitude of changes in such phenomena which are not capable of direct measurement. Fir example, price level, cost of living and ups and downs in business activities are phenomena in which changes in directly measurable factor affecting price level, cost of living and business activities, inded numbers help us to measure relative changes in corresponding phenomenon which is otherwise not directly measurable.

Uses of Index Numbers :

The important uses of index numbers are described below :

1. They are economic barometer :- Index numbers are mainly used in business and economics. Like barometers are used in physics ton measure atmospheric pressures, index numbers measure the level of business and economic activities and are, therefore, termed as 'econoic barometers' or 'barometers of eco nomic activity' For instance, price index numbers are those that relate to changes in the level of prices of commodity or a group of commodities over a period of time. Price index numbers are useful in studying price movements and determininig their effect on economy. It is often useful to compare changes in general price level with changes in related series, such as ban deposits, bank loans, etc.,for formulating economic policies.

2. The measure comparative changes :- The important purposes of index numbers is to measure the relative change in a variable or a group of related variables in respect

to time or place. The changes in the phenomena like price level, cost of living, etc., are not capable of being measured directly and are, therefore, measured with the help of index numbers. Indices of physical changes resulting over a period of time in production, sales, imports, etc., are extremely helpful in analysing the movements in these characteristics over time.

3. They help in forecasting : Many governmental and private agencies are engaged in computation of index numbers for purpose of forecasting business and economic condition. For example, index number of industrial and agriculture production not only reflect the trend but can also help in forecasting future production Similarly, index numbers of unemployment in a country not only reflect the trends in the phenomenon but are useful in determining factors leading to unemployment. The analysis os such trend and factors in unemployment activity help in framing a suitable employment policy.

4. They measure the purchasing power of money :- Consumer price index numbers are useful in finding the intrinsic worth of money as they are used for adjusting the original data related to wages for price changes. In other words, index numbers are helpfur for transforming nominal wages into real wages. Based on this aspect, the Government of India of the states use consumer price index numbers for determining the amount of additional wages or dear ness allowances to be given to their employess to compensate for changes in the price level or cost of living. Thus, most of the government now have index linked salary structures and additional dearness allowance is granted to employees for a point rise iun consumer price index.

5. They measure the real gross national product (G.N.P.) : Index numbers are also used for determining the real gross national product (G.N.P) or income calculated at current prices. Real G.N.P. is determined by price index of the current year., i.e.,

Real G.N.P. = G.N.P. at current price X 100
Current price index

Problems in the Construction of Index Number

Before constructing index numbers, a careful study of the following related problems be made :

1. The purpose of the index number : Index numbers are constructed for serving specific purposes, Therefore, it is important to know what kind of changes we are trying to measure and how we intend to use them. Obviously, a clear defination of the purpose and objective is the first major problem in the construction of index numbers. For example, if price index is to be constructed for measuring the cost of living of middle class families in a region, care must be taken to include items which are consumed by these families. Similarly, for measuring prices and not wholesale prices of these items.

2. Selection of items : Having defined the purpose of index numbers, the next problem relates to the Selection of items, In this regard, a decision about the number of items included, the more representative shall be the index but at the same time cost andtime involved in the construction of index number will increase.Therefore, the number of tiems selected should neither be too small nor too large. secondly, one should ensure that the items selected represent the tastes, habits and customs of the people for whom the index is being constructed. For instance, while computing cost of living index for middle class families, gold, car, etc. will not be relevant items. Thus, only relevant standardised items, which are easy to define and describe, should be included in the constructrion of index numbers so that they reflect the change that we wish to measure.

3. Data for index numbers : The data used in index numbers are usually concerned

with the prices and quantities consumed of the selected items for different points of time. As such, we always face the problems of selecting a reliable source of data. Thus, the data should be collected from standard trade journals, official publication, chamber of commerce and other government agencies. Data are also collected through field studies or sample surveys. Here the samples selected should be representative of the class to which they belong and then only the resulting data is expected to be reliable, accurate and homogeneous. For uniformity, it is often desirable to group the items into homogeneous groups of subgroups. For instance, in measuring price changes, domestic items may be grouped into cereals, mill. edible oils, clothing, electricity and fuel, etc. Similarly, items with elastic demand. Type of price quotations is another consideration while collecting data. A wholesale price index needs wholesale price quotations while retail price quotations will be desirable will be desirable in the construction of cost of living index number.

4. Choice of base poeriod : The period with which the comparison of the relative changes in the level of a phenomenon are made is termed as 'base period of 'reference period'. The index for the base period is always taken as 100. For reliable and precise comparisons of the relative changes, a base year should be a sufficiently 'normal' year. It should be period free from all abnormalities like economic boom or depression, labour strikes, wars, earthquakes, etc. In other words. the base period free from all abnormalities, etc. In other words. the base period should be more or less stable and free from unusual ups and downs. It is also desirable that the base period should not be too distant from the given period with which relative changes are measured. In case the base period is too distant, it is desirable to shift t it, For example, it seems undesirable to compare prices of commodities must have changed since then. Notable technological developments, rising income of the people, changing patterns of consumption, quality

of goods, changes in habits and tastes of the people are some factors which compel the shifting of the base period. The base period may be of two types :

(i) Fixed base period.

(ii) chain base period.

{i} Fixed base period : If the base period or reference period is kept fixed for all current periods of comparison, it is called the fixed base period. For example, the year 1951, being the first year of planning process, may be taken as the base period for studying relative planning development in the current years/

{ii} Chain base period : In chain method, the change in the level of the phenomenon for any given period is compared with the level of the phenomenon in the preceding period and not to the base period.

5. Choice of an average : We observed that index numbers are special typeof averages.An such, the choiceof a suitable average is also important in the construction of index numbers. Arithmetic mean, median and geometric mean are the commonly used averagesin index numbers. Out of the three averages, arithmetic mean and a median are comparatively easier to calculate. However, mediancompletely ignores the extreme observations while the arithmetic mean is unduly affected by such observation. The geometic mean is the most suitable average in the course struction of index numbers in view of the following properties :

{i} Geometric mean gives more importance to smaller items and less to larger items and is, therefore, least affected by the values of the extreme items.

{ii} Geometric mean gives equal weithts to equal ration of changes.

In spite of theoretical justification for its suitability, geometric mean is not a

common average in the construction of index numbers. It is view of its difficult computational process. For simplicity in calculations. arithmetic mean is used instead, however, geometric mean is recommended for greater accuracy and precisicion.

6. Selection of weights : unweighted index numbers give equal importance to all commodities. However, all items or commodities included in the construction of an index number are not of equal importance. For example, in the construction of cost of living index, sugar cannot be given the same importance as the cereals. In order to allow each commodity to have a reasonable influence on the index, we make use of weighted index numbers which give appropriate weight to different commodities according to their importance. The selection of appropriate weight is again a difficult task.

The method of asigning weight are :

{a} Implicit (or arbitrary) {b} Explicit (or actual).

In implicit weighting, a commodity or its variety is included a a number of times according to its importance. on the other hand, in explicit weighting, some actual criteria is used for assigning weights to different commodities included in the index number.

Generally, the weights for various commodities are decided according to {i} Value or produced, {ii} Value or quantity consumed, {iii} Value or quantity sold). When the quantity (value) is the basis of weight, we call it quantity as weights. On the other hand, in the method of averaging price relatives, values are used as weights

7. Selection of suitable formula Selection of a suitable formula for construction of any index number also poses some problems. There are various formulae for calculating index numbers such as the aggregate method or the average of relatives

method in simple aggregate method the price of each commodity is given in usual units and this leads to the dominance of a particular quantity in the index. The difficulty is remove by considering the average of relatives methods. But in this method we assume that each item is purchased for an equal amount of money in the base year i.e. the value of all the items in the base year is the same. This leads to the concepts of weighted in ex formula where items are waited according to their relative importance in this regard a separate section is devoted for method of constructing index No. there suitability in a particular situation.

Method of constructing index numbers

A number of formulae have been developed for constructing indexnumbers which may be grouped into the following categories-

1. Simple or unweighted index numbers

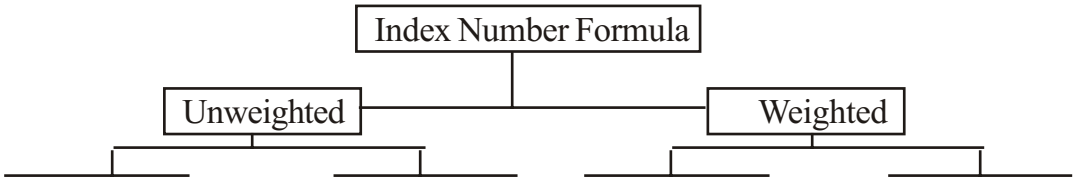
2. Weighted index numbers

Various index formulae in each of the above to categories may be further classified as-

(a) Simple aggregative method

(b) Method of simple average relatives

The following chart may be used define the above classification-



Simple aggregative, Simple average of relatives, Simple aggregative, Simple aggregatives relatives

Notations :

For proper understanding of various index number formula, the understanding of the following notations is necessary:

Po(j) : Price of jth commodity in the base year, (j=1,2,.....N)

Pn(j) : Price of jth commodity in the current year, (j=1,2.....N)

qo (j) : Quantity of jth commodity consumed or purchased in the base year:,(j1,2,.....N)

qn(j) : Quantity of jth commodity consumed or purchased in the current year., (1,2.....N)

w(j) : Weight assigned to jth commodtity according to its relative importance, (j=1,2.....N)

Pon : Price index for the current year (n) with respect to the base year (O)

Pno : Price index for the base year (O) with respect to the current year (n)

Qon : Quantity index for the current year (n) with respect to the base year (O)

Qno : Quantity index for the base year (O) with respect tothe current year (n)

Von : Value index for the current year (n) with respect to the bas year (O)

Vno : Value index for the base year (o) with respect to the current year (n).

Price Relatives, Quantity Relatives and Value Relatives

Price Relatives :

Price relative is one of the simpleast example of an index number. It is defined as the ratio of the price of a single commodity in the current year (n) to its price in the base year (o). Therefore, the price relative of period (n) with respect to period (o) is

$$\text{Price Relative} = \text{Pon} = \frac{Pn}{Po} \quad \text{.....(1)}$$

It is customary to express price relative as percentage by multiplying it by 100. Thus, the price relative in (1) expressed in percentage becomes

$$\text{Oruce Relative} = \text{Pon} = \frac{Pn}{Po} \times 100 \quad \text{.....(2)}$$

Here it is important to note that the price relative for a given period with respect to same period is always 1 or 100 in percentage terms. In particular, we can say that the price relative of the base year is always 100. In defining the price relatives, the prices are assumed constant at one time point. However, if they very over a period, an appropriate average of the prices over the given duration is used for the purpose.

Illustration 1. Let the prices of certain item in 1985 and 1990 were Rs. 110 respectively Then taking the year 1985 as base,m the price relative for the year 1990 will be -

$$\text{Price Relative for the year 1990} = \text{Pon} = \frac{Pn}{Po} \times 100 = \frac{110}{70} \times 100 = 157.14\%$$

As Pn= the price in the current year,i.e., in 1990=110 Rs.

and Po=the price in the base year 1985 = 70 Rs.

Illustration 2. The Price of two commodities in the year 1990 and 1991 are shown in the table. The price relatives of the two commodities are also given assuming 1990 as the base year -

Table Showing Price Relative

Price			Price Relative
	1990 (Po)	1991(Pn)	Pon=Pn/PoX100
A	55	65	118.18%
B	76	89	117.11%

Quantity Relatives

Quantity relative are another type of index numbers used for measuring changes in quantum or volumes of a commodity, such as quantity of production, consumption. exports, imports, sales, etc. Thus, in this case the commodity is used in a more general sense. It may mean volume, of exports, imports, sales, production, number of passengers travelling by railways and so on. As in the case of prices, quantities too are assi, ed constant for any period, otherwise, an appropriate average is used for the purpose to mae this assumption valid. Now, the quantity relative is defined as the ratio of the quantity of a single commodity in the current year (n) to its quantity in the base period (o), Thus,

Quantity Relative=Qon=qn/qo.....(3)

Expressing this ratio in percentages, we get,

Quantity Relative=Qon=qn/qo X 100 .....(4)

Quantity relatives too have the same properties as that pertaining to price relatives.

Illustration 3. Let us consider the data on production of wheat of a country in millions of tons for three years-

Years	1989	1990	1991
Production of Wheat (Millions of tons)	1090	988	1306

Based on the above data, the quantity relatives for the year 1990and 1991 taking 1989 as the base are shown in the following table.

Years	1989	1990	1991
Production of wheat	1090(qo)	988(ql)	1306 (q2)

Quantity Relative	Q00=$\frac{1090}{1090}$X100	$\frac{988}{1090}$X100Q02	$\frac{1306}{1090}$X100
	=100%	=90.64%	=119.82%

Value Relatives

A value relative is yet another type of index number. It is used in a situation when one is willing to compare changes in the monetary value of consumption, sale export, import of commodity on two or more points of time. If p and q are the price and quantity of a commodity produced, consumed or sold, during a period, then v = pq gives the total money value of the transaction. Her a also p and q are assumed constants at a time point, otherwise a suitable average of prices and quantities over the given period is taken to make this assumption valid. In usual notations, sthen we define.

Vo= poqo : total money value during the base period.

vn = pnqn : total money value during the current period.

There fore, the value relative of the current year is the ratio of vn to vo, i.e.

$$\text{Value Relative} = \frac{V0n}{} = \frac{vn}{vo} \qquad \text{.....(5)}$$

When expressed in percentage terms, one gets

$$\text{Value Relative} = V0n = \frac{vn}{vo} \times 100 \qquad \text{.....(6)}$$

Illustration 4. From the price per unit and the units of quantity consumed of two commodities in the years 1980 and 1985 given as under compute value relatives :

Commodity	Price per unit (Rs.) in		Quantity consumed (Units) in	
	1980	1985	1980	1985
A	7	10	40	52
B	9	12	75	80

Solution :

	Price Per unit (Rs.)		Quantity consumed In (unit)		Values in (Rs.)		Value Relatives Qon=$\frac{vn}{vo}$ X 100
Commodity	1980	1985	1980	1985	1980	1985	
	Po	Pn	qo	qn	vo=poqo	vn=pnqn	
A	7	10	40	52	280	520	$\frac{520}{280}$ x 100 =185.71%
B	9	12	75	80	675	960	$\frac{960}{675}$ X 100 =142.22%

Properties of Relatives

Iff pa, pb, pc,.....be the respective prices of the commodity in periods a, b,c....., then, with usual notations, the price relative have the following properties :-

1. Identity property :

According to this property, the price relative for a given period with respect to the same period is 1 or 100% That is -

$$Paa = \frac{pa}{pa} = 1 \quad \text{or} \quad Paa = \frac{pa}{pa} \times 100 = 100\%$$

or

$$Pbb = \frac{pb}{pb} = 1 \quad \text{or} \quad Pbb = \frac{pb}{pb} \times 100 = 100\%$$

2. **Time Reversal property**

This property states that if the current period and the base period are interchanged, then the product of the corresponding price relatives is unity. In other words, the corresponding price relatives are reciprocal of each other. Symbolic.

$$Pab = \frac{pb}{pa} \quad \text{and} \quad Pba = \frac{pa}{pb}$$

: $Pab \times Pba = \frac{Pb}{Pa} \times \frac{Pa}{pb} = 1$

or $Pab = 1/Pab$ and $Pba = 1/Pab$

3. Cyclical or Circular Property

According to this property, if the periods a., b, and c are in circular order then the product of the three relative defined with respect to the preceding period as base period is unity. In Symbols :

$$Pab.\ Pbc.\ Pca = 1$$

as $Pab = \frac{pb}{pa}$, $Pbc = \frac{pc}{pb}$, $Pca = \frac{pa}{pc}$

The property can be extended to any number of periods which are in circular order, For example, for four periods a, b, c, and d which are in circular we have

$$Pab.\ Pbc.\ Pcd.\ Pda = 1$$

4. **Modified Cyclical or circular property**

Using the concepts in properties 2 and 3, we have the following modified circular property of the price relatives.

$$Pab.\ Pbc = Pac$$

In case of four periods,

$$Pab,\ Pbc.\ Pcd = Pad$$

Illustration 5. Let us consider the following data related to prices of a commodity of three years-

Year	1980	1985	1990
Price (Rs.)	54	67	90

With the help of given data, let us check the properties of the pr relaives as discussed above. Symbolic, the given can be put as-

a = 1980 b = 1985, c = 1990

and pa = 54, pb = 67, pc = 90

Indentity property :

Here, $Paa = \frac{pa}{pa} \times 100 = \frac{54}{54} \times 100 = 100\ \%$

$Pbb = \frac{pb}{pb} \times 100 = \frac{67}{67} \times 100 = 100\%$

$Pcc = \frac{pc}{pc} \times 100 = 90 \times 100 = 100\%$

Times Reversal Property :

$Pab = \frac{pb}{pa} = \frac{67}{54}$; $Pba = \frac{pa}{pb} = \frac{54}{67}$

Obviously Pab = 1/Pba.

Similarly, we can observe that Pbc = 1/Pcb and Pac = 1/Pca

Cyclical or circular property :

Since, $Pab = \frac{pb}{pa} = \frac{67}{54}$, $Pbc = \frac{90}{67}$ $Pca = \frac{54}{90}$

$Pab \times Pbc \times Pca = \frac{67}{54} \times \frac{90}{67} \times \frac{54}{90} = 1.00$

Modified Circular Property :

Let us consider $Pac \times Pbc = \frac{67}{54} \times \frac{90}{67} = \frac{90}{54}$

Also $Pac = \frac{pc}{pa} = \frac{90}{54}$

$Pab \times Pbc = Pac$

Simple or Unweighted Index Number

Simple aggregate Method

This is the simplest method of computing index numbers. According to this method the total of the current year prices for various commodities is expressed as a percentage of the total of the base year prices for these commodities. In Symbols,

Simple aggregate price index $Pon = \frac{\Sigma Pn}{\Sigma Po} \times 100$

Here Pon = Current year index number

Pn = the total of commodity prices in the current year

Po = the total of commodity prices in the base year.

Example 1 : Using simple Aggregate method obtain index numbers for 1990 taking 1988 as the base year from the following data -

Commodities		A	B	C	D	E
Price (Rs.)	1988	100	80	160	220	40
	1990	140	120	180	240	40

Solution : Computing Index number (Simple Aggregate Method)

	Price (Rs.)	
Commodities	1988 (Po)	1990 (Pn)
A	100	140
B	80	120
C	160	180
D	220	240
E	40	40
Total	Σ Po = 600	ΣPn = 720

: Price Index for 1990 $= \frac{Pon}{\Sigma Po} = \Sigma Pn \times 100 = \frac{720 \times 100}{600} = 120$

Which shows a net increase of 20% in the price of commodities in the year 1990 as compared to 1988.

Example 2 : From the following data construct price index numbers for the years 1985 and 1990 by simple aggregate method taking 1980 as the base year -

		Price (in Rs.)		
Commodities	Unit	1980	1985	1990
Wheat	Quinital	200	250	275
Rice	Quintal	300	350	450
Arhar	Quintal	600	700	750

Milk	Litre	6	7	8
Clothing	Metre	30	35	40

Solution :

Construction of Price Indices (Simple Aggregate Method)

		Price (In Rs.)	
Commodities	1990 (P0)	1985 (P1)	1990 (P2)
Wheat	200	250	275
Rice	300	350	450
Arhar	600	700	750
Milk	6	7	8
Clothing	30	35	40
Total	Σ Po = 1136	Σ P1 = 1342	ΣP2=1523

: Price Index for 1985 = $P_{01} = \frac{\Sigma P1}{\Sigma P0} \times 100 = \frac{1342}{1136} \times 100 =$ 118.13

Price Index for 1990 = $\frac{\Sigma P_2}{\Sigma P_0} \times 100 = \frac{1523}{1136} \times 100 =$ 134.07

Price Index for various years thus becomes -

Year	:	1980	1985	1990
Index number (base 1980)	:	100	118.13	134.07

Example 3 : Prices of and article for six years are given below :

Year :	1980	1981	1982	1983	1984	1985
Prices (in Rs.) :	10	14	16	20	22	26

Obtain price index numbers (i) assuming 1980 as the base year, (ii) assuming the average price of the six years as base.

Solution : The computation of the price index numbers in the two cases is shown

in the following table -

Construction of Price Index Num

Year	Price	Index Numbers (base 1980) (pn/po) X 100	Index Numbers (Average Price 18 as 100)
1980		100	$\frac{10}{18}$ X 100 = 55.5
1981	14	$\frac{14}{10}$ X 100 = 140	14 X 100 = 77.77
1982	16	$\frac{16}{10}$ X 100 = 160	$\frac{16}{18}$ X 100 = 88.88
1983	20	$\frac{20}{10}$ X 100 = 200	$\frac{20}{18}$ X 100 = 111.11
1984	22	$\frac{22}{18}$ X 100 = 220	$\frac{22}{18}$ X 100 = 122.22
1985	26	$\frac{26}{10}$ X 100 = 260	$\frac{26}{18}$ X 100 = 144.44

Average price of the six years = $\frac{10 + 14 + 16 + 20 + 22 + 26}{6}$

$= \frac{108}{6}$ = 18 Rs.

Limitations of the simple Aggregate Method

Although the method is simple, it has the following two limitations -

1. The above formula does not consider the relative importance of various commodities involved.
2. In this formula, the prices of various commodities are generally given in different units, e.g., wheat may be quoted in Rs. per quintal ; milk, petrol in Rs. per litre ; cloth in Rs. per meter and so on. Thus the particular units used in the price quotations may affect the value of the index number.

Simple Average of price relatives Method

In this method, we first obtain price relatives for all commodities included in the index numbers. As usual, the price relative off the current year expressed as a percentage of the price of the base year is pn X 100. Now these commodity price relatives may be averaged by using averages like arithmetic mean, median mode or geometric mean. Howver, if we use arithmetic mean, the formula for computing index number becomes -

$$Pon = \Sigma \frac{\frac{pn}{Po}}{N} X\ 100$$

Here, $\Sigma \frac{Pn}{Po}$ = the sum of commodity price relatives.

N = The number of commodities.

Similarly, the geometric mean of the relatives can be used to find the index number of the current year as-------

$$Pn = \text{Anti log}\ \Sigma \left\{\frac{\log Pn}{Pn} \times 100\right\} = \text{Anti log}\ \left\{\frac{\Sigma \log R}{N}\right\} ; \text{where } R = \frac{Pnx}{Po}\ 100$$

Example 4 : Construct index number for 1990 taking 1988 as the base year from the following data by using average for price method.

Commodities		A	B	C	D	E
Price in Rs.	1998	100	80	160	220	40
	1990	140	120	180	240	40

Construction of Price Indices (Simple average of price relation)

Commodities	Price in 1998	Price in 1990	Price Relatives
A	100	140	140X100=140.00
B	80	120	120X120=150.00
C	160	180	180x100=112.50
	(Base) Po	(Current Year) Pn	(Pn/Pnx 100)
D	220	240	240x100=109.10

E	40	40	40x100=100.00
N=5			ΣPn x 100 = 611.60

Price Index No. for 1990 = $\left(\text{Pon}\right) = \left(\frac{\text{Pn}}{\text{Po}} \times 100\right) = \frac{611.60}{\text{N}} = 122.32$

Example 5 : Use the following information to construct index number for 1990 taking the price of 1985 as base. Use (i) simple aggregate method and (i) simple average of relative method in the construction.

Commodity	A	B	C	D	E
Price in Rs.	1985	12	25	10	6
	1990	15	20	15	15

Solution :

Construction of Index numbers :

Commodity	Price in 1985 (base year)		Price in 1990 (Current Year)	
	Price Po	Price Relatives	Price Pn	Price Relatives Pnx100
A	12	100	15	$\frac{15}{12} \times 100 = 125.00$
B	25	100	20	$\frac{20}{25} \times 100 = 80.00$
C	10	100	12	$\frac{12}{10} \times 100 = 120.00$
D	5	100	10	$\frac{10}{5} \times 100 = 200.00$
E	6	100	10	$\frac{15}{6} \times 100 = 250.00$
N=5	Σ Po=58	S Pn=72	$\frac{\text{Pnx}}{\text{SPo}}$100=775	

Price index for 1990

(i) By aggregate method = Pon = $\frac{\Sigma \text{Pn}}{\Sigma \text{P0}} \text{ X } 100 = \frac{72}{58} \text{ X } 100 = 124.14$

(ii) By simple average of price relatives $P_{on} = \Sigma \left(\frac{\frac{P_n}{P_o} \times 100}{N} \right) \frac{775}{5} = 155.50$

Example 6. Using arithmetic mean, median and geometric mean, construct index numbers by the simple average of relative method from the following data for 1990 and 1991 with 1989 as the base year.

Price (in Rs. per unit)

Commodity	1989	1990	1991
A	100	120	150
B	40	45	60
C	30	35	45
D	10	12	15
E	20	22	23

Solution :

Computation of price Indices [(Using Mean, Median and G.M. (base 1989)]

Articles Prices (Rs.) Price Relatives for 1990 Price Relative for 1991

	P0	P1	P2	R1=P1/P0 X 100	log 21	R2=P2/PoX100	log R2
A	100	120	150	$\frac{120}{100}$ X 100 = 120.0	2.0792	$\frac{150}{100}$ X 100 = 150.00	2.1761
B	40	45	60	45/50 X 100 = 112.5	2.0511	60/40 X 100 = 150.0	2.1761
C	30	35	45	35/30 X 100 = 116.7	2.0671	45/30 X 100 = 150.0	2.1761
D	10	12	15	12/10 X 100 = 120.0	2.0792	15/10 X 100 = 150.0	2.1761
E	20	22	23	22/20 X 100	=110.0	2.0414 23/20 X 100=	2.060

N=5 ΣR1 = 597.2 ΣLog R1 = 10.318 ΣR2=715 Σlog R2= 10.7651

(i) Price Index by using arithmetic mean of the relatives.

Price Index for 1990 = P0 = $\frac{\Sigma P1/P0}{N} \times 100 = \frac{\Sigma R1}{N} = \frac{579.2}{} = 115.80$

Price Index for 1991 = P0 n = $\frac{\Sigma P2/P0}{N} \times 100 = \frac{\Sigma R2}{N} = 715.0 = 143.0$

(ii) Price Index by median of the relatives.

Arranging the price relative in ascending order, and selecting the size of $\frac{N+1th}{2}$ = 3rd term, we have

Price Index for 1990 = 116.70

Price Index for 1991 = 150.00

(iii) Price Index by using G.M. of the relatives.

Price Index for 1990 = anti log ($\frac{\Sigma \log R1}{N}$) = anti log ($\frac{10.3180}{5}$)

anti log [2.0636] = 115.8

Price Index for 1991 = anti log ($\frac{\Sigma \log R2}{N}$) = anti log ($\frac{10.7651}{5}$)

anti log [2.1530] = 142.2

Thus, price indices for various years can be summarised as -

Used Average	Price indices (Years)		
	1989	1990	1991
A.M.	100	115.8	143.0
Median	100	116.7	150.0
G.M.	100	115.8	142.2

Example 7. Taking of I year as base, construct the index numbers for II and III years from the following data. Use the simple average of relatives method.

Year Articles (Rate per Ruppes)

	A	B	C
I	4 kg	2 kg	1 kg
II	2.5 kg	1.6 kg	1 kg
III	2.0 kg	1.25 kg	0.8 kg

Solution : In this example, prices are given in quantity per rupee'. Thus, before computing price relatives, these are to be converted into 'rupees per unit of quantity'. Considering the unit of the quantity as 'quintal' the prices in 'rupees per quintal' are given below -

Year Price (in Rupees per quintal)

	A	B	C
I	25	50.0	100
II	40	62.6	100
III	50	80.0	125

Now, the price index numbers for II and III years (I years as base) can be constructed as usual. The procddure is clarified in the following table-

Construction of price Relatives

Article	Price (Rs.) I	II	II	Price Relatives for II year (1 year as base) R1=P1/Po X 100	Price Relatives for III Year (I year as base) R2=P1/P0 X 100
A	25	40.0	50.0	160.0	200.0
B	50	62.5	80.0	125.0	160.0
C	100	100.0	125.0	100.0	125.0
Total N =3				S R1 = 385.0	485.0

Price Index for II year = $P01 = \frac{\Sigma P1}{P0} X100 = \frac{\Sigma R1}{N} = \frac{385.0}{3} = 128.33$

Price Index for III year = $P02 = \frac{\Sigma P2}{Po} X 100 = \frac{\Sigma R2}{N} = \frac{485.0}{3} = 161.67$

Merits and Demerits of Simple Averages of Relative Method

Merits

1. The Index number is not influenced by extreme items. All items are given equal weightage.
2. The method provides an index which is not affected by the particular units used in price quotations.

Limitations.

1. In this method, we face with the problem of selecting a suitable average.
2. All commodity price relatives are given equal weightage which may not be true always.

❄ ❄ ❄

Lesson : 8

WEIGHTED INDEX NUMBERS

Author : **Vetter:**

Dr. S. S. Tasak **Dr. R. K. Mittal**

As observed, simple or unweighted index number assign equal importance to all the commodites or items included in the index. However, various commoditis included are not of equal importance. To overcome this disadvantage of the simple index numbers, we weight the price of each commodity by a suitable factor. This factor is generally taken as the quantity or volume of the commodity sold or consumed during the base year, the current year or some typical year (taken as an average over a number of years). In this way, the importance of the commodities as also reflected in the index number. As indicated earlier, weighted index number too can be classified as :

1. Weighted aggregate index
2. Weight Average of Relatives

Weighted Aggregate Index Numbers

In this method, appropriate weights are assigned to various commodities which reflect their relative importance in the group. A reasonable assumption is to consider the quantities consumed or produced as weights. If W is the weight attched to commodity then a general weighted price index can be formulated as under :-

$$\textbf{Weighted aggregate price Index} = P_{0n} = \frac{\sum P_a W}{\sum P_0 W} \times 100 \qquad \ldots.(1)$$

Where symbols have their usual meanings.

By using different types of weight in (9), a number of formulas have been developed for the construction of index numbers, These formulas are -

1. Laspeyers' index or base year method.
2. Pasche's index or given year method.
3. Marshall - Edgeworth's index number.
4. Walsh's index number.
5. Bowley's index number.
6. Fisher's ideal index number.

Laspeyres' Index Number

Taking the quantity of the base year i.e. q_0 as weight in formula (1), we get Laspeyre's price index number as -

$$\textbf{Laspeyres' Price Index} = P_{0n} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 \quad \ldots.(2)$$

Paasche's Index Number

According to this formula, the quantity of the given year, i.e. q_n is taken as weight. So using q_n for W in (9), we gent Paache's index as -

$$\textbf{Paasche's Price Index} = P_{0n} = \frac{\sum P_a q_n}{\sum P_0 q_n} \times 100 \quad \ldots.(3)$$

Marshall-Exgeworth's Index Number

In this formula, the average of the base year and given year quantity is taken as weight. Thus, on putting $w = (p_0 + q_n)/2$ in (1), we gets

$$\textbf{Marshall-Edgeworth's Price Index} = P_{0n} = \frac{\sum P_n \left(\frac{q_0+q_n}{2}\right)}{\sum P_0 \left(\frac{q_0+q_n}{2}\right)} \times 100$$

$$= \mathbf{P_{0n}} = \frac{\sum P_n(q_0+q_n)}{\sum p_0(q_0+q_n)} \times 100 \qquad \ldots.(4)$$

Walsh's Index Number

Taking as weights the geometric mean $\sqrt{[q_0 q_1]}$ of the base and given year quantities, i.e., putting $W = \sqrt{[q_0 q_1]}$ in formula (1), one gets

$$\textbf{Walsh's Price Index} = \mathbf{P_{0n}} = \frac{\sum P_n[\sqrt{(q_0+q_n)}]}{\sum P_0[\sqrt{(q_0+q_n)}]} \times 100 \qquad \ldots.(5)$$

Bowley's Index Number

Taking the grithmetic mean of Laspeyres' and Paache's index numbers in (2) and (3) respectively, one gets

$$\textbf{Bowley's Price Index} = \mathbf{P_{0n}} = \frac{1}{2}\left[\frac{\sum p_n q_0}{\sum p_0 q_0} + \frac{\sum p_n q_0}{\sum p_0 q_n}\right] \times 100 \qquad \ldots.(6)$$

Fisher's Ideal Index Number

We define Fisher's ideal price index as the geometric mean of Laspyre's and Paasche's index numbers in (10) and (11). Therefore,

$$\textbf{Fisher's Price Index} = \mathbf{P_{0n}} = \sqrt{\left[\frac{\sum p_n q_0}{\sum p_0 q_0} + \frac{\sum p_n q_0}{\sum p_0 q_n}\right]} \times 100 \qquad \ldots.(7)$$

As will be discussed later, Fisher's ideal index satisfies both the time reversal test and factor reversal test which provides this index number a theoretical advantage over other formulas. In view of these properties, it is called an index formula.

Use of Laspeyres's and Paasche's Index Numbers

The main advantage of Laspeyre's index formula is that the weight q_0 for

the base year remains the same throughout. Thus, while constructing the index number, only the changes in prices are experienced. On the other hand, for Paasche's index, weights q_n are also obtained for each given year. In case the base year is the typically selected normal year, then the use of baseyear quantities provides more stability to the index number. Obviously, the use of Laspeyres' index number is advantageous when the base year is stable and normal. On the other hand, if the conditions have been changing fast over the years, then considering current year's quantities (q_n) as weights, represents a more realistic picture. As such, Paasche's index is more suitable in this situation.

Example 8 : Construct an index number for 1990 (base year = 1980) from the following data using weighted aggregative index number.

Commodity	Weight	Prices (Rs.)	
		Base year (1980)	Current year (1990)
A	30	4.25	5.20
B	40	2.90	3.75
C	15	2.15	1.95
D	15	8.85	8.10

Solution : **Construction of Index**

Commodity	Weight	Prices (Rs.)		P_0W	P_nW
		Base year (1980) P_0	Current year (1990) P_n		
A	30	4.25	5.20	127.50	156.00
B	40	2.90	3.75	118.00	150.00
C	15	2.15	1.95	32.25	29.25
D	15	8.85	8.10	132.75	121.50
Total	100			$\Sigma P_0 W$ =410.5	$\Sigma P_n W$ 456.73

Weighted Index Number for 1990 = P_{0n} = $\frac{\Sigma P_n W}{\Sigma P_0 W} \times 100 = \frac{456.75}{410.50} \times 100 =$ **111.27**

Example 9 : Construct price index numbers for one year 1990 (base year - 1985) from the following data by (i) Laspeyres' Method (ii) Paashe's method (iii) Marshall-Edgeworth's method (iv) Fisher's method.

Commodity	Base year (1985)		Current year (1990)	
	Price	Quantity	Price	Quantity
A	10	30	12	50
B	8	15	10	25
C	6	20	6	30
D	4	10	6	20

Construction of Price Index Numbers

Commodity	Base year		Current year					
	Price P_0	Quantity q_n	Price P_0	Quantity q_n	P_0q_n	P_nq_0	P_0q_n	P_nq_0
A	10	30	12	50	300	500	360	600
B	8	15	10	25	120	200	150	250
C	6	20	6	30	120	180	120	180
D	4	10	6	20	40	80	60	120
Total					Σp_0q_0 =580	Σp_0q_n =960	Σp_nq_0 =690	Σp_n =1150

(i) Laspeyre's Method - Using formula (2),

Price Index = P_{0n} = $\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times 100 = \frac{690}{580} \times 100 =$ **118.96**

(ii) Paasche's Method - Using formula (3),

$$\textbf{Price Index} = P_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times 100 = \frac{1150}{960} \times 100 = \textbf{119.79}$$

(iii) Bowley's Method - Using formula (6),

$$\textbf{Price Index} = P_{0n} = \frac{1}{2}\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right] \times 100$$

$$= \frac{1}{2}[(118.96 + 119.79)] \times 100 = 119.37.$$

(iv) Marshall Edgewo - Method Using formula (4),

$$\textbf{Price Index} = P_{0n} = \frac{\Sigma p_n (q_0 + q_n)}{\Sigma p_0 (q_0 + q_n)} \times 100 = \frac{\Sigma p_n q_0 + \Sigma p_n q_n}{\Sigma p_0 q_0 + \Sigma p_0 q_n} \times 100$$

$$= \frac{690+1150}{580+960} \times 100 = \textbf{119.48}$$

(v) Fisher's Method - Using formula (7),

$$\textbf{Price Index} = P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_0}{\Sigma p_0 q_n}\right]} \times 100$$

$$= \frac{690}{580} \times \frac{1150}{960} \times 100 = \sqrt{[1.189 \times 1.197]} \times 100$$

$$= \textbf{119.29.}$$

Example 10 : From the following data calculate price index numbers for the year 1980 with 1970 as the base year by using (i) Laspeyre's method (ii) Paashe's method (iii) Marshall Edgeworth method (iv) Fisher's method.

Commodity	1970		1980	
	Price(P_0)	Quantity(q_0)	Price(P_n)	Quantity(q_n)
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

Solution :

Construction of Price Index Numbers

Commodity	1970		1980					
	P_0	q_n	P_n	q_n	P_0q_n	P_nq_0	P_0q_n	P_nq_0
A	20	8	40	6	160	120	320	240
B	50	10	60	5	500	250	600	300
C	40	15	50	15	600	600	750	750
D	20	20	20	25	400	500	400	500
Total					Σp_0q_0 =1660	Σp_0q_n =1170	Σp_nq_0 =2070	Σp_nq_0 =1790

(i) On using formula (1), Laspoyre's price index will be :

$$= P_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times 100 \qquad = \frac{2070}{1660} \times 100 = \mathbf{124.7}$$

(ii) On using formula (2), Pasche's price index becomes

$$= P_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times 100 \qquad = \frac{1790}{1470} \times 100 = \mathbf{121.77}$$

(iii) Onusing formula (4), Marshall-Edgeworth price index can be computed as

$$= P_{0n} = \frac{\Sigma p_n (q_0+q_n)}{\Sigma p_0 (q_0+q_n)} \times 100 \qquad = \frac{\Sigma p_n q_0 + \Sigma p_n q_0}{\Sigma p_0 q_0 + \Sigma p_0 q_n} \times 100$$

$$= \frac{2070+1790}{1660+1470} \times 100 \quad = \frac{3860}{3130} \times 100 = \mathbf{123.32}$$

(iv) On using formula (7), Fisher's price index is

$$P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_n 0 q_n}\right]} \times 100 = \sqrt{[1.247 \times 1.2177]} \times 100$$

$$= \mathbf{123.33.}$$

Weighted Average of Relative Method

As discussed earlier, weighted average of relative method is used to overcome the limitations of the simple average of relatives. Weighted arithmetic mean is the most common weighted average for computing index number, although weighted geometric mean can also be used.

The general formula for weighted average of relatives can be written as -

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0} \times 100\right) W}{\Sigma W} \times 100 = \frac{\Sigma RW}{\Sigma W} \quad \ldots. (8)$$

$$\text{Where } R = \left(\frac{P_n}{P_0} \times 100\right) \quad \ldots. (9)$$

Similarly, the index number based on geometric mean of price relatives is

$$P_{on} = \text{Antilog}\left(\frac{\Sigma W \log R}{\Sigma W}\right)$$

In this method, however each price relative is generally weighted by the total of the commodity in terms of some monetary value, say in Rs. etc. The value of the commodity is obtained by multiplying the price p of the commdity by the quantity q. Thus, the weight are given by the value, i.e., W = pq. Depending on whether the base year values p_0q_0 are taken as weights, the price index formulae obtained by using weighted average of relatives are -

1. Weight arithmetic mean of price relatives using base year value weight is :

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right)(p0q0)}{\Sigma p0q0} \times 100 = \frac{\Sigma P_n q_0}{\Sigma P_0 q_0} \quad \ldots (10)$$

2. Weighted arithmetic mean of price relative using current year value weights is -

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right)(pnqn)}{\Sigma pnqn} \times 100 \quad \ldots (11)$$

$$\text{or } P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right) W}{\Sigma W} \times 100$$

where $W = (p_n q_n)$ = Value in the given current year.

Remark. Here it is important to see that formulae in equation (11) is the same as Laspeyre's formulae given in equation (2).

Example 11 : Construct an index number for the following data using weighted average of price relative method.

Commudity	Base year Price (Rs.)(P.)	Current year Price(Rs.)(P.)	Weights
A	42.5	52.0	30
B	29.5	37.5	40
C	21.5	19.5	15
D	88.5	81.0	15

Solution : Computing weighted average of price relative index.

Commudity	Base year Price (Rs.)P.	Current year Price(Rs.)P.	Weights $R=p_n/p_0$ X 100	Weights W	RW
A	42.5	52.0	122.4	30	3672.0
B	29.5	37.5	127.1	40	5084.0
C	21.5	19.5	90.7	15	1360.5
D	88.5	81.0	91.5	15	1372.5
				$\Sigma W=100$	ΣRW-11489.0

Thus, the weighted average of price relative index

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right) W}{\Sigma W} \times 100 = \frac{\Sigma RW}{\Sigma W} = \frac{11489.0}{100}$$

= 114.89

Example 12 : An enquiry into the family budget of middle class family gave the following information -

Item	Food	Rent	Clothing	Fuel	Others
Expenditure % :	30%	15%	20%	10%	25%
Price (Rs.) in 1985 :	100	20	70	20	40
Price (Rs.) in 1986	90	20	60	15	35

Compute the price index for 1986 by using (1) weighted arithmetic mean of price relatives (ii) weighted geometric mean of price relatives.

Solution : Computation of weighted index

Item	Weight W	P0	Pn	Price Relative $R=P_n/P_0X100$	WR	Log R	W log R
Food	30	100	90	90.0	2700.00	1.9542	58.626
Rent	15	20	20	100.0	1500.00	2.0000	30.000

Clothing	20	70	60	85.7	1714.00	1.9330	38.660
Fuel	10	20	15	75.0	750.0	1.8751	18.751
Others	25	40	55	137.5	3437.50	2.1383	53.457
	$\Sigma W=$ 100				$\Sigma WR=$ 10101.50		$\Sigma W \log R=$ 199.494

(i) Index based on weighted A.M. of relatives :

$$P_{0n} = \frac{\Sigma RW}{\Sigma W} = \frac{10101.5}{100} = \mathbf{102.02}$$

(i) Index based on weighted G.M. of relatives :

$$P_{0n} = \text{antilog}\left(\frac{\Sigma W \log R}{\Sigma W}\right) = \text{antilog}\left(\frac{199.494}{100}\right) = \text{antilog}\,[1.99494]$$

$$= 98.33$$

Example 13 : From the information on the next page construct index number of 1990 by the method of weighted average of relavites using (i) base year value as weights (ii) given year values weights.

Article	1985		1990	
	Price(P_0)	Quantity(q_0)	Price(P_n)	Quantity(q_n)
A	8	50	20	40
B	6	10	18	2
C	4	5	5	2

Solution : Constructionof Price Index (weighted average relative method).

Article	1985 P_0	1985 q_0	1990 P_n	1990 q_n	$V_0=P_0q_0$	$V_n=P_nq_n$	$R=P_n/P_0$	RV_0	RV_n
A	8	50	20	40	400	800	2.5	1000.0	2000.0
B	6	10	18	2	60	36	3.0	180.0	108.0
C	4	5	5	2	20	16	2.0	40.0	32.0
Total					ΣP_0q_0 =480	ΣP_0q_n =852		ΣRV_0 =2070	ΣRV_n =1790

(i) Price index when base year value weights are used:

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right)(p0q0)}{\Sigma p0q0} \times 100 = \frac{\Sigma RV_0}{\Sigma V_0} \times 100 = \frac{1220.0}{480.0} \times 100$$

$= 473.45.$

(ii) Price index when given year value weights are used:

$$P_{0n} = \frac{\Sigma\left(\frac{P_n}{P_0}\right)(pnqn)}{\Sigma pnqn} \times 100 = \frac{\Sigma RV_n}{\Sigma V_n} \times 100 = \frac{\Sigma RV_n}{\Sigma V_n} \times 100$$

$= 473.45.$

Quantity or Volume Index numbers

Instead of comparing price changes over a period of time, we may be interested in analysing changes in quantity of production or consumption of certain commodities over a given period of time. In order to study such changes, we construct quantity index numbers. In this regard, a few quantity index numbers (Q_{0n}) formula may be listed as under.

(i) Simple aggregative quantity index :

$$Q_{0n} = \frac{\Sigma q_n}{\Sigma q_0} \times 100 \quad \ldots (12)$$

(ii) Simple average of relatives method :

$$Q_{0n} = \frac{\Sigma (q_n/q0)}{N} \times 100 \quad \ldots (13)$$

(iii) Laspeyres' Quantity index : Using base year prices as weights, Laspeyres' quantity index is given by

$$Q_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times 100 \quad \ldots (14)$$

(iv) Paasche's quantity index : Using given year prices as weights, Paasche's quantity index can be formulate as

$$Q_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times 100 \quad \ldots (15)$$

Flasher's and Marshall-Edgeworth's formula for quantity index numbers can also be obtained on the similar pattern.

While the price index number measure the change in the value of a fixed aggregate of goods at varying prices, the quantity index number measures the change in value of a varying aggregate of goods at fixed prices. Thus, the price index number enable us to know the amount of expenditure in a given year if the same volume of quantity is consumed at varying prices. On the other hand, the quantity index number tells us how much we shall spend in the given year if varying quantities of commodities are bought at the same price.

Example 14 : Compute (i) Laspeyres' (ii) Paasche's and (iii) Fisher's quantity index numbers from the following data :

Article	1982		1984	
	Price(P_0)	Quantity(q_0)	Price(P_n)	Quantity(q_n)
A	5	10	4	12
B	8	6	7	7
C	6	3	5	4

Solution : Computation of Price Index Numbers

Commodity	1982		1984					
	P_0	q_n	P_n	q_n	P_0q_n	P_nq_0	P_0q_n	P_nq_0
A	5	10	4	12	50	60	40	48
B	8	6	7	7	48	56	42	49
C	6	3	5	4	18	24	15	20
Total					Σp_0q_0 =116	Σp_0q_n =140	Σp_nq_0 =97	Σp_nq_n =117

Using (14), Laspeyres' quantity index is -

$$= Q_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times 100 = \frac{97}{116} \times 100 = \mathbf{83.62}$$

Using (23), Paasche's quantity index is -

$$= Q_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times 100 = \frac{117}{140} \times 100 = \mathbf{83.57}$$

Finally, Fiasher's Quantity Index will be

$$Q_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]} \times 100$$

$$= \sqrt{\left[\frac{97}{116} \times \frac{117}{140}\right]} \times 100 = \sqrt{\left[\frac{11349}{13697}\right]} \times 100 = 91.02$$

Example 15 : Use the following data to find Fisher's price and quantity indices-

Article	Base year		Current year	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

Solution : Computation of Fisher's price and quantities indices.

Article	Base year		Current year					
	P_0	q_n	P_n	q_n	P_0q_n	P_nq_0	P_0q_n	P_nq_0
A	6	50	10	56	300	336	500	560
B	2	100	2	120	200	240	200	240
C	4	60	6	60	240	240	360	360
D	10	30	12	24	300	240	360	288
E	8	40	12	36	320	288	480	432
Total					Σp_0q_0 =1360	Σp_0q_n =1344	Σp_nq_0 =1900	Σp_nq_n =1880

Fisher's Price Index $P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} \times 100$

$$= \sqrt{\left[\frac{1900}{1360} \times \frac{1880}{1344}\right]} \times 100$$

$$= \sqrt{[1.95219]} \times 100 = \mathbf{139.79}$$

Fisher's Quantity Index $Q_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]} \times 100$

$$= \sqrt{\left[\frac{1900}{1360} \times \frac{1880}{1344}\right]} \times 100$$

$$= \sqrt{[0.9778328]} \times 100 = 0.9889 \times 100 = \mathbf{98.89}$$

Tests of Adequacy of Index Numbers

Earlier, while discussing the relatives, we observed that relatives follow certain properties. When these properties are true for an individual commodity, these should also hold good for a group of commodites. As such, the index number as an aggregative relative should also satisfy these properties. In view of this, we now discuss these properties in the light of various index number formulae. A good index number should satisfy the following tests or properties.

1. Unit Test
2. Time Reversal Test
3. Factor Reversal Test
4. Circular Test

Unit Test

According to unit-test, an index number should be independent of the unit in which prices and quantities of various commodities and quoted. This test is satisfied by all the index number formulae except the simple aggregative index.

Time Reversal Test

According to Prof. Fisher "the formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as base." In other words, if the two periods, the base and the reference period are interchanged, the product of the two index numbers should be unity, i.e, they should be reciprocal of each other. Symbolically,

$$P_{0n} \text{ X } P_{n0} = 1 \text{ or } \qquad P_{0n} = 1/P_{n0} \qquad \ldots (16)$$

An index formula satisfying the criteria in equation (16) is said to satisfy the time reversal test.

Let us see whether Laspeyres' index number satisfy this property or not. For which, Laspeyres' formula is -

$$= \mathbf{P_{0n}} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{x } 100$$

$$= \mathbf{P_{n0}} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \text{x } 100$$

$$\textbf{Now} \quad = \mathbf{P_{0n} X P_{no}} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{x } 100 \text{ x } \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \text{x } 100 = 1.$$

So Laspeyres' formula does not satisfy the time reversal test. Similarly, Paasche's index formula does not satisfy this test. However, if we consider Fisher's index

$$\mathbf{P_{0n}} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{ X } \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]}$$

$$\textbf{and thus} \quad \mathbf{P_{0n}} = \sqrt{\left[\frac{\Sigma p_o q_n}{\Sigma p_n q_n} \text{ X } \frac{\Sigma p_0 q_0}{\Sigma p_n q_0}\right]}$$

$$\text{Now } P_{on} \times P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]} \times \sqrt{\left[\frac{\Sigma p_0 q_n}{\Sigma p_n q_n} \times \frac{\Sigma p_0 q_0}{\Sigma p_n q_0}\right]}$$

thus, Fisher's Index satisfies time reversal test.

Similarly, students can see that **Marshall-Edgeworth's** and **Walsh's** index numbers also satisfy time reversal tests.

Factor Reversal Test

We known that if the two factors p and q are inter-changed in a price index fromula (P_{0n}), we get quantity index formula (Q_{0n}). Then we expect that the product of P_{0n} and Q_{0n} should be equal to the true value ratio.

index should give the true ratio of value in the given year (n) to the value in the base year (0). Symbolically, the factor reversal test is satisfied if

$$P_{0n} \times Q_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} = \frac{\text{Value in the given year (n)}}{\text{Value in the base year (o)}} \qquad \ldots(17)$$

Now let us consider the case of Laspeyres' Index for which

$$P_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \quad \text{and} \quad Q_{0n} = \frac{\Sigma q_n p_0}{\Sigma q_0 p_0}$$

$$\text{Thus, } = P_{0n} \times Q_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma q_n p_0}{\Sigma q_0 p_0} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

Thus, **Laspeyre's index** does not satisfy factor reversal test. However, in the case of Fisher's index, we have,

$$P_{on} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]} \text{ and } Q_{0n} \sqrt{\left[\frac{\Sigma q_n p_n}{\Sigma q_0 p_0} \times \frac{\Sigma q_n p_n}{\Sigma q_n p_n}\right]}$$

$$P_{on} \text{ X } Q_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{ X } \frac{\Sigma p_n q_n}{\Sigma p_n q_n}\right]} \text{ X } \sqrt{\left[\frac{\Sigma q_0 p_0}{\Sigma q_0 p_0} \text{ X } \frac{\Sigma q_n p_n}{\Sigma q_n q_n}\right]}$$

$$= \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}$$

Thus, **Fisher's index also satisfies the factor reversal test.** Now since Fisher's index number is one which satisfies both the time reversal and factor reversal test, so it is called **Fisher's ideal index number.**

Circular Test :

This is another test for the adequacy of the index number. This is based on the shifting of the base period and thus is an extension on the time reversal test. According to this test, an index number should also work in a circular way, i.e. the test requres that

$$P_{01} \text{ x } P_{12} \text{ x } P_{20} = 1 \qquad \ldots(18)$$

For, four time points, we have

$$P_{01} \text{ x } P_{12} \text{ x } P_{23} \text{ x } P_{30} = 1$$

In particular, for two time points, 0 and n, one gets

$$P_{0n} \text{ x } P_{n0} = 1$$

which is nothing but time reversal test discussed above. It can be observed that none ..... weighted index numbers satisfies this test. However, for three 0, 1 and 2, if the test is applied to index obtained by simple aggregate method, one gets -

$$P_{01} X P_{12} X P_{20} = \frac{\Sigma p_1}{\Sigma p_0} x \frac{\Sigma p_2}{\Sigma p_1} x \frac{\Sigma p_0}{\Sigma p_2} = 1$$

Similarly, if we apply circular test to index numbers obtained by fixed weight aggregative method, we get -

$$P_{01} X P_{12} X P_{20} = \frac{\Sigma p_1 q}{\Sigma p_0 q} x \frac{\Sigma p_2 q}{\Sigma p_1 q} x \frac{\Sigma p_0 q}{\Sigma p_2 q} = 1.$$

Thus, for index obtained by simple aggregate method or by fixed weight aggregative method, the circular test is satisfied.

Example 16. Calculate Laspeyre's and Paasche's price indices for the year 1980 from the following data. Prove that both the formulae do not satisfy the Time Reversal Test.

Commodity	Price (Rs.)		Quantities (kgs)	
	1979	1980	1979	1980
A	2.0	2.50	3	5
B	2.5	3.00	4	6
C	3.0	2.50	2	3
D	1.0	0.75	1	2

Solution : Computation of Laspeyre's and Paasche's indices

Article	Base year		Current year					
	P_0	q_n	P_n	q_n	P_0q_n	P_nq_0	P_0q_n	P_nq_n
A	2.0	3	2.50	5	6.0	7.5	10.0	12.5
B	2.5	4	3.00	6	10.0	12.00	15.0	18.0
C	3.0	2	2.50	3	6.0	5.00	9.0	7.5
D	1.0	1	0.75	2	1.0	0.75	2.0	1.5
Total					Σp_0q_0 =23.0	Σp_nq_n =25.25	Σp_0q_n =36.0	Σp_nq_n =39.5

Laspeyrs' index = P_{0n} $= \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} x\ 100 = \frac{25.25}{23.00} x\ 100 = 109.78$

Paasche's index = P_{0n} $= \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} x\ 100 = \frac{39.50}{36.00} x\ 100 = 109.72$

Time Reversal Test : This test is satisfied if $P_{0n} \times P_{n0} = 1$

$$P_{0n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_n} \text{ and } P_{n0} = \frac{\Sigma p_n q_0}{\Sigma p_n q_0}$$

$$\therefore \quad P_{0n} \times P_{n0} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times \frac{\Sigma p_0 q_0}{\Sigma p_n q_0} = \frac{39.50}{36.00} \times \frac{23.00}{25.25} = 1$$

Thus, Paasche's indedx also does not satisfy the Time Reversal Test.

Example 17. The following table gives the prices and quantities of 5 commodities in the base and current year. Use it to verify whether Fisher's ideal index satisfies the time reversal test.

Commodity	Base year		Current year	
	Unit price (Rs.)	Quantity (kgs.)	Unit price (Rs.)	Quantity (kgs.)
A	5	50	5	70
B	5	75	10	80
C	10	80	12	100
D	5	20	8	30
E	10	50	5	60

Solution : Computation Fisher's Ideal Index

Commodity	Base year		Current year					
	P_0	q_0	P_n	q_n	P_0q_0	P_nq_n	P_nq_n	P_nq_n
A	5	50	5	70	250	350	250	350
B	5	75	10	80	375	400	750	800
C	10	80	12	100	800	1000	960	1200
D	5	20	8	30	100	150	160	240
E	10	50	5	60	500	600	250	300
Total					Σp_0q_0 =2025	Σp_nq_n =2500	Σp_0q_n =2370	Σp_nq_n =2890

$$\text{Fisher's ideal index } = P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]}$$

$$= \sqrt{\left[\frac{2370}{2025} \text{ X } \frac{2890}{2500}\right]} \text{ x } 100$$

$= \sqrt{[1.352948]} \text{X } 100 = 1.1632 \text{ X } 100 =$ **116.32**

For time reversal test -

$$P_{on} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{ X } \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} = \sqrt{\left[\frac{2370}{2025} \text{ X } \frac{2890}{2500}\right]}$$

$$P_{on} = \sqrt{\left[\frac{\Sigma p_0 q_n}{\Sigma p_n q_n} \text{ X } \frac{\Sigma p_0 q_0}{\Sigma p_n q_0}\right]} = \sqrt{\left[\frac{2500}{2890} \text{ X } \frac{2025}{2370}\right]}$$

$$P_{on} \text{ X } P_{no} = \sqrt{\left[\frac{2370}{2025} \text{ X } \frac{2890}{2500}\right]} = \sqrt{\left[\frac{2500}{2890} \text{ X } \frac{2025}{2370}\right]} = 1.0$$

Thus, Fisher's Index satisfies the time reversal test.

Example 18. In the above example, verify that Fisher's index also satisfies the factor reversal test.

Solution : The factor reversal test is satisfied if

$$P_{0n} \text{ X } Q_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_o}$$

Let us consider -

$$P_{on} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \text{ X } \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} = \sqrt{\left[\frac{2370}{2025} \text{ X } \frac{2890}{2500}\right]}$$

$$Q_{on} = \sqrt{\left[\frac{\Sigma q_n p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}\right]} = \sqrt{\left[\frac{2500}{2025} \times \frac{2890}{2370}\right]}$$

$$\therefore P_{on} \times Q_{on} = \sqrt{\left[\frac{2370}{2025} \times \frac{2890}{2500}\right]} = \sqrt{\left[\frac{2500}{2025} \times \frac{2890}{2370}\right]}$$

$$= \frac{2890}{2025} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_o}$$

Thua, Fisher's index also satisfies the factor reversal test.

Example 19. From the following prove that Fisher's ideal index satisfies both the time reversal and the factor reversal test -

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	6	50	10	60
B	2	100	2	120
C	4	60	6	60

Solution : Computation of Fisher's Ideal Index Number

Commodity	Base year		Current year					
	P_0	q_0	P_n	q_n	P_0q_0	P_0q_n	P_nq_0	P_nq_n
A	6	50	10	60	300	360	500	600
B	2	100	2	120	200	240	200	240
C	4	60	6	60	240	240	360	360
Total					Σp_0q_0 =700	Σp_0q_n =840	Σp_nq_0 =1000	Σp_nq_n =1200

Fisher's index (Price) = $P_{0n} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} \times 100$

$$= \sqrt{\left[\frac{1000}{700} \times \frac{1200}{840}\right]} \times 100 = \mathbf{142.86}$$

Time Reversal Test. This test is satisfied if $P_{0n} \times P_{n0} = 1$.

Here $P_{on} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} = \sqrt{\left[\frac{1000}{700} \times \frac{1200}{840}\right]}$

and $P_{n0} = \sqrt{\left[\frac{\Sigma p_0 q_n}{\Sigma p_n q_n} \times \frac{\Sigma p_0 q_0}{\Sigma p_n q_0}\right]} = \sqrt{\left[\frac{700}{1000} \times \frac{840}{1200}\right]}$

$\therefore P_{on} \times P_{no} = \sqrt{\left[\frac{1000}{700} \times \frac{1200}{840}\right]} \times \sqrt{\left[\frac{700}{1000} \times \frac{840}{1200}\right]} = 1.00$

Thus, Fisher's Index satisfies the **Time Reversal Test.**

Factor Reversal Test : The test is satisfied if - $P_{on} \times Q_{on} = \frac{\Sigma P_n Q_n}{P_o Q_o}$

Here $P_{on} = \sqrt{\left[\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right]} = \sqrt{\left[\frac{\Sigma 1000}{700} \times \frac{1200}{840}\right]}$

and $Q_{0n} = \sqrt{\left[\frac{\Sigma q_n p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}\right]} = \sqrt{\left[\frac{840}{700} \times \frac{1200}{1000}\right]}$

$\therefore P_{on} \times Q_{on} = \sqrt{\left[\frac{1000}{700} \times \frac{1200}{840}\right]} \times \sqrt{\left[\frac{840}{700} \times \frac{1200}{1000}\right]}$

$$= \frac{1200}{700} = \frac{\sum p_n q_n}{\sum p_0 q_o}$$

Thus, Fisher's index also satisfies **Factor Reversal Test.**

Link and Chain Relatives

Let $p_1, p_2, p_3, \ldots$ denote prices of a commodity during successive intervals of time 1, 2, 3, ... respectively then the price relatives of each time interval with respect to the preceding time interval as base, i.e., $p_{12}, p_{23}, p_{34}, \ldots$, etc., are called **link relatives.** Now using this series of price relatives and circular property of relatives, one can get the price relative for a given period with respect to any other period as base. To put it other way, let us consider the modified circular property of price relatives which states

$$p_{13} = p_{12}, p_{23} \quad \ldots(19)$$

or $$p_{14} = p_{12}, p_{23}, p_{34} \quad \ldots(20)$$

That is, on using (19), the price relative of period 3 with respect to period 1 as base can be obtained from the link relative p_{12} and p_{23}. Similarly, link relatives p_{12}, p_{23}, and p_{34} can be used to get p_{14}, i.e., the price relative of period 4 with respect to period 1as base.

Thus, the price relatives with respect to a fixed base period can be obtained by using link relatives. These link relavies of one period with respect to preceding as base are chained together by successive multiplication to get a chain price relative of a period with respect to a fixed base period. In view of this, it is called a chain relative or chain index. The concept of link and chain relatives, can be similarly used in the case of quantity and value relatives.

Illustration 20. If the prices of a commodity during 1984, 1985, 1986 and 1987 are Rs. 30, 35, 38 and 42 respectively, then the price link relatives can

be obtained as shown in the table -

Year	1984(1)	1985(2)	1986(3)	1987(4)
Price(Rs.)	$30(p_1)$	$35(p_2)$	$38(p_3)$	$42(p_4)$
Link relatives		$P_{12}=\frac{P_2}{P_1}=\frac{35}{30}$	$P_{23}=\frac{P_3}{P_2}=\frac{38}{35}$	$P_{34}=\frac{P_4}{P_3}=\frac{42}{38}$

Now on using these link relative and their circular property, the price relative with respect to a fixed base can be obtained. For example, the price relative for the year 1986 with base year 1984 can be enumerated as

$$P_{13} = P_{12} \text{ X } P_{23} = \frac{35}{30} \quad \frac{38}{35} = \frac{38}{30} = \mathbf{126.67\%}$$

Similarly, the price relative of 1987 with base year as 1984 becomes.

$$P_{14} = P_{12} \text{ X } P_{23} \text{X } P_{34} = \frac{35}{30} \text{X} \frac{38}{35} \text{X} \frac{42}{38} = \frac{42}{30} = \mathbf{140\%}$$

Thus, the construction of chain indices involves the following steps-

1. Express the figures for each period as percentage of the preceding period to get the link relatives (L.R.)

2. These link relatives are chained together by successive multiplication to get the chain indices or relatives of a period with respect to a fixed base period Symbolically, the chain relatives or indicies are computed as

P_{01} = First link relative

$P_{02} = P_{01} \text{ X } P_{12}$

$P_{03} = P_{01} \text{ X } P_{12} \text{ X } P_{23} = P_{02} \text{ X } P_{23}$

$P_{01} = P_{01} \text{ X } P_{12} \text{ X } P_{23} \text{ X } P_{34} = P_{23}$

$P_{0k} = P_{01} \text{ X } P_{12} \text{ X } \ldots \text{ X } P_{(k-1)} = P_{0(k-1)} \text{ X } P_{(k-1)k}$

or use the following formula formula for computing chain index (C.I.)

$$C.I. = \frac{\text{Current year L.R. X Preceeding C.I.}}{100} \quad ...(21)$$

It is notable here that the meanings of the terms fixed based index (F.B.I), link relative or index (L.R.). chain index (C.I.)and chain base index (C.B.I.), should be clearly understood. In these terms link relative (L.R.) and chain base index (C.B.I.) convey the same meaning. While, for an index series, chain indies are the same as the fixed base index.

Further, in some specific situations, we also need to convert chain base index numbers (C.B.I.) to fixed base index (F.B.I.) for which the following formula is applicable -

$$\text{Current year F.B.I.} = \frac{\text{Current year C.B.I. X Previous year F.B.I.}}{100} \quad ...(22)$$

The computation procedure will be more clear from the following examples-

Example 20. From the following data of wholesale prices of a certain commodity, construct (i) Fixed Base Index (1979 = 100) and (ii) Chain index numbers.

Year :	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Price :	75	50	65	60	72	70	69	75	84	80

Solution : Using the formulae-

(i) **Fixed Base Index of a given year =** $\frac{\text{Price of the given year}}{\text{Price of the base year}} X\ 100$

(i) **Chain Index** = $\frac{\text{Link relative of the given year X Chain index of the preceeding year}}{100}$

Using formula (i) and (ii), the fixed base and chain indices are given in the following table -

Year	Price	Fixed Base Index (1979=base)	Link Relatives (L.R.)	Chain Index (C.L.)
1979	75	100	100	100
1980	50	$\frac{50}{75}$ X 100=66.67	$\frac{50}{75}$ X 100=66.67	$\frac{66.67 \text{ X } 100}{100}$ =66.67
1981	65	$\frac{65}{75}$ X 100=86.67	$\frac{65}{50}$ X 100=130.00	$\frac{130 \text{ X } 66.67}{100}$ =86.67
1982	60	$\frac{60}{75}$ X 100=80.00	$\frac{60}{65}$ X 100=66.67	$\frac{66.67 \text{ X } 86.67}{100}$ =80.00
1983	72	$\frac{72}{75}$ X 100=96.00	$\frac{72}{60}$ X 100=120	$\frac{120 \text{ X } 80}{100}$ =96.00
1984	70	$\frac{70}{75}$ X 100=93.33	$\frac{70}{72}$ X 100=97.22	$\frac{97.22 \text{ X } 96}{100}$ =93.33
1985	69	$\frac{69}{75}$ X 100=92.00	$\frac{69}{70}$ X 100=98.57	$\frac{98.57 \text{ X } 93.33}{100}$ =92.00

1986	75	$\frac{75}{75} \times 100 = 100.00$	$\frac{75}{69} \times 100 = 108.69$	$\frac{108.69 \times 92}{100} = 100$
1987	84	$\frac{84}{75} \times 100 = 112.00$	$\frac{84}{75} \times 100 = 112.00$	$\frac{112 \times 100}{100} = 112$
1988	80	$\frac{80}{75} \times 100 = 106.67$	$\frac{80}{84} \times 100 = 95.24$	$\frac{95.24 \times 112}{100} = 106.67$

Remark : It may be noted that the chain indices are the same as the fixed base indx numbers.

Example 21. Use the following chain base index numbers (C.B.I.) to obtain the fixed base index numbers.

Year	:	1978	1979	1980	1981	1982	1983
Chain Base Index	:	105	75	71	105	95	90

Solution : Using formula (22), the required incides are shown in table.

Computation of fixed bas index numers

Year	Chain base Index (C.B.I.)	Fixed base Index number (F.B.I.)
1978	105	= 105.5
1979	75	$\frac{75 \times 105}{100} \times 78.75$
1980	71	$\frac{71 \times 78.75}{100} \times 55.91$

1981	105	$\frac{105 \times 55.91}{100} \times 58.71$
1982	95	$\frac{95 \times 58.71}{100} \times 55.77$
1983	90	$\frac{90 \times 55.77}{100} \times 50.20$

Example 22. Compute chain Index numbers with 1986 prices as base from the following table giving the average wholesale price of the commodities A,B,C for years 1987 to 1990.

Commodities	1986	1987	1988	1989	1990
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

Solution : Computation of Chain indicies

Commodity	Price relative based on precending year				
	1986	**1987**	**1988**	**1989**	**1990**
A	100	$\frac{16}{20} \times 100 =$	$\frac{28}{16} \times 100 =$	$\frac{35}{28} \times 100$	$\frac{21}{35} \times 100 =$
B	100	$\frac{30}{25} \times 100 =$	$\frac{24}{30} \times 100 =$	$\frac{36}{24} \times 100$	$\frac{45}{36} \times 100 =$
C	100	$\frac{25}{20} \times 100 =$	$\frac{30}{25} \times 100 =$	$\frac{24}{30} \times 100$	$\frac{30}{24} \times 100 =$

Total of link Relatives	300	325	375	355	310
Average of link relatives	100	108.33	125	118.33	103.33
Chain Relatives	100	$\frac{108.33 \times 100}{100} = 108.33$	$\frac{125 \times 108.33}{100} = 135.41$	$\frac{118.33 \times 135.41}{100} = 160.23$	$\frac{103.33 \times 160.23}{100} = 165.5$

Example 22. From the chain base index given below, perpare fixed base index numbers.

Year :	1981	1982	1983	1984	1985
Index :	110	160	140	200	150

Solution : Conversion from chain base to fixed base index

Year	Chain base index	Conversion	Fixed index
1981	110		110.0
1982	160	(160x110)/100	176.0
1983	140	(140x176)/100	246.4
1984	200	(200x264.4)/100	492.8
1985	150	(150x492.8)/100	739.2

Example 24. From the following fixed base number index (F.B.I.) prepare chain base indes (C.B.I.) :

Year :	1980	1981	1982	1983	1984	1985
(C.B.I.) :	220	250	300	280	350	415

Solution :

Conversion from F.B.I. to C.B.I.

Year	Chain base index	Conversion	Fixed index
1980	220	-	100.00
1981	250	(250x220)/100	113.64
1982	300	(300x250)/100	120.00
1983	280	(280x300)/100	93.33
1984	350	(350x280)/100	125.00
1985	415	(415x350)/100	118.57

Base Shifting, Splicing and Deflating of Index Numbers

Base Shifiting :

In reference to index numbers, base shifting means to prepare a new indes series with a new base period in place of an old one. Thus, a base shifting is nothing but a procedure of recasting an index series by shifiting its base period to some recent or more relevent base period. Base shifting is necessary in the following situations -

1. When the base period of the index series is too old or too distant from the current period.

2. When we wish to compare two or more index series with different base periods then, for making valid comparisions, it is necessary that the given index series be expressed with a common base peirod.

In base shifting procedure, the index number of the new base year is taken as 100 and then the remaining index numbers in the series are expressed as percentage of the index number selected as new base.

Thus, the index series may be recast by using the following formula -

$$\textbf{Recasted Index No. of any year} = \frac{\text{old Index No. of the year}}{\text{Index No. of new base year}} \times 100 \quad \text{.....(23)}$$

Example 25. Assuming 1979 as the base prepare new index numbers from the indices given below :

Year	:	1976	1977	1978	1979	1980
Indices	:	100	110	125	250	300

Base shifting from 1976 to 1978

Year	Indices (Bease 1976)	Base Shifting	New Indices (base 1979)
1976	110	(100 X 100)/250	40
1977	110	(110x100)/250	44
1978	175	(175x100)/250	70
1979	250	(250x100)/250	100
1980	250	(300x100)/250	120

Example 26. Reconstruct the following index series using 1980 as base :

Year	:	1976	1977	1978	1979	1980	1981	1982
Index No.	:	110	130	150	175	180	200	220

Solution :

Base Shifting

Year	Indices No.	Base Shifting	New Indices (base 1980)
1976	110	(110 X 100)/180	61.11
1977	130	(130x100)/180	72.22
1978	150	(150x100)/180	83.33
1979	175	(175x100)/180	97.22
1980	180	(180x100)/180	100.00
1981	200	(200x100)180	111.11
1982	220	(220x100)180	122.22

Splicing

Splicing means combining two or more index series. For retaining continuity in comparison between two or more index series, we splice or combnine them into a new single index series. For clarity, suppose there is an index series' A ' with base period 1970 which discontinued in 1975 and then a new index series. 'B' is prepared with base period 1970 whihc was discontinued in 1975 and then a new index series 'B' is prepared with base period 1975. Then for comparing the two index series, we can splice them into a new continous in dex series in the following manner :-

1. Splice the index series B' to 'A' to obtain a new continuous index series with base 1970. This splicing procedure is known as forward splicing. In fact, in this splicing the base 1975 of in dex series 'B' has to be shifted to base 1970.

2. Splice the index series 'A' to 'B' to get a new continous index series with base 1975. this splicing procedure is known as backward splicing. In other words, in this type of splicing the base period 1970 of index serice 'A' is to be shifted to base 1975. Thus, the procedure in splicing is very much alike in that involved in base shifting.

The formula used for forward splicing is :-

$$\text{Required Index} = \frac{\text{Old index number on the existing base x Index number to be spliced}}{100} \quad \ldots(24)$$

Also the formula used for backward splicing is :-

$$\text{Needed Index} = \frac{100}{\text{Old index number on the existing base x Index number to be spliced}} \quad \ldots(25)$$

The following examples will clarify the procedure.

Example 27. Splice the following two index series, series A forward and the series B backwards.

Year : 1983 1984 1985 1986 1987 1988

Series A : 110 130 150

Series B : 100 110 140 150

Solution : Splicing Two Index Series

Year	Series		Index Numbers spliced forward to Series A	Index number splied backward to series B
	A	B		
1983	100			(100x100)/150=66.67
1984	130			(100x130)/150=86.67
1985	150	100	(150x100)/100=150	(100x150)/150=100
1986		110	(150x110)/100=165	
1987		140	(150x140)/100=210	
1988		150	(150x150)/100=225	

Thus,

Year :	1983	1984	1985	1986	1980	1988
Forward Splice Series :	100	130	150	165	210	225
Backward Splice Series :	66.67	86.67	100	110	140	150

Example 28. We have the following three index series.

I-Series

Year	:	1980(Base)	1981	1982
Index No.	:	100	120	200

II-Series

Year	:	1982(Base)	1983	1984
Index No.	:	100	110	130

III-Series

Year	:	1984	1985	1986
Index No.	:	100	130	140

Obtain a continuous series with the base 1984 by splicing the three series.

Solution : **Splicing Three Series**

Year	Indices			Special Index Series (base 1984)
	I	II	III	
1980	100			(100x76.92)/200=38.10
1981	120			(120x76.92)/200=46.15
1982	200	100		(100x100)/130=76.92
1983		110		(110x110)/130=84.62
1984		130	100	(100x130)/130=100
1986			130	130
1986			140	140

Thus,

Year	**:**	1980	1981	1982	1983	1984	1985	1986
Continous Series base(1984)	**:**	38.10	46.15	76.92	84.62	100	130	140

Deflating :

Deflating means "making allowance for the effect of changing price level." An increase in price level of consumer goods over a period means areduction in the purchasing power of the people. For example if the price of rice rises from Rs. 500 per quintal in 1990 to Rs. 1000 per quintal in 1992, this simply means that in 1992 the person can buy only half the amount of rice in 1992 if he decideds to spent the same amount which he was speding in 1990. In other words, the value of the rupees is 50 paisa in 1992 as compared to that in 1990 i.e. the purchasing power of the money in 1992 is half as compared to that in 1990. Therefore, the purchasing power is given by the reciprocal of the index number and consequently the real income (or wages) is obtained by dividing the nominal income of the period by the corresponding index number and expressing the ratio in percentage. Thus.

$$\text{Real wages} = \frac{\text{Money or normal wages}}{\text{Price Index}} \times 100$$

The real wages is also known as deflated wages (or income). In this way, we observe that, in the deflating procedure, the relevant price index is the deflator and a deflated value, in general, can be computed by useing the formula -

$$\textbf{Deflated value} = \frac{\text{Current Value}}{\text{Deflator}} \times 100$$

$$= \frac{\text{Current Value}}{\text{the relevent index number}} \times 100$$

Example 29: The following table gives the money wages and cost of living index number based on 1979.

Year	: 1979	1980	1981	1982	1983	1984	1985
Forward Splice Series	: 65	70	75	80	90	100	120
Backward Splice Series	: 100	110	120	130	150	160	200

Solution :

$$\text{Real Wages} = \frac{\text{Normal Wages}}{\text{Index No.}} \times 100$$

Computation of Real Wages

Year	Wages (Rs.)	Index (1979=100)	Deflected wages or Real Wages [Col(2)/Col.(3)]/100
(1)	(2)	(3)	(4)
1979	65	100	(65x100)/100 = 65
1980	70	110	(70x100)/110 = 63.64
1981	75	120	(75x100)/120 = 62.50
1982	80	130	(80x100)/130 = 61.54
1983	90	150	(90x100)/150 = 60.00
1984	100	160	(120x100)/160 = 62.50
1985	120	200	(120x100)/120 = 60.00

Example 30. The following table gives the monthly average salary of a teacher and general price indices for a period of six years.

Year	:	1980	1981	1982	1983	1984	1985
Income (Rs.)	:	360	420	500	550	600	640
General Price Index	:	100	104	115	160	280	290

Find real average salary and construct real wage index based on 1980.

Solution :

Computation of Real Income and Real Income Indices

Year	Income (Rs.)	Price Index	Real Income	Real Income Index
1980	360	100	(360x100)/100 = 360.00	100
1981	420	104	(420x104)/100 = 403.80	112.2
1982	500	115	(500x115)/160 = 434.80	120.8
1983	550	160	(420x100)/160=343.80	95.5
1984	600	280	(420x100)/280=214.30	59.5
1985	640	290	(420x100)/290=220.70	61.3

Example 31. Compute the index of real wages from the following data using 1983 as base year.

Year	:	1980	1981	1982	1983	1984	1985
Average Monthly Wages (Rs.)	:	120	132	143	150	171	200
Price Index	:	100	120	130	150	190	200

Solution : Computation of Real wage indices (Base 1983)

Year	Wages (Rs.)	Price Index	Real Wages	Real wages Index
				(Base = 1983)
1980	120	100	(120x100)/100 = 120	120
1981	132	120	(130x104)/120 = 110	110
1982	143	130	(143x115)/130 = 110	110
1983	150	150	(150x100)/150=100	100
1984	171	190	(171x100)/190= 90	90
1985	200	200	(200x100)/200= 100	100

Limitation of Index Numbers

Although index numbers are found to be very useful for measuring relative change in some phenomenon, they are not without limitation. These limitations are :

1. Index numbres in easure only approximate relative changes in two periods. They are capable of measuring changes in characteristics which can be quantified and vary with time.

2. Index numbers do not use complete data as only a limited number of representative items are included in their construction. Therefore, they do not reflect the true picture.

3. Selection of the base year is also a difficult task in the construction of index numbers as the selection of a 'normal year' is a subjective matter.

4. Determination of quality of the product is yet another important consideration in the construction of index numbers. However, it is a difficult task in modern times when qualities of different products undergo quick changes.

5. The index number is formed to serve only a specific purpose and, as such, its use is limited only tothe phenomenon under study.

6. Index numbres are subjected to certain errors, namely-sampling errors, miscellaneous errors and incorrect classification of items. Sampling errors crop up in the selection of items, errors arising due to incomplete information, faulty price quotations while, lack of representative character of items come under miscelleneous errors. Classification of people into a specific class is an

important problem in the construction of cost of living index numbers for a particular class of people. The classification may be faulty.

In spite of these limitations, index number are regularly constructed and widely used for studying the related problems.

❄ ❄ ❄

Lesson : 9

TIME SERIES ANALYSIS

Author : **Vetter:**

Prof. Ved Paul **Dr. B. S. Bodla**

Introduction : A time series is an arrangement of statistical data in a chronological order, i.e., in accordance with its time of occurrence. It reflects the dynamic pace of movements of a phenomenon over a period of time. Most of the series relating to Economics, Business and Commerce are all time series spread over a long period of time. Accordingly, time series have an important and significant place in business and economics, and basically most of the statistical techniques for the analysis of time series data have been developed by economists. However, these techniques can also be applied for the study of behaviour of any phenomenon collected chronologically over a period of time in any discipline relating to natural and social sciences, though not directly related to economics or business.

Time series analysis is a quantitative method we use to detect patterns of change in statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future. Thus, time series analysis helps us cope with uncertainty about the future. Moreover, the analysis of time series on major national aggregates such as population, national income, capital formation, etc., provides the most crucial information about the success and

weakness of a growth strategy which may have been adopted in the past, and can serve as the basis of setting targets for the future.

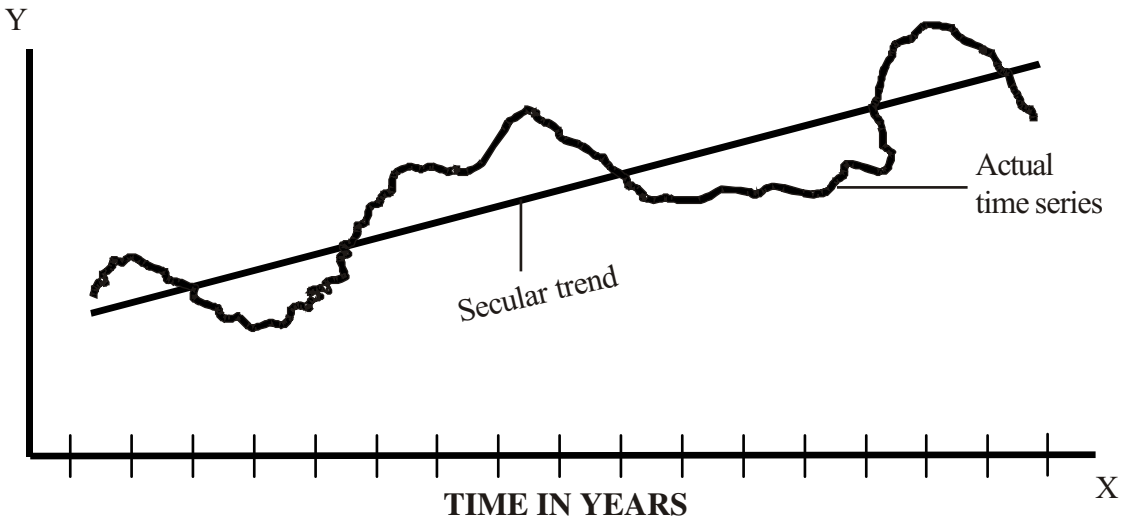
Components of a Time Series

As discussed in Section (4.1) above, a time series refers to any group of statistical information accumulated at regular intervals. Interestingly, if the values of a phenomenon are observed at different periods of time, the values so obtained will indicate appreciable variations or changes. These variations are due to the fact that the values of the phenomenon are affected not by a single factor but due to the cumulative effect of a multiplicity of factors pulling it up and down. For example, the price of a particular product depends on its demand, various competitive products in the market, raw materials and transportation expenses, investment and so on. There are four kinds of changes, or variations (known as components of time series), involved in time series analysis. they are :

1. Secular trend
2. Cyclical fluctuations
3. Seasonal variations
4. Irregular variations

Secular Trend

The general tendency of the time series data to increase or decrease or stagnate during a long period of time is called the secular trend or simple trend. In brief, the trend is the long-term movement of a time-series. The steady increase in the cost of living recorded by the Consumer Price Index is an example of secular trend. From year to year, the cost of living varies a great deal, but if



(FIG. 1)

examined over a long-term period, the trend toward a steady increase is observed. Figure 1 shows a trend in an increasing but fluctuating time series.

There are four reasons why it is useful to study secular trend.

1. The study of trend enables us to describe a historical pattern. Many times a past trend is used to evaluate the success of a previous policy. For instance, an evaluation of the effectiveness of a recruiting programme by a university may be done on the basis of examination of its past enrolment trends.

2. This helps in business forecasting and planning future operations. For example, if the time series data for a particular period regarding a particular phenomenon, say, growth rate of the India's population, is observed, we can estimate the population for some future time.

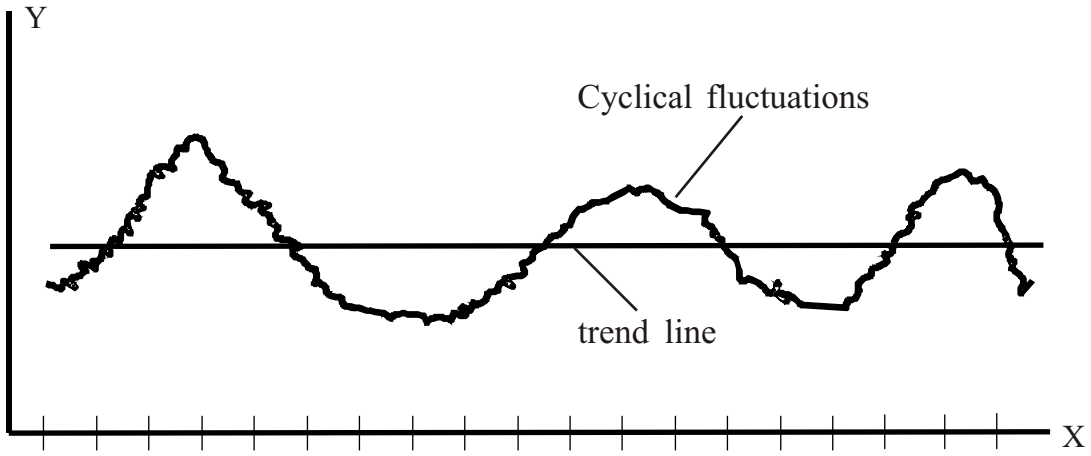
3. Trend makes it easier for us to study the other three components of the time series. This can be done by isolating trend values from the given time series.

4. Trend analysis enables us to compare two or more time series over

different periods of time and draw important conclusions about them.

Cyclical Variations

The oscillatory movements in a time series with period of oscillation greater than one year are termed as cyclical variations. The most common example of cyclical fluctuation is the business cycle. These variations are the upswings and downwings in the time series that are observable over extended period of time. Neither the amplitude nor the frequency of occurrence of these cycles is uniform. Empirical studies based on the analysis of time series data on a large number of major economic aggregates for developed countries have shown that the length of time interval after which cycles occur ranges from 8 to 10 years. Figure 2 illustrates a typical pattern of cyclical movements. A

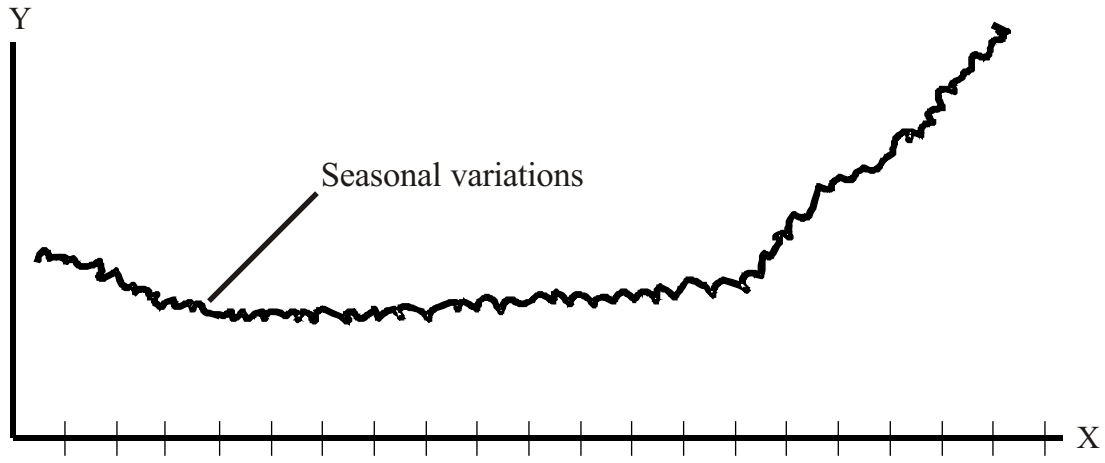


TIME IN YEARS
(FIG. 2)

knowledge of the cycle component enables a businessman to have an idea about the periodicity of the booms and depressions and accordingly he can take timely steps for maintaining stable market for his product.

Seasonal Variations

As we might expect from the name, seasonal variation involves patterns of change within a year that tend to be repeated from year to year. Some examples are the production of soft drinks, which is high during the summer and low during the winter; substantial increases in the number of flu cases every winter;



TIME IN MONTHS
(FIG. 3)

woollen garments sales, which is high from October month to January and low during rest of the months.

In each of these examples, note that there are systematic causes of these fluctuations, such as the weather, holidays and government accounting procedures and so forth. These systematic causes occur regularly. Some other causes like national level festivals including Dussehra, Holi, Diwali also bring a shift in sales of departmental stores. Since these variations are regular pattern, they are useful in forecasting the future. Fig. 3 illustrates a typical pattern of seasonal variations.

Random or Irregular Variations

These variations are purely random and are the result of such unforeseen and unpredictable forces which operate in absolutely erratic and irregular manner. These powerful variations are caused by numerous non-recurring factors like floods, famines, wars, earthquakes, strikes, revolution, etc., which behave in a very unpredictable manner. The collapse of OPEC in 1986, the Iraqi situation in 1990 on gasoline prices in the United States, Security Prices fluctuations in India in March-April 1992 on account of SCAM are some examples of irregular variations.

Models of Time Series Analysis

The following are the two models commonly used for the decomposition of a time series into its components :

1. $0_t = T_t + S_t + C_t + I_t$..... Additive Model

2. $0_t = T_t \times S_t \times C_t \times I_t$ ..... Multiplicative Model

Where 0_t is the time series value at time t, and T_t, S_t, C_t and l_t represent the trend, seasonal, cyclical and random variations at time t. In these models $S = S_t$, $C = C_t$ and $I = I_t$ are absolute quantities which can take positive and negative values so that

$\sum S = \sum S_t = 0$, for any year

$\sum C = \sum C_t = 0$, for any cycle and

$\sum I = \sum I_t = 0$, in the long-term period.

The first model assumes that the economic time series is additive and is made

up of the four components T, S, C and I. Here it is assumed that the four components are independent of one another. Independence is said to exist when the pattern of occurrence and the magnitude of movements in any particular component are not affected by the other components. As a concrete example, the production of beer has been increasing over last many years. This additivity assumption implies that this steady increase in the production of beer has no effect on the seasonal variation of the production of beer. It also implies that the causes for the increase in the production of beer are different from the causes of the seasonal variation of beer. It may be noted that when the time series data are recorded against years, the seasonal component would vanish and in that case the additive model will take the form.

$$0_t = T_t + C_t + R_t$$

The second model - the multiplicative model, is used where it is assumed that the forces giving rise to the four types of variations are interdependent, so that the overall pattern of movements in the time series is the combined result of the interaction of all the forces operating on the time series. According to this assumption, the original magnitude of the time series are the product of its four components. The reason for using this model is that it allows convenient isolation of the components. If the decomposition of a time series is done by taking logarithms, the multiplication model will be expressed as

$$\text{Log } 0_t = \text{Log } T_t + \text{Log } C_t + \text{Log } S_t + \text{Log } R_t$$

Thus, we see that the four components of a time series relating to economic and business phenomenon conform to the multiplicative model. In practice, additive model is rarely used.

Decomposition of Time Series into Its Four Components

We have seen that there are two models of time series which can be used for decomposition of it. It could be observed from both of these models that decomposition of a time series requires estimation of its four components and then separating them from each other so as to be able to understand the pattern of variations in each component independently. We shall follow the Pearson's approach based on the multiplication model for decomposition of time series.

The first component to be estimated is the trend variations. Trend variations are estimated by fitting a trend line on the time series data. After estimating the trend variations, these are then separated from the time series is known as detrending. Detrending requires dividing both sides of Multiplicative model by the trend values T_t, so that

$$\frac{0_t}{T_t} \quad C = \quad _t . S_t . R_t$$

After isolating trend, we shall compute and separate the seasonal varations and the resulting multiplicative model would be expressed as

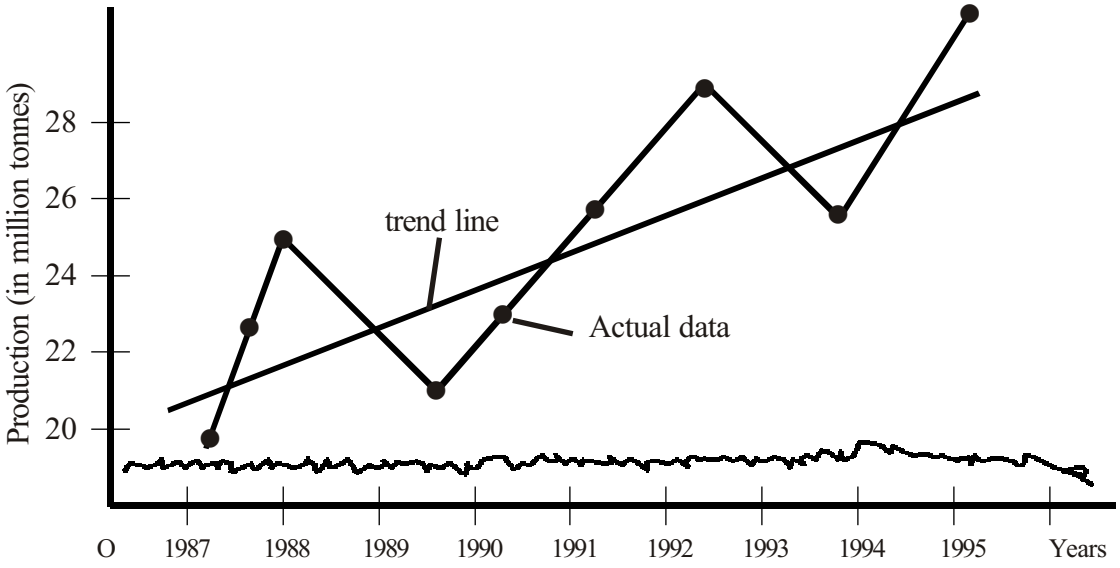
$$\frac{0_t}{T_t S_t} = C_t . R_t$$

So far we have discussed ways of separating the trend T and the seasonal variation S. The cyclical variations and irregular variations can be examined easily with reference to the pattern of their occurrence and amplitude. If we

ignore the random variations then the values obtained in the following equation may be taken to reflect cyclical variations.

$$\frac{0_t}{T_t, S_t} = C_t . I_t$$

To arrive at better estimates of cyclical fluctuations, the irregular component (I) should be eliminated from the $C_t . I_t$ value obtained in above equation. However, the extent to which their elimination is possible, they tend to become



(FIG. 4)

marginal in the process of deseasonalisation.

Estimation of Trend

Of the four components of a time series, secular trend represents the long-term direction of the series. There are various ways to describe the trend or fitting a straight line, such as the freehand curve method, the method of semi averages. the method of moving averages, and the method of least squares. In

this lesson we shall discuss each of these methods in estimation of trend. The general formula for a straight line is Y = a+bx where x is called the independent variable, and Y is called the dependent variable is the Y-intercept of the straight line and b is the slope of the trend line.

The Free-hand Method

The simplest method of finding a trend line when given a set of time series data is the free hand method. According to this method first of all we shall plot the data on a graph and then, by observation, will fit a straight line through the plotted points in a way such that the straight line shows the trend of the time series.

Illustration-1: Fit a trend line to the following data by free hand method

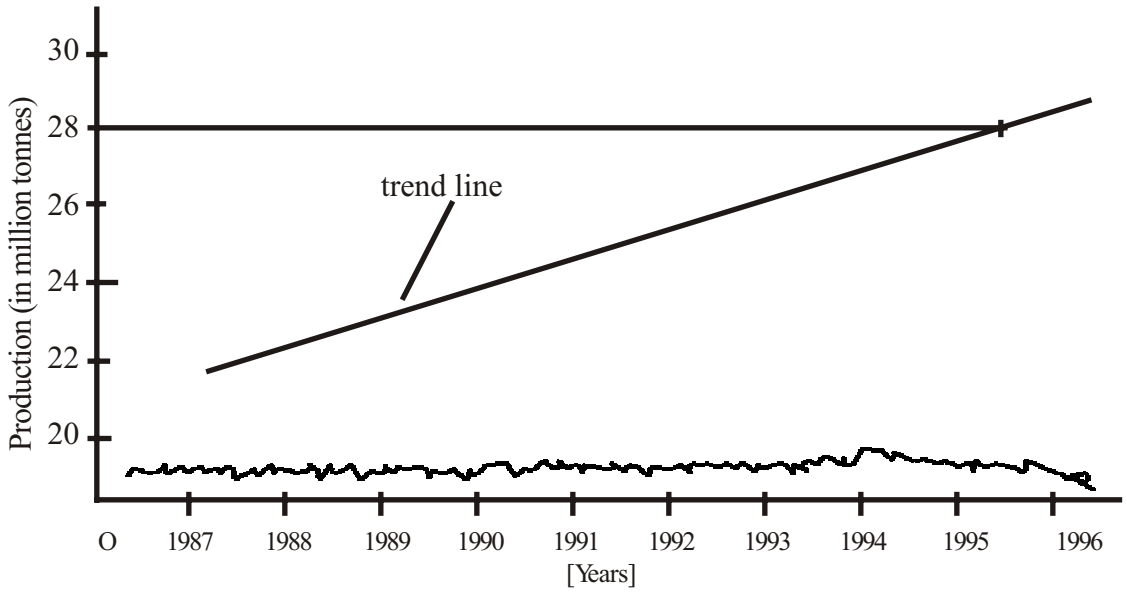
Year	*X*	*Production of Steel (in million tonnes)*	*Year*	*X*	*Production of Steel (in million tonnes)*
1987	-3	20	1992	2	25
1988	-2	22	1993	3	26
1989	-1	24	1994	4	25
1990	-0	21	1995	5	28
1991	1	23			

Solution :

From figure (4) it is obvious that this is not an accurate way of fitting a straight line or a curve to the data as it gives only rough idea regarding trend.

Finding the trend Equation : It requires selection of two points on the straight line. An important feature of time series is that the data are given in order of time. In illustration 1, it starts from 1987 and goes up to 1995 in one year time intervals. Let us start at 1987, and call it the 'origin', and designate it as zero. Then 1988 is 1, 1989 is 2 and so forth, as shown in illustration and also figure. In this way the origin may be placed at any year. If we let 1990 be the origin, then 1987 is -3, 1988 is -2, 1989 is -1, 1990 is 0, 1991 is 1, and so on.

Now, let us assume that the trend line goes through the points for 1987 and 1995 (illustration 1). It becomes a problem of finding the equation for the straight line going through the two points 1987 and 1995. The coordinates of the two points selected now become (-3, 20) and (5, 28). Substituting the values



(FIG. 5)

of these coordinates into the equation for a straight line, we find.

$20 = a + (-3b)$

$28 = a + 5b$

Solving these two equations gives a = 17, b = 1. Thus the equation for the trend line is

$Y_c = 17 + 1X$ ..... Y_c = Estimated value.

The interpretation of the equation is, when X = -3 (1987),

$Y_c = 17 + (1)(-3) = 17-3 = 14$

which indicates that the estimated production by the trend line is 14 million tonnes.

Method of Semi averages

This method divides the time series into two parts, finds the averge of each part, and then fits a trend line through these averages. Note that in case number of years are even, middle year is left out for computing the semi averages.

Illustration-2: Using the statement of illustration-1 estimate the value for 1996 by applying method of semi averages.

Solution :

Here n = 9, and hence the two parts will be 1987 to 1990 and 1992 to 1995. As n is a odd number we will ignore the middle year of the series, i.e. 1991.

Year	***Production of Steel (in million tonnes)***	***4 years semi-total***	***Semi averges (A.M.)***
1987	20		
1988	22		

1989	24	87	21.75
1990	21		
1991	23		
1992	25		
1993	26		
1994	25	104	26.00
1995	28		

Here these miaverage 21.75 is to be plotted against the middle of the years 1988 and 1989 and value 26 against the middle of the years 1993 and 1994. The graph is shown in Fig (5).

From the graph we see that the estimated value for 1996 is 27.7.

This method assumes the presence of linear trend which may not exist. Semi averages are affected by extreme values. This is a crude and simple way of fitting a trend line, but its simplicity is its advantage.

Moving Average Method

This is a very simple and flexible method of measuring trend. When a trend is to be determined by this method, the average value for a number of years (or months) is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the average. The averaging process smoothens out fluctuations in the given data.

While applying moving average method, the choice of the length of the period for a moving average is necessary because this would determine the extent to which variations would be smoothened in the process of averaging. The period

of moving average is to be decided in the light of the length of the cycle. These averages donot yield an equation which could be used for forecasting the values of a time series variable for the future.

Illustration 3 : Estimate the trend values using the data given below by taking a three-yearly moving average.

Year	*Value*	*Year*	*Value*
1987	3	1992	11
1988	4	1993	09
1989	8	1994	10
1990	6	1995	14
1991	7		

Solution :

Year	*Value*	*Three-yearly moving-Total*	*Three yearly moving average*
1987	3		
1988	4	15	5
1989	8	18	6
1990	6	21	7
1991	7	24	8
1992	11	27	9
1993	9	30	10
1994	10	33	11
1995	14		

Note : If the moving average is an even period moving average, the moving total and moving average which are placed at the centre of time span fromwhich they are computed fall between two time periods.

The moving average method is applicable not only to trend lines but also to all kinds of data that show regular periodic fluctuations. We shall use it also to eliminate seasonal fluctuations.

The Method of Least Squares

The method of least square is the most widely used method of fitting a straight line to a series of data. Estimation of trend values by this method makes use of the general equation for estimating a straight line.

$$Y_c = a+bx.$$

The values of the two constants, a and b, in the Equation are obtained by solving simultaneously the two normal equations.

$$\sum Y = na + b\sum x$$

$$\sum XY = a \sum X + b \sum X^2$$

Where n represents number of years for which data are given. We can measure the variable X from any point of time in origin such as the first year. But the calculations are simplified when the mid-point in time is taken as the origin because in that case the negative values in the first half of the series balance out the positive values in the second half so that $\sum x = 0$. Since $\sum X = 0$ the above two normal equations would take the form

$$\sum Y = na \qquad \text{..... (i)}$$

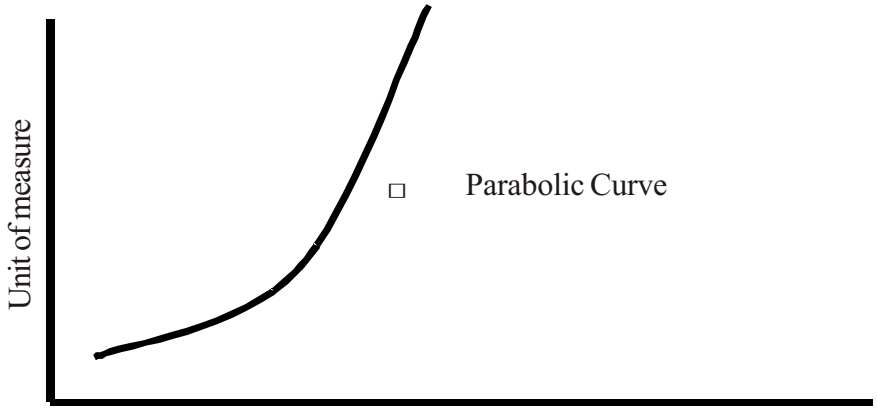
$\sum XY = b\sum X^2$ ..... (ii)

The values of a and b can now be determined easily.

Since $\sum Y = na$, $\therefore\ a = \sum Y/N$

Since $\sum XY = b\sum X^2$, $\therefore\ b = \sum XY/\sum X^2$.

It should be noted that in case of odd number of years, when deviations are taken from the middle year. $\sum X$ would always be zero provided there is no gap in the data given. However, in case of even years also $\sum X$ will be zero if the X



(FIG. 6)

origin is placed midway between the two middle years.

Illustration : 4

Fit a trend line to the following data by the least squares method

Year	:	1991	1993	1995	1997	1999
Production (in '000 tons)	:	36	42	46	54	32

Estimate the production in 1996 and 2002.

Solution :

Let the trend line be given by the equation :

$Y = a+bx$

Where origin is at 1995.

Computation for Straight Line Trend

Year t	***Production (in '000 tons) (Y)***	***X=t-1989***	***XY***	***X^2***	***Trend values Y_c***
1991	36	-4	-144	16	41.2
1993	42	-2	-84	4	41.6
1995	46	0	0	0	42.0
1997	54	2	108	4	42.4
1999	32	4	128	16	42.8
n=5	$\sum Y=210$	$\sum X=0$	$\sum XY=8$	$X^2 = 40$	

Since $\sum X = 0$, $a = \sum Y/N$, $b = \sum XY/\sum X^2$

We have

$\sum Y=210$, $N = 5$, $\sum XY = 8$, $\sum X^2 = 40$

$\therefore\ a = \frac{210}{5} = 42,\quad b = \frac{8}{40} = 0.2$

Substituting values of a and b in the least square equation we get.

$Y_c = 42 + 0.2X$

By substituting X = -4, -2, 0, 2, 4, in the above equation we will obtain the estimated values for the years 1991, 1993, 1995, 1997 and 1999 respectively.

Thus

$Y_{1991} = 42 + (0.2)(-4) = 42 - 0.8$

$= 41.2$

$Y_{1993} = 42 + (0.2)(-2) = 42 - 0.4 = 41.6$

$Y_{1999} = 42 + (0.2)(0) = 42.0$

$Y_{1996} = 42 + (0.2)(2) = 42.4$

$Y_{2002} = 42 + (0.2)(4) = 42.8$

The estimated production in 1996 is obtained on taking

X = t-1995 = 1996 - 1995 = 1

$Y_{1996} = 42 + (0.2)(1) = 42.2$

The estimated production in 2002 is obtained on taking

X = t - 1995 = 2002 - 1995= 7

Hence $Y_{2002} = 42 + (0.2)(7) = 43.4$.

Second Degree Polynomial Trend

In the previous lesson, we have described the method of fitting a straight line to a time series. But many time series are best described by curves, not straight lines. The linear model does not adequately describe the change in the variable

as time changes in case of non-linear trend. Most often a parabolic curve is used to overcome this problem. The parabolic curve is described mathematically by a second degree equation. A hypothetical parabolic curve is illustrated in Fig. (6). The general form for an estimated second-degree equation is :

$$Y_c = a+bX = CX^2$$

where : Y_c is the estimated value of the dependent variable; a, b and c are numerical constants, and X represents the coded value of time variable.

It must be noted that if the third term CX^2 is introduced in Eq. , it will give a parabolic trend. In the above Eq. the value of c reveals whether the resultant second degree curve is concave or convex. Here the value of c also determines the extent towhich the curve departs from linearity.

The derivation of the second-degree equation is beyond the scope of this lesson. However, we can determine the value of the numerical constants (a, b and c) from the following three equations :

$$\sum Y = an + c\sum X^2$$

$$\sum X^2Y = a \sum X^2 + C\sum X^4$$

$$b = \frac{\sum XY}{\sum X^2}$$

After finding the values a, b and c by solving the above equations, we substitute these values in the equation for a second-degree parabola. A problem involving a parabolic trend is considered below in illustration.

Illustration 6 : Fit an equation of the form $Y = a+bX + cX^2$ to the data given below :

Years	1991	1992	1993	1994	1995
Consumption of wheat (in Qtls.)	25	28	33	39	46

Solution

Calculations for Second Degree Trend

Years	*Consumption (in Qtls.)* Y	X	X^2	X^4	XY	X^2Y
1	25	-2	4	16	-50	100
2	28	-1	1	1	-28	28
3	33	0	0	0	0	00
4	39	1	1	1	39	39
5	46	2	4	15	92	184
n=5	$\sum Y=171$	$\sum X=0$	$\sum X^2=10$	$\sum X^4=34$	$\sum XY=53$	$\sum X^2Y=351$

We are to fit $Y_c = a+bX+cX^2$

The first step in fitting a second degree equation is translate the independent variable (time) into a coded time variable X. Not that the coded variable X is listed in one year intervals because there is an odd number of elements in our time series. Now find values of a, b and c by solving the three equations meant

for the purpose.

Three equations are

$$\Sigma Y = an + c\Sigma X^2 \quad \text{..... (1)}$$

$$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4 \quad \text{..... (2)}$$

$$b = \frac{\Sigma XY}{\Sigma X^2} \quad \text{..... (3)}$$

By substituting the given values in the above equations, we get :

171 = 5a + 10c ..... 1 (a)

351 = 10a + 34c ..... 2 (a)

b = Í 0 ¸ ÷ 5.3 ..... 3 (a)

Now we shall find a and c by solving equations 1(a) and 2(a).

1. Multiply equation 1(a) by 2 and subtract equation 2(a) from equation 1(a)

342 = 10a + 20c

-351 = -10a - 34c

———————

-9 = -14c ..... (4)

From equation (4) we find c

c = -9/-14 = 0.64

2. Substitute the value for c into equation 1(a).

$171 = 5a+(10)\ (0.64)$

$171 = 5a + 6.4$

$164.6 = 5a$

Lastly, we shall put these numerical values in the general equation as follows :

$Y_c = a + bx + cx^2$

$32.92 + 5.3x + 0.64x^2$

Exponential Trend :

$Y_c = ab^x$ ..... (5.5)

Where :

a and b are the two constants, and

X represents the values assigned to time.

In general, the exponential trend is applicable, where growth in the time series data is nearly at a constant rate per unit of time (expressed in percentage). When the aggregate variable related to national product, population, or production in the country as a whole or in a region are given we normally use exponential trend.

Taking logarithm of both sides of Eq. (5), we get

$\text{Log } Y = \text{Log } a + X \log b$ ..... (6)

When plotted on a semi logarithmic graph, the curve gives a straight line. However, on an arithmetic chart the curve gives a non-linear trend.

To obtain the values of the two constants, a and b, we need to solve simultaneously the two normal equations :

$$\sum \text{Log } Y = n \log a + \log b \sum X$$

$$\sum (X \log Y) = \log a \sum X + \log b \sum X^2$$

When deviations are taken from middle year, i.e., $\sum X = 0$, the above equation takes the following form :

$$\sum \log Y = n \log a$$

$$\text{or } \log a = \frac{\sum \log y}{n}$$

and $\sum (X, \log Y) = \log b \sum X^2$

$$\therefore \log b = \frac{\sum (X \log Y)}{\sum X^2}$$

The rate of growth implicit in a semilogarithmic trend is often of interest. It is derived by solving the equation for compound interest - $\log (1+r) = b_1$

Here b_1 is the slope and r is the rate of growth.

Illustration 7 : The sales of a company in lakhs of rupees for the years 1988 to 1994 are given below :

Years :	1988	1989	1990	1991	1992	1993	1994
Sales :	16	23	33	46	66	95	137

Estimates sales for the year 1995 using an equation of the form $Y = ab^x$, here X = years and Y = sales.

Solution

Fitting Equation of Form $Y = ab$

Year X	***Sales Y***	***X (Coded)***	***Log Y***	X^2	***X Log Y***
1988	16	-3	1.2041	9	-3.6123
1989	23	-2	1.3617	4	-2.7234
1990	33	-1	1.5185	1	-1.5185
1991	46	0	1.6627	0	0
1992	66	1	1.8195	1	1.8195
1993	95	2	1.9777	4	3.9554
1994	137	3	2.1367	9	6.4101
		$\sum X=0$	$\sum \text{Log}Y=11.6819$	$\sum X^2=28$	$\sum X\text{Log}Y=4.3308$

$$\text{Log } a = \frac{\sum \text{Log}Y}{n} = \frac{11.6809}{7} = 1.6687$$

$$\text{Log } b = \frac{\sum X \text{ Log}Y}{\sum X^2} = \frac{4.3308}{28} = 0.1547$$

We know, Log Y = Log a + X Log b

$\therefore$ log Y = 1.6687 + 0.1547 X

For 1995, X would be +4. When X=4, Log Y will be -

Log Y = 1.6687 + (0.1547) (4) = 2.2875

Y = Anti log 2.2875 = 193.86

Thus the estimated sales for the year 1995 is Rs. 193.86 lakhs.

Do Yourself

1. What is a time-series ? What are its main components ?

2. What do you mean by decomposition of a time series ?

3. Distinguish between additive and multiplicative model in the analysis of time series.

4. What is 'Secular Trend' ? Discuss the various ways of estimating the trend value.

5. Fit a trend line from the following data by using semiaverage method:

Year	:	1993	1994	1995	1996	1997	1998
Profits (in '000 Rs.)	:	100	120	140	150	130	200

Answer : Joining the points (1994, 120) and (1997, 160) we get the trend line.

6. The following table shows the number of salesmen working in a certain concern.

Year	:	1994	1995	1996	1997	1998

No. of salesmen : 28 38 46 40 56

Use the method of least squares to fit a straight line trend and estimate the number of salesmen in 1999.

(Ans. Trend values 30, 35.8, 41.6, 43.4, 53.2 and $Y_{1999} = 49$)

7. Using three-year moving averages, determine the trend and short-term fluctuations. Plot the original and trend values on the graph paper.

Year	*Production (in '000 tons)*	*Year*	*Production (in '000 tons)*
1978	21	1983	22
1979	22	1984	25
1980	23	1985	26
1981	25	1986	27
1982	24	1987	26

Ans. : Using additivemodel, short-term fluctuations are 0, -0.3, 1.0, 0.3, -1.7, 0, 0.7

8. Fit a parabolic curve of the second degree to the data given below and estimate the value for 2000 and comment on it.

Year : 1994 1995 1996 1997 1998

Sales (in '000 Rs.) : 10 12 13 10 8

(Answer : $Y_{2000} = 7.226$)

9. The sales of a company for eight years are given below :

Year : 1988 1989 1990 1991 1992 1993 1994 1995

Sales (in '000) : 52 45 98 92 110 185 175 220

Estimates sales figure for 1996 using an equation of the form $Y = ab^x$ where X=years and Y= sales. (Ans. Sales for 1996-294.1)

❄ ❄ ❄

Lesson : 10

PROBABILITY

Author : **Vetter:**

Prof. Ved Paul **Dr. B. S. Bodla**

Objective: The main objective of this lesson is to make the students learn about the basic concepts of probability with reference to its applications in statistical analysis.

Structure

4.1 Backdrop

4.2 Meaning and Definition

4.3 Approaches to Probability

4.4 Probability Theorems for Problems Solving

4.5 Permutation and Combinations

4.6 Summary

4.7 Self- Assessment Exercise

4.8 Suggested Readings

4.1 Backdrop

The word probability or chance is very common in day to day life of human being. For instance , we come across statements like " India may win the World Cup"; "It is likely that Mr X may not come for teaching statistics class today"; "Probably Iraq may win the Gulf War against USA". All these terms possible, probable, likely, etc, convey the same sence i.e. the event is not certain to take place or there are some uncertainties about the happening of the events. In simple words, the word probability

thus refers that there is uncertainty about the happening of event.

The theory of probability has its origin in the games of chance related to gambling such as throwing a die, tossing a coin, drawing cards from a pack of cards etc. Jerane Cardon (1501-76), an Italian mathematician, was the first scholar to write a book on the subject entitled "Book on Games of Chance" which was published after his death in 1663. During the last quarter of the eighteenth century, the study of the games of chance no longer remained dependent on the initiative of the gamblers. The subject became an an area of academic interest and a number of scholars addressed themselves to the field of probability. In the early nineteenth century the famous French mathematician Laplace, and the German mathematician Gauss carried the knowledge of the subject many important steps forward. With the expansion of national economics, some great persons like De Moivre (1718), James Bernoulli (1713), Bayes (1768) etc., were inspired to develop the theory of probability further and apply it in different fields of decision making. In later years, R.A. Fisher, Karl Pearsons, J. Neyman etc. developed a sampling distribution theory based on the laws of probability.

Today a comprehensive theory of probability exists and in the words of Emile Borel, "Probability theory is of interest, not only to card and dice players, who were its godfathers, but also to all men of action, heads of industries or heads of armies, where success depends on two sorts of factors the one known or calculable, the other uncertain and probabilitical."

4.2 Meaning and Definition

One of the major reasons for the evolution and development of the theory of probability is its presence in almost every aspect of practical life. A phenomenon is random if chance factors determine its outcome. All the possible outcomes may be known in advance, but the particular outcome of a single trial in any experimental operation cannot be pre-determined. Nevertheless, some regularity is built into the

process so that each of the possible outcomes can be assigned a probability fraction. The simplest example of a random phenomenon is the result of the toss of a coin. Though all the possible outcomes are known, i.e., head or tail, but chance factors determine the outcome of any single toss. There is no deterministic regularity here, that is, one cannot say for sure that head or tail, will come up on a particular toss. Similarly, in the roll of a cubic die, we cannot predetermine which side will turn up; eventhough, all the possible outcomes are known. The existence of random phenomena is found in so many diversified fields that it is imperative particularly for students in the social sciences to study the theory of probability.

Probability is especially important in statistics because of the many principles and procedures that are based on this concept. Indeed, probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phrases like: 'You had better take an umbrella because it is likely to rain.' 'His chances of winning are pretty small.' It is very likely that it may rain by the evening. You are probably right.' or 'There are fifty-fifty chances of his passing the examination.' In each of these phrases an idea of uncertainty is acknowledged. Goethe remarked that, "There is nothing more frightful than action in ignorance." Reasoning in terms of probabilities is one weapon by which we attempt to reduce this uncertainty or ignorance. The use of word 'probability' in statistics, however is somewhat different. It is more precise than what it means in popular usage. In statistics, a probability is a numerical value that measures the uncertainty that a particular event will occur.

4.3 Approaches to Probability

There are basically four approaches for measuring the probability. They represent different conceptual aspects for understanding the gambut of probability. They are:

- Classical Approach.
- Empirical Approach.
- Subjective Probability.

- Modern Approach

Let us discuss them in detail for the sake of smooth understanding of the students.

- **Classical Approach to Probability**

Since the theory of probability had its origin in gambling games, the method of measuring probabilities, which was just developed, was particularly appropriate for gambling situations. This method of measuring probability is called classical or a priori concept of probability. In such situations, the probability of an event is simply the ratio of number of favourable outcomes of an event to the number of possible outcomes, where each outcome is equally likely to occur. In other words, if there are several equally likely events that may happen, the probability that any one of these events will happen, is the ratio of the number of cases fovourable to its happening to the total number of possible cases. If there are 'n' possible outcomes favourable to the occurrence of an event 'A' and 'm' possible outcomes unfavorable to the occurrence of A, and all of these possible outcomes are equally likely and mutually exclusive, then the probability that A will occur denoted by p (A) is

p(A)= n/n+m = Number of outcomes favourable to occurrence of A / Total number of possible outcomes

and the probability, that A will not occur, denoted by q (A) is

q(A) = m/n+m = Number of outcomes not favourable to occurrence of A / Total number of possible outcomes

For example, if we toss a coin, the probability of the head coming up is : p = 1/2, because the number of favourable event is 1, and the total number of possible outcomes is 2. The probability of head not coming up is:

$$q = 1/2$$

p and q of an event is equal to 1. (p+q = 1). then, 1 - p = q, 1 - q = p.

Let us Understand the Fundamental Concepts Used in Probability Theory

(i) Random Experiment

Probabilities are obtained for the outcomes of the situations, which are called random experiments. The term 'experiment' is used in Statistics in a much broader sence than in Chemistry or Physics. The tossing of a coin, for example, is considered a statistical experiment. An experiments has two properties : (a) Each experiment has several possible outcomes and that can be specified in advance. (b) we are uncertain about the outcome of each experiment. While tossing a coin, it can be specified that head or tail will turn up, but we are not certain whether the outcome of a particular toss will be head or tail. We use the word 'experiment' because the outcome is yet to be determined, whereas, the objective 'random' signifies that any particular outcome is uncertain.

(ii) Sample Space

A sample space of an experiment is the set (or collection) of all possible outcomes. The sample space for tossing a coin contains just two outcomes, head or tail; thus the sample space = (Head, Tail). The sample space of throwing a die will be (1, 2, 3, 4, 5, 6). Each possible outcome given in the sample space is called element or sample point. If two coins are to be tossed once, the four possible outcomes of this experiment will be :

Coin 1 \ Coin 2	H	T
H	HH	HT
T	TH	TT

The sample space of this experiment = (HH., HT, TH, TT).

If a pair of dice is to be cast once, the 36 possible outcomes of this experiment, will be :

Outcome of	Outcome of Second Die					
First Die	1	2	3	4	5	6
1	(1, 1)	(1,2)	(1, 3)	(1.4)	(1, 5)	(1, 6)
2	(2, 1)	(2,2)	(2, 3)	(2.4)	(2, 5)	(2, 6)
3	(3, 1)	(3,2)	(3, 3)	(3.4)	(3, 5)	(3, 6)
4	(4, 1)	(4,2)	(4, 3)	(4.4)	(4, 5)	(4, 6)
5	(5, 1)	(5,2)	(5, 3)	(5.4)	(5, 5)	(5, 6)
6	(6, 1)	(6,2)	(6, 3)	(6.4)	(6, 5)	(6, 6)

(iii) Event

An event is any element or point of the sample space in which we are interested. An event is called simple event, if it contains one element, if it is made up of more than sample point, it is called compound or complex event. A simple event is not decomposable, while a compound event can be decomposed into a number of disjoint simple events.

(iv) Mutually Exclusive Events

Mutually exclusive events are such events where the occurrence of one event prevents the possibility of the other to occur. In simple words, when several events are mutually exclusive, at the most one event may occur. A very simple example of a collection of mutually exclusive events is given by the coin toss. There are two possible events, a head or a tail. Since both events cannot occur on the same toss, they are mutually exclusive, the occurrence of one event rules out the occurrence of the other.

(v) Equally Likely Events

Events are said to be equally likely if after all relevant evidence has been taken

into account, one of them may not be expected rather than the other. For example, head and tail are equally likely events in tossing an unbiased or symmetrical coin.

(vi) Exhaustive Events

The events are said to be exhaustive if at least one of them necessarily occurs. In other words events are defined to be exhaustive if they between themselves exhaust all possible outcomes of the random experiment. For example, throwing of a die consists of 6 exhaustive events.

(vii) Independent Events

Events are said to be independent, if the occurrence of one does not affect the occurrence of any of the others. Two events are independent when they have no influence on each other. The result of the first toss of a coin does not affect the result of successive tosses at all.

(viii) Dependent Events

If the occurrence of the 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 4/54 or 1/12; if it happens that king is drawn and is not replaced in the pack, the probability of drawing again a king would be 3/51. Thus, the outcome of the first event has affected the outcome of the second event. So they are dependent events.

(ix) Complementary Events

To any event A, there is an event denoted by 'not A' or A and called the complementary of A. A contains all the outcomes of the experiment which are not in A. Thus 'No head' or 'At least one head' are complementary events in two tossings of a coin.

From the concepts discussed above, it may be observed, that a probability will always be a number between 0 and 1 inclusively. This is because the numerator in the

probability fraction can never be negative nor can it be larger than the denominator. Two important observations follow from this definition. First, an event that is certain to occur will have the same value in both the numerator and the denominator, for the same events will result from all experiments. The probability in such a case will be 1. At the other extreme in impossible event's frequency ratio will always be 0 in the numerator, for such an event will occur in name of the experiments. Thus:

p (certain event) = 1, p (impossible event) = 0

Expressions of Probability

Probabilities can be expressed either as ratios, fraction or in percentages. For example, the probability of getting a head in a toss of coin can be express as 1/2 or .5 or 50%.

Illustration 4.1

(a) What is the chance of drawing a king in a draw from a pack of 52 cards?

Solution: Total number of cases that can happen = 52.

No. of favourable cases = total number of kings in a pack of cards = 4

The Probability (p) = 4/52 or 1/13.

(b) An urn contains two blue balls and three while balls. Find the probability of a blind man obtaining one blue ball in a single draw.

Solution:

p = 2/(2+3) = 2/5.

(c) If two dice are thrown - (i) What is the probability of throwing two sixes ? (ii) What is the probability of throwing a total of 9? (iii) What is the probability of not throwing a total of 9?

Solution: The total number of outcomes in the throw of two dice will be 36.

Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd	Ist	2nd
1	1	2	1	3	1	4	1	5	1	6	1
1	2	2	2	3	2	4	2	5	2	6	2
1	3	2	3	3	3	4	3	5	3	6	3
1	4	2	4	3	4	4	4	5	4	6	4
1	5	2	5	3	5	4	5	5	5	6	5
1	6	2	6	3	6	4	6	5	6	6	6

(i) In the throw of two dice, two sixes can come only once, when the total number of outcomes will be 36. Hence the probability of coming of two sixes in the throw of two dice will be 2/36.

(ii) In the throw of two dice, total of 9 can come in this way : 3,6 ; 4 ,5; 5, 4; 6, 3. Hence, the probability of coming of a total of 9 in the throw of two dice will be 4/36 or .

(iii) The probability of not throwing a total of 9 in the throw of two dice will be

$$1 - 1/9 = 8/9.$$

(d) What is the probability that a vowel selected at random in any English book is an 'I'.

Solution: Total number of equally likely events = 5

Number of favourable events = 1

$$p = 1/5.$$

(e) What is the probability of a king in a pinochle deck. (A pinochle deck consists of 2 aces. 2 kinds, 2 queens, 2 jacks, 2 tens and 2 nines of each suit. There are no cards of lower value).

Solution: Total number of cases = Total No. of cards = 48

Total number of favourable cases = Total number of king in the deck = 8

The p = 8/48 = 1/6.

(f) The ten digits 0 to 9 are stamped on 12 discs, there being one digit on a disc and the discs are thoroughly mixed in a box. If a disc is drawn at random, find the probability that the disc has an odd digit on it.

Solution: Total number of cases = 10.

Total number of favourable cases (1, 3, 5, 7 and 9) = 5

The p = 5/10 or 1/2 or 0.5.

(g) Find the probability of drawing a black card in a single random draw from a well-shuffled pack of ordinary playing cards.

Solution: Total number of outcomes = 52

No. of favourable outcomes = 26

Hence, p (drawing a black card) = 26/52 = 1/2.

(h) Find the probability of drawing a face card in a single random draw from a well-shuffled pack of ordinary playing cards.

Solution: There are 52 mutually exclusive equally likely outcomes.

The number of favourable outcomes (face cards - include the jack, the queen and the king in each) is 12. Thus

p (drawing a face card) = 12/52 = 3/13.

By finding probabilities, for different events, a probability table can be constructed. In such a table, the probabilities of happening of all possible events can be seen simultaneously.

Illustration 4.2

There are 50 balls, each ball having two colours, one black or white and the other red, orange or green as shown in the following table:

	Red	Orange	Green	Total
Black	3	12	15	30
White	7	3	10	20
Total	10	15	25	50

(It means there are three balls, which are black and red, 12 balls are black and orange and so on.)

If of these balls, one ball is selected at random, find the probability of each type of ball being drawn up.

Solution:

Probability Table

	Red	Orange	Green	Total
Black	0.06	.24	.30	0.60
White	0.14	.06	.20	0.40
Total	0.20	.30	.50	1.00

- **Empirical Approach to Probability**

The classical or a priori approach to probability, while useful for solving problems involving games of chance, suffers from serious difficulties, and does not provide answers to wide range of other types of problems. For example, it can tell the probability producing defective items in a production process. Such sort of questions can be answered with reference to empirical data. The probability of an event can be obtained on the basis of past records of the frequency distribution. For example, if a train comes daily. Past records show that in the last 365 days it was late on 13 days, then the probability of its late coming is (p) = 13/365.

According to Van Mises. "If an experiment be repeated a large number of times under essentially identical situations, the limiting value of the ratio of the number of

times the event A happens to the total number of trials of the experiments as the numbers of trials increases indefinitely, is called the probability of the occurrence of A" Thus

$$P(A) = m/n$$

It is assumed that the limit is finite and unique.

Here

m = no. of times an event A occurs

n = no. of times the experiment is performed

Illustration 4.3

(a) The following table gives a distribution of wages:

Weekly wages:	30-35	35-40	40-45	45-50	50-55	55-60	65-65	65-70
No. of workers:	9	108	488	230	112	30	16	7

An individual is taken at random from the above group. Find the probability (i) his wages were under 40, (ii) his wages were 55 or over, and (iii) his wages were either between 45-50 or 35-40.

Solution

(i) Total wage earners = 9 + 108 +488 + 230 +122 + 30 + 16 + 7 = 1000

No of wage earners earning wages below 40 is = 9 + 108 = 117

Thus (p) = 117/1000 = .117

(ii) No. of wage earners earning wages over 55 is = 30 + 16 + 7 = 53

Hence, (p) 53/1000 = .053

(iii) No. of wage earners earning was between 45-50 or 35-40 is

230 + 108 = 338.

Hence (p) = 338/1000 = .338

(b) The manufactures of 'Bajaj' scooter give choice to their customers to have either a double seated scooter or a single seated scooter. On analysis

of the booked orders for those scooters, they find that 75% of their customers are men and 25% women. 80% of the men customers prefer double seated scooters and rest one seated. 90% of their women customers prefer one seated scooters and rest two seated scooters. In what proportion, the manufacturers should manufacture these two scooters?

Solution

Men customers preferring tow seater	= .75 x .8 = .600
Women customers preferring two seater	= .25 x .1 = .025
Total	= .625
Men customers preferring one-seater	= .75 x .2 = .150
Women customers preferring one-seater	= .25 x .9 = .225
Total	= .375

Ratio .675 : .375 or 5 : 3

(c) In a sample of 100 radios it was found that :

No. of defects	0	1	2
No. of Radios	10	85	5

What is the probability that a radio selected at random will have zero defect?

Solution

10/100 = 1/10 or 0.1 0r 10%

- **Subjective Approach to Probability**

The subjective or personalistic approach to probability is of recent origin. According to this concept, the probability of an event is the degree of belief or degree of confidence placed in the occurrence of an event by a particular individual based upon the evidence available to him. This evidence may consist of relative frequency of oc-

currence of data or any other quantitative or qualitative information. To forecast the demand, predicting price etc. is done on the basis of subjective probabilities. The probability is determined between 0(0 = impossible) and 1(1 = certain event).

Illustration 4.4

(a) A job applicant assigns probabilities as follows:

The probability, p(A), of being offered a job at a company. A is 0.6; the probability, p(B); of being offered a job at a company B is 0.5; the probability of being offered a job at both companies is 0.4. What, consequently, is the probability of being offered a job with at least one of the two companies?

Solution

	Given				Total Completed		
	B	B'	Total		B	B'	Total
A	0.4	—	0.6	A	0.4	0.2	0.6
A'	—	—	—	A'	0.1	0.3	0.4
Total	0.5	—	1.0	Total	0.5	0.5	1.0

The probability of his getting job at least in one of the companies, i.e.;

$$p(AB+A'B+AB') = 0.4+0.1+0.2 = 0.7.$$

(b) You have noticed that your officer is happy on 60 percent of your calls, so you assign a probability of his being happy on your visit as 0.6 or 6/10. You have noticed also that if he is happy, he accedes to your requests with a probability of 0.4 whereas if he is not happy, he accedes to the requests with a probability of 0.1. You call one day and he accedes to your request. What is the probability of his being happy?

Solution

Probability Table

	Happy	Not happy	Total
Request Accepted	.24	.04	.28
Request not Accepted	.36	.36	.72
Total	.6	.4	1

(p) of being happy and accepting the request = .6 x .4 = 0.24

(p) of not being happy and accepting the request = .4 x .1 = .04

The chances of his accepting the request = 0.28

and the chances of his accepting the request when he is happy = .24

Hence, the probability of his being happy having accepted the request = .24/.28 or 6/7 or 0.857.

- **Modern Approach to Probability**

In this approach no precise definition of probability is given, but the theory is based on certain axioms or postulates. The axioms are:

To every event A, there corresponds a real value P(A) called probability of the happening of the event A, which satisfies the following three axioms:

(i) 07 P(A)7 1:

(ii) P(S) = 1, P(∅) = 0 S = certain event
∅ = impossible event

(iii) If A_1, A_2…… An are mutually exclusive events, then

(A_1, or A_2 or…….or A_n) = $P(A_1) + P(A_2) + …. + P(A_n)$ in particular,

P(A) = The sum of the probabilities of simple event, comprising the event A = Number of sample points in A / Total no. of sample points in S

4.4 Probability Theorems for Problems Solving

The solution to many problems involving probabilities requires a thorough understanding of some of basic rules that govern the manipulation of probabilities. They are generally called probability theorems. Let us discuss them in detail:

(1) Addition Theorem: The theorem is defined as follows:

"If two events are mutually exclusive and the probability of the one is p_1 while that of the other is p_2, the probability of either the one event or the other occurring is the sum p_1+p_2"

Proof of the Theorem: If an event A can happen in 'a 1' ways and B in 'a 2' ways, then the number of ways in which either event can happen is 'a 1 +a 2'. If the total number of possibilities is 'n', then by definition the probability of either the first or the second event happening is

$$a\ 1 + a\ 2/n = a\ 1/n + a\ 2/n$$

Since $a\ 1/n = P(A)$ and $a\ 2/n = P(B)$

Hence $P(AorB) = P(A) + P(B)$.

The theorem can be extended to three or more mutually exclusive events. Thus

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

For example, the probability of getting spot (1) in a throw of a single die is 1/6, the probability of getting spot (3) is also 1/6 and the probability of getting spot (5) too is 1/6. The probability of getting and odd number (1, 3 and 5) in a throw of a single die will be the addition of their respective probabilities, that is, 1/6+1/6+1/6 = 3/6 or 1/2.

The addition theorem will hold good only if:

(i) Items are mutually exclusive,

(ii) Mutually exclusive items belong to same set.

Illustration 4.5

(a) A bag contains 4 white, 2 black, 3 yellow and 3 red balls. What is the probability of getting a white or red ball at random in a single draw of one.

The probability of getting one white ball = 4/12

The probability of getting one red ball = 4/12

The probability of one white or red ball = 4/12+3/12 = 7/12 or 7/12x100 = 58.3%.

(b) A card is drawn at random from an ordinary pack of 52 playing cards. Find the probability that a card drawn is either a spade or the ace of diamonds.

The probability of drawing a spade = 13/52.

The probability of drawing and ace of diamonds = 1/52.

Probability of drawing a spade or an ace of diamond = 13/52 + 1/52 = 14/52

A total of 7 can come in 6 different ways (1/6, 2/5, 3/4, 4/3, 5/2, 6/1)

A total of 11 can come in 2 different ways (5/6, 6/5)

The probability of getting a total of 7 = 6/36 or 1/6

The probability of getting a total of 11 = 2/36 or 1/18

The probability of getting either 7 or 11= 1/6 + 1/18 = 4/18 or 2/9.

The addition theorem will hold good only if the events are mutually exclusive. If events contain no sample point in common, then some adjustment is necessary under such a case :

p [(A) or (B)] = p (A)+p(B)-p(A and B)

The following example will make it clear.

A bag contains 25 balls, numbered from 1 to 25, one is to be drawn at random. Find the probability that the number of the drawn ball will be a multiple of 5 or 7.

The probability of the number being multiple of 5 (5, 10, 15, 20, 25) = 5/25.

The probability of the number being multiple of 7 (7, 14, 21) = 3/25

Thus the probability of the number being a multiple of 5 or 7
will be = 5/25+3/25= 8/25.

In the above illustration, find the probability that the number is a multiple of 3 or5:
The probability of the number being multiple of 3 (3,6 9, 12, 15, 18,21, 24) = 8/25
The probability of the number being multiple of 5 (5, 10, 15, 20, 25) = 5/25.
Joint probability 8/25+5/25 = 13/25; but this answer is wrong, because item No. 15 is not mutually exclusive. Hence the correct probability will be

= 8/25+5/25-1/25 = 12/25.

Hence, p (A+B) + p (A) + p(B)-p(AB). The following diagram will make it clear.

Similarly, when three events are not mutually exclusive, then:
p(A+B+C) =p(A)+p (B)+p (C) -p (AB) -p (AC)-p (BC)+p (ABC)

Illustration 4.6

What is the probability of drawing a black card or a king from a pack of ordinary playing cards?

Number of black cards		number of kings		number of black kings
26	+	4	-	2

Hence, (p) = 26/52+4/52-2/52 = 28/52.

It is also essential that mutually exclusive items must belong to the same set. To illustrate this point, let us look at the following example:

Suppose the probability of a man dying between his 40th and 41st birth days is 0011, and the probability of his marrying between his 41st and 42nd birthdays is 0.009. These events are mutually exclusive but it cannot be said that the probability of a man dying in his 40th year and of marrying in his 41st year is .011+.009 = .02. These two events do

not belong to the same set.

(2) Multiplication Theorem: According to this theorem. "If two events are mutually independent, and the probability of the one is P_1 while that of the other is P_2 the probability of the two events occurring simultaneously is the product of P_1 and P_2". For example, the probability of head coming up in a toss of a coin is 1/2 and the probability of 4 coming in a throw of a die is 1/6. If a coin and a die are thrown together, the probability of head coming up in the toss of coin and 4 coming up in the throw of a die will be 1/2x1/6 =1/12.

Proof of the Theorem: If an event A can happen in 'n 1' ways of which 'a 1' are successful and the event B can happen in 'n 2' ways of which 'a 2' ways are successful, we combine the successful events of both A and B events where the total number of successful happening is 'a 1 x a 2'. Similarly, the total number of possible cases is 'n1x n2'. Then by definition, the probability of occurrence of both event is

a 1 x a 2/n 1 x n 2 = a 1/n 1 x a 2/n 2

Since a 1/n 1 = P(A) and a 2/n 2 = P(B)

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

The theorem can be extended to three or more independent events. Thus

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

(a) What is the probability of throwing two 'fours' in two throws of a die?

The probability of a 'four' in first throw = 1/6

The probability of a 'four' in second throw = 1/6

The probability of two 'fours' = 1/6 x 1/6 = 1/36.

(b) What is the probability of getting all the heads in four throws of a coin?

The chance of getting head in the Ist throw = 1/2

The chance of getting head in the 2nd throw = 1/2

The chance of getting head in the 3rd throw = 1/2

The chance of getting head in the 4th throw = 1/2

Thus the probability of getting heads in all the throws;

=1/2x1/2x1/2x1/2 = 1/16.

(c) Suppose it is 9 to 7 against a person A who is now 35 years of age lying till he is 65 and 3 to 2 against a person B, now 45 living till he is 75; fund the chance that one at least of these persons will be alive 30 years hence.

The chance that A will die within 30 years is 9/16 and the chance that B will die within 30 years is 3/5. The events are independent, therefore, the chance that both will die is 9/16 x 3/5 = 27/80.

Then chance that both will not be dead i.e., at least one will be alive is 1 - 27/80 = 53/80.

(d) A problem in statistids is given to three students A, B, C, whose chances of solving it are 1/2, 1/3, 1/4 respectively. What is the probability that the problem will be solved?

Probability that student A will fail to solve the problem = 1 - 1/2 = 1/2

Probability that student B will fail to solve the problem = 1 - 1/3 = 2/3

Probability that student C will fail to solve the problem = 1 - 1/4 = 3/4

Since the events are independent, the probability that all the students A, B, C will fail to solve the problem = 1/2 x 2/3 x 3/4 = 1/4.

So, the probability that the problem will be solved = 1 - 1/4 = 3/4.

This problem can also be solved in the following way :

	Condition	**Probability**
(i)	A solves, B solves C solves	= 1/2 x 1/3 x 1/4 = 1/24
(ii)	A solves, B solves, C fails to solve	= 1/2 x 1/3 x 3/4 = 3/24
(iii)	A solves, B fails to solve, , C solves	= 1/2 x 2/3 x 1/4 = 2/24
(iv)	A fails to solve, B solves, C solves	= 1/2 x 1/3 x 1/4 = 1/24

(v) A solves, B fails to solve, C fails to solves = 1/2 x 2/3 x 3/4 = 6/24

(vi) A fails to solves, B solves, C fails to solve = 1/2 x 1/3 x 3/4 = 3/24

(vii) A fails to solves, B fails to solve, C solves = 1/2 x 2/3 x 1/4 = 2/24

(viii) A fails to solves, B fails to solve, C fails to solve = 1/2 x 2/3 x 3/4 = 6/24.

The problem is solved in all the conditions, except that of (viii). If the probabilities of (i) to (vii) are added, that will give the probabilities of problem being solved. The total comes to 18/24 or 3/4.

Illustration 4.7

A Helicopter is equipped with three engines that operate independently. The probability of an engine failure is 0.01. What is the probability of successful fight if only one engine is needed for the successful operation of the aircraft?

Solution

Since the flight is unsuccessful only when all the three engines fail, then the probability of unsuccessful flight is:

.01 x .01 x.01 = .000001.

The probability of successful flight = 1 - .000001 = .999999.

Illustration 4.8

Five cards are to be drawn in succession and without replacement from an ordinary deck of playing cards.

(a) What is the probability that there will be no ace among the five cards drawn?

(b) What is the probability that the first three cards are aces and the last two cards are kings?

(c) What is the probability that only the first three cards are aces?

(d) What is the probability that an ace will appear only on the fifth draw?

Solution

(a) The probability that there will be no ace among the five cards:

p = 48/52 x 47/51 x 46/50 x45/49 x 44/48 = 205476480/311875200.

(b) The probability that the first three cards are aces and the last two cards are kings :

p = 4/52 x 3/51 x 2/50 x4/49 x 3/48 = 288/311875200.

(c) The probability that only the first three cards are aces:

p = 4/52 x 3/51 x 2/50 x48/49 x 47/48 = 54144/311875200.

(d) The probability that an ace will appear only on the fifth draw:

p = 48/52 x 47/51 x 46/50 x45/49 x 4/48 = 18679680/31875200.

The multiplication theorem will hold good only if the events belong to the same set. In order to show the importance of this fact, Moroney in his book "facts from Figures" gives an interesting example. He observes, "Consider the case of a man who demands the simultaneous occurrence of many virtues of an unrelated nature in his young lady. Let us suppose that he insists on a Grecian nose, platinum-blonde hair, eyes of odd colours - one blue, one brown, and finally a first class knowledge of statistics. What is the probability that the first lady he meets in the street will put ideas of marriage into his head? It is difficult to apply multiplication theorem in this case, because events do not belong to the same set.

(3) Conditional Theorem : If sub-event are not independent, and the nature of dependence is known, we have the theorem of conditional probabilities. This theorem is more or less corollary of the multiplication theorem. The theorem is that the probability that both of two dependent sub-events can occur is the product of the probability of the first sub-event and the probability of the second after the first sub-event has occurred. In notation p(A and B) = p (A) x p(B/A) is the conditional probability of B when A has already happened. For example, if out of a pack of cards shuffled or

each time 'king' turns out first and the card is not restored, then in a second reshuffling the probability of 'king' turning up again - 4/52 x 4/51 = 12/2652, since there are 4 kings at the first shuffle of 52 cards and 3 kings only at the second shuffle of 51 cards. The term 'condition probabilities' is often known as probabilities due to partial exhaustion of a sample space.

(4) Bayes' Theorem: Probabilities can be revised when new information pertaining to a random experiment is obtained. The notion of revising probabilities is a familiar one, for all of us, even to those with no previous experience in calculating probabilities have lived in an environment ruled by whims of chance and have made informal probability judgements. We do also intuitively revise these probabilities upon observing certain facts and change our actions accordingly. Our concern for revising probabilities arises from a need to make better use of experimental information. This is referred to as Bayes' Theorem after the Reverend Thomas Bayes, who proposed in the eighteenth century, that probabilities be revised in accordance with empirical findings.

Quite often the businessman has the extra information on a particular event or proposition, either through a personal belief or from the past history of the event. Probabilities assigned on the basis of personal experience, before observing the outcomes of the experiment are called prior probabilities. For example, probabilities assigned to past sales records, to past number of defectives produced by a machine, are examples of prior probabilities. When the probabilities are revised with the use of Bayes; rule, they are called posterior probabilities. Bayes' theorem is very useful in solving practical business problems in the light of additional information.

Suppose, a random experiment having several mutually exclusive events E_1, E_2…… and the probabilities of each event $P(E_1)$, $P(E_2)$……. have been obtained. These probabilities are referred to as prior probabilities, because they represent the chances that events before the results from empirical investigation are obtained. The investigation itself may have several possible outcomes, each statistically dependent upon Es.

For any particularly result which we may designate by the letter R, the conditional probabilities $P(R/E_1)$, $P(R/E_2)$……are often available. The result itself serves to revise the event probabilities upward or downward. The resulting values are called posterior probabilities since they apply after the information result has been learned. The posterior probability values are actually conditional probabilities of the form $P(E_1/R)$, $P(E_2/R)$ that may be found according to Bayes' Theorem. The posterior probability of E, for a particular result R of an empirical investigation may be found from:

$$P(E_2/R) = P(E_2)\, P(R/E_1)/ [P(E_1)\, P(R/E_1) + P(E_2)\, P(R/E_2)\text{……..}]$$

Illustration 4.9

Box 1 contains three defective and seven non-defective items, and Box 2 contains one defective and nine non-defective items. We select a box at random and then draw one item fromthe box.

(a) What is the probability of drawing a non-defective item?

(b) What is the probability of drawing a defective item?

(c) What is the probability that box 2 was chosen, given a defective item is drawn?

Solution

$P(B_1)$ = Probability that box 1 is chosen = 1/2,

$P(B_2)$ = Probability that box 2 is chosen = 1/2,

P(D) = Probability that a defective item is drawn,

P(ND) = Probability that a non-defective item is drawn.

(a) P(ND) = P (Box 1 and non defective) + P(Box 2 and non defective)

= (1/2 x 7/10) + (1/2 x 9/10) = 16/20

(b) P(D) = P (Box 1 and defective) + P(Box 2 and defective or

= P(D) = (1/2 x 3/10) + (1/2 x p/10) = 4/20

(c) By Bayes' theorem

$P(B_1/D) = PB_1$ and D) /P(D) = 3/20 / 4/20 = 3/4.

$P(B_1)$ and $P(B_2)$ are called prior probabilities and $P(B_1/D)$ and $PB_2/D)$ are called posterior probabilities. The above information is summarized in the following table:

Event	Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
B_1	1/2	3/10	3/20	3/4
B_2	1/2	1/10	1/20	1/4
Total			**4/20**	**1.0**

Illustration 4.10

A box contains four fair dices and one crooked die with a loaded weight which makes the six-face appear on two-thirds tosses. You are asked to select one, die at random and toss it. If the crooked die is indistinguishable from the fair die and the result of your toss is a six-face; what is the probability that you tossed the crooked die?

Solution

P (Fair dice) = 4/5

P (Crooked die) = 1/5

The probability of tossing a six face in a fair die = 1/6

The probability of tossing a six face in a crooked die = 2/3

The probability of tossing a six face when die is crooked = 1/5 x 2/3 = 2/15.

The probability of tossing a six face when die is fair = 1/5 x 1/6 = 4/30.

The probability of tossing a six face = 2/15 + 4/30 = 4/15

The posterior probability that die tossed is crooked is : 2/15 / 4/14 = 1/2 or 50%.

Illustration 4.11

Urn A_1 contains 8 black and 2 white marbles. Urn A_3 contains 3 black and 7 white marbles, and urn A_3 contains 5 white and 5 black marbles. A fair die is to be cast. If the

die turns up 1, 2 or 3 then a marble will be selected from A1. If the die turns up 4 or 5 a marble will be selected from A_2. Finally, a marble will be selected from A_3. If the die turns up 6. Given that the marble selected is black, what is the probability that the marble was from urn A_2?

Solution

Probability of marble being chosen from urn $A_1 = 3/6$

Probability of selecting a black marble from $A_1 = 8/10$

Hence, the joint probability of 1, 2 or 3 coming up in the fair die and then drawing a black marble from $A_1 = 3/6 \times 8/10 = 24/60$.

Similarly, the probability of 4 or 5 turning up in the die and drawing a black marble from urn $A_2 = 2/6 \times 3/10 = 6/60$

and the probability of 6 turning up in the die and drawing a black marble from urn $A_3 = 1/6 \times 5/10 = 5/60$

The probability of drawing a black marble from any of these urn is

$$24/60 + 6/60 + 5/60 = 35/60.$$

Assuming that the marble selected is black, the probability that the marble was chosen from urn A2 is : P 6/60 / 35/60 = 6/35.

Illustration 4.12

Urn A contains 6 green and 4 red marbles, and urn B contains 2 green and 7 red marbles. A marble is to be selected at random from A and placed in B. One marble is then selected from B. Given that the marble selected from B is green, what is the probability that the marble selected from A will also be green?

Solution

Probability of marble selected from B is green, if the marble selected from A and placed in B is green = $6/10 \times (2+1)/(9+1) = 6/10 \times 3/10 = 18/100$. Probability of

marble selected from B is green; if the marble selected from A and placed in B is red

= 4/10 x 2/9+1 = 4/10 x 2/10 = 8/100.

The joint probability of green marble selected from B = 18/100 + 8/100 = 26/100.

The probability, given that the marble selected from B is green, the marble selected from A will also be green.

=18/100 / 26/100 = 18/26.

Illustration 4.13

In a factory, machines M_1, M_2 and M_3 manufacture respectively, 30, 30 and 40 percent of the total output. Of their output 1, 3, and 2 percent are defective items. An item is drawn from day's output and is found defective. What is the probability that it was manufactured by M_1 by M_2, by M_3?

Solution

(P) that an item is manufactured by M_1 = 30/100 = .3, and the item is defective:

.3 x .01 = .003.

(P) that an item is manufactured by M_2 = 30/100 = .3, and the item is defective:

.3 x .03 = .009.

(P) that an item is manufactured by M_3 = 40/100 = .4, and the item is defective:

.4 x .02 = .008.

Probability of defective item = .003+.009+.008 = .02

Probability that the defective item is manufactured by M_1 = .003/.02 = or 3/20.

Probability that the defective item is manufactured by M_2 = .009/.02 = or 9/20.

Probability that the defective item is manufactured by M_3 = .008/.02 = or 8/20.

Illustration 4.14

A can hit a target 3 times in 5 shots, B 2 item in 5 shots, C 3 times in 4 shots.

They fire a volley. What is the probability that 2 shots hit?

Solution

Fire a volley means that A, B and C all try to hit the target simultaneously. Twp shots hit the target in one of the following ways:

(a) A and B hit and C fails to hit.

(b) A and C hit and B fails to hit.

(c) B and C hit and A fails to hit.

The chance of hitting by A = 3/5 and of not hitting by him = 1- 3/5=2/5

The chance of hitting by B = 2/5 and of not hitting by him = 1- 2/5=3/5

The chance of hitting by C = 3/4 and of not hitting by him = 1- 3/4=1/4

The probability of (a) = 3/5 x 2/5 x 1/4 = 6/100

The probability of (b) = 3/5 x 3/4 x 3/5 = 27/100

The probability of (c) = 2/5 x 3/4 x 2/5 = 6/100

Since (a), (b) and (c) are mutually exclusive events, the probability that two shots hit

6/100 + 27/100 + 12/100 = 45/100 = 9/20 or 9/20 x 100 = 45%

The classical or a prior probability measures have two very interesting characteristics. First, the objects referred to as fair coins true dice or fair deck of cards are abstractings in the sense that no real world object exactly possesses the features postulated. Secondly, in order to determine the probabilities, no coins had to be tossed, no dice rolled nor cards shuffled. That is no experimental data were required to be collected; the probability calculations were based entirely on logical prior (thus a priori) reasoning.

It may be possible that the results of a few trials of an experiment may be different than the expected on the basis of probability. If a coin is tossed 10 times, it may be that head may turn up 7 times and tail 3 times whereas, according to the prior probability the head should turn 5 times and tail also 5 times. But in 500 or 1000 trials, the results may be much nearer to the probable results.

4.5 Permutation and Combinations in the Theory of Probability

Knowledge of permutaions and combinations is essential to solve the problems related to probability determination. So, we have discussed these concepts hereunder:

Permutations: Sometimes we are interested in the total number of different ways in which items can be arranged so that the order of components is important, yet no two arrangements are similar. Arrangements of this sort are called permutations. For example, if seven alphabets - A, B, C, D, E, F, G, are to be arranged by taking two letters at a time, but under no circumstances may an arrangement contain the same 2 letters (like AA, or BB etc.) then the following permutations are possible:

AB	AC	AD	AE	AF	AG
BA	BC	BD	BE	BF	BG
CA	CB	CD	CE	CF	CG
DA	DB	DC	DE	DF	DG
EA	EB	EC	ED	EF	EG
FA	FB	FC	FD	FE	FG
GA	GB	GC	GD	GE	GF

Hence, there are 7 x6 = 42 permutations,

Thus, following formula can give the number of permutations

$$\text{Perm.} = n(n - 1)$$

If 26 letters are to be arranged in this manner, the total number of ways will be 26 (26 - 1) = 650.

The permutation can be shown in a tree-diagram also. For example, three chairs x, y and z can be arranged n(n - 1)(n - 1) or 3(3 - 1)(3 - 2) = 6 ways.

The tree diagram shows this:

TREE DIAGRAM

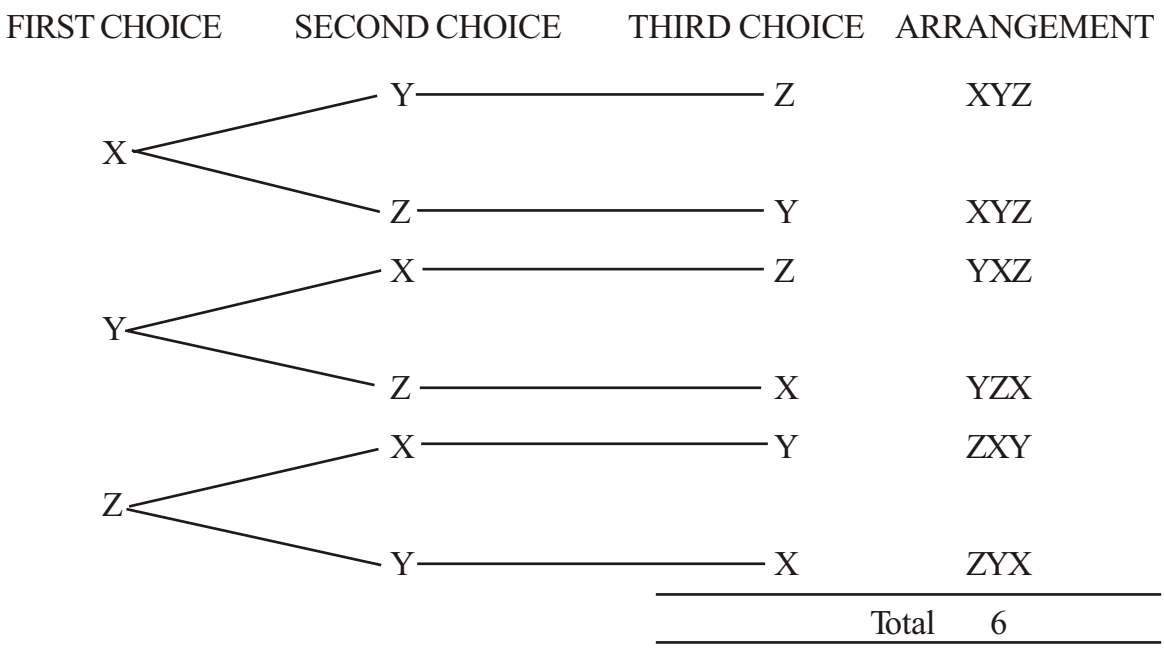


Illustration 4.15

If a man has the choice of traveling between Hisar and Delhi by 8 trains, in how many possible ways he can complete the return journey, using a different train in each direction?

Solution

For the outward journey he has the choice of using all the 8 trains. Having completed the outward journey, he will be left with only 7 trains to complete the return journey. Thus, the total number of ways in which he can complete the journey are 8(8 - 1) = 56.

Some general rules regarding permutation are as follows:

For finding permutation of doing 'n' function in 'r' ways, the formula is

Prem. = n(n -1) (n - 2) x (n - 3) x (n - 4)….

This formula can also be written as follows:

P = n !*/(n - r) !

Illustration 4.16

(a) If a person is given one cup of coffee of each of 5 brands and asked to rank these according to preference. How many possible ranking can there be?

$n P_r = n ! / (n - r) !$ or $5 !/(5 -5) = 5 \times 4 \times 3 \times 2 \times 1/1 = 120$

It can also be calculated:

Prem. $= n(n - 1) (n - 2) (n - 3) (n - 4) = 5 \times 4 \times 3 \times 2 \times 1 = 120$

(b) There are six doors in a room. Four persons have to enter it. In how many ways they can enter from different doors?

$n(n - 1) (n - 2) (n - 3)$ $\quad n = 6, r = 6$

$=(6 - 1) (6 - 2) (6 - 3)$ $\quad {}^{4}p_4 = n !/ (n - r) !$

$= 6 \times 5 \times 4 \times 3 = 360$ $\quad = 6!/(6-4)!\ 6x5x4x3x2x1/2x1 = 6x5x4x3 = 360$

(c) In how many ways first, second and third prizes can be distributed to three of 10 competitors?

$n = 10, r = 3$

$p_3 = 10 ! / (10 - 3) = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 / 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 10 \times 9 \times 8 = 720$ ways.

(d) Four strangers board a train in which there are 6 empty seats. In how many different ways can they be seated?

$n = 6, r = 4$

${}^{6}p_4 = n !/ (n - r) !/ 6 !/ (6 - 4) ! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 / 2 \times 1$

$= 6 \times 5 \times 4 \times 3 = 360$ ways

(e) In how many ways can 12 seats be occupied by 6 women?

$n = 12, r = 6$

${}^{13}p_4 = n!/(n - r) ! = 12 !/ (12 - 6) ! = 12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 / 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 12 \times 11 \times 10 \times 9 \times 8 \times 7 = 665280.$

(f) How many of officers can we possibly have if we have a set of 10 persons and

have to different sets fill the offices of chairman, Dy. Chairman, Secretary and Treasurer.

p_4 = 10 ! / (10 - 4) = 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 / 6 x 5 x 4 x 3 x 2 x 1= 5040.

If there are n things in which p are of one kind, q are of the other kind and r of a third kind, then the number of permutations will be:

Prem. = n ! / P x q x r

Illustration 4.17

(a) In how may ways 12 students of MBA (DE) be allotted to three tutorial groups of 2, 4 and 6 respectively?

n = 12, p = 2, q = 4, r = 6

n ! / p !q!r! = 12 ! / 2!4! 6 = 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) (4 x 3 x 2 x 1) (6 x 5 x 4 x 3 x 2 x 1) = 12 x 11 x 10 x 9 x 8 x 7 / 4 x 3 x 2 x 1 x 2 x 1 = 13860.

(b) In how may ways can the letters of the word 'BASKET' be arranged?

There are 6 letters.

Hence Perm. = 6 ! = 6 x 5 x 4 x 3 x 2 x 1 = 720 ways.

(c) how many ways can the letter of the words 'BETTER' be arranged?

In this letter E comes twice. 'T' comes twice, 'B' and 'R' come only once.

Hence, Perm. = n ! / p !q!r! = 6 ! / 2!2!1 = 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) x (2 x 1) x 1 = 720/ 4=108.

(c) In how many ways can the letters of the words 'MACMILLAN', 'BANANA' AND 'STATISTICALLY' can be arranged?

'MACMILLAN' N= 9

A comes 2 times

M comes 2 times

L comes 2 times

others come only once

Permutation = 9 ! / 2! 2! ! ! ! ! = 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1/ (2 x 1) (2x 1) (2 x 1) = 15120.

'BANANA' n = 6 A = P = 3

N = q = 2

Permutation = n ! p!q! = 9 ! / 2! 2! 2! ! ! ! ! ! ! = 6 x 5 x 4 x 3 x 2 x 1/ (3 x 2 x 1) (2x 1) = 60.

'STATISTICALLY' n = 13 T = p = 3

S = q = 2

A = r = 2

I = 1 = 2

L = k = 2

others come only one:

Permutation = n ! / p!q!r!l!k! = 13 ! / 3! 2! 2! 2! 2! = 13 x 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1/ (3 x 2 x 1) (2 x 1) (2 x 1) (2 x 1) (2 x 1) = 64864800.

4.6 Summary

The word probability or chance is very common in day to day life of human being. Probability is especially important in statistics because of the many principles and procedures that are based on this concept. Indeed, probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phrases like: 'You had better take an umbrella because it is likely to rain.' 'His chances of winning are pretty small.' It is very likely that it may rain by the evening. You are probably right.' or 'There are fifty-fifty chances of his passing the examination.' In each of these phrases an idea of uncertainty is acknowledged. Goethe remarked that, "There is nothing more frightful than action in ignorance." Reasoning in terms of probabilities is one weapon by which we attempt to reduce this uncertainty or ignorance. The use of word 'probability' in statistics, however is somewhat different. It is more precise than what it means in popular usage. In statistics, a probability is a numerical value that measures the uncertainty that a particular event will occur. There are basically three methods of

measuring probabilities. They represent differentconceptual approaches. They are: Classical Approach, Empirical Approach, Subjective Probability. and Modern Approach.

4.7 Self- Assessment Exercise

1. Define probability and explain the importance of this concept in statistics.
2. Explain what do you understand by term 'probability'. State and prove the addition and multiplication theorems of probability.
3. Explain the concept of independence and mutually exclusive events in probability. State theorems of total and compound probability.
4. Define probability and enunciate the Multiplication Law of probability, giving suitable examples.
5. What are the different schools of thought on the interpretation of probability? How does each school define probability ? Explain with suitable examples.
6. (a) When are the events said to be independent in the probability sense? Give examples of dependent and independent events.

   (b) Differentiate between the circumstances when the probabilities of two events are : (i) added, and (ii) multiplied.
7. (a) If A and B are events, define the compound events: A+B, (i.e., the union of the two events and A x B, i.e., the intersection of the two events. Prove that

   $$p(A+B)= p(A)=p(B) - p(AB)$$

   and establish a similar rule for p(A+B+C). State the general result for n such events.

   (b) When a soldier fires at target, the probability that he hits the target is 1/2 for soldier A, 1/2 for soldier B, 2/3 for soldier C, and 1/12 for soldier D. If all the four soldiers A, B, C and D fire at the target simultaneously, calculate probability that the target is hit by some one or more.

   [(b) 373/648.

8. Explain why there must be mistake in the following statement:
"A quality control engineer claims that the probability that a large consignment of glass bricks contains 0, 1, 2, 3, 4 or 5 defectives are .11, .23, .37, .16, .09 and .05 espectively.
[The total of probabilities of all mutually exclusive events cannot exceed 1. Here it is 1.01].

9. One bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black ball. If one ball is drawn from each bag, find the probability that (a) both are white, (b) both are black, and (c) one is white and one is balck. [(a) 1/4, (b) 5/24, (c) 13/24.

10. A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit 2 out of 3 shots. Find the probability of the target being hit at all when they both try.
[11/12]

11. The odds are 7 to 5 against A, a person who is now 30 years old living till he is 70 years and the odds are 2 to 3 in favour of B who is now 40 years of age living till he is 80 years. Find the chance that one at least of these two persons will be alive 40 years hence.
[13/20]

12. It is given that in two towns the number of rainy days in a year are respectively 20 and 30. What is the probability that on a particular day in that year there is (a) no rain in both towns, (b) rain in one town only, (c) rain in both towns.
[(a) 69/73 x 67/73, (b) 4/73 + 6/73, (c) 4/73 x 6/73]

4.8 Suggested Readings

1. Hooda, R P : Statistics for Business and Economics
2. Gupta, B N : Statistics Theory and Practice
3. Gupta, S P : Statistical Methods
4. Kappor, V K : Statistics Theory, Methods and Applications
5. Bhardwaj, R S: Business Statistics

❄ ❄ ❄

Lesson : 11

PROBABILITY DISTRIBUTION

Author : **Vetter:**

Dr. B. S. Bodla **Dr. R. K. Mittal**

The objective of this lesson is to make students familier with behaviour of the variable whose values depend upon chance factor. In this lesson we shall describe three probability distributions named Binomial, Poisson and and Normal distribution. All these distributions are widely used to explain the behaviour of business and econimic variables.

1.0 Random variable :

It is a numerical form of the outcomes of a random experiment. For instance, let us consider the case of tossing a fair coin twice, its sample space is (TT, HT, TH, HH). If we count the number of heads appear on the top, then number of heads could take up values 0, 1 and 2. Let us denote the number of heads with symbol X, then we may say X is a variable whose value is determined by chance. Such variables whose values are determined by chance are called random variables. Corresponding to each value of a random variable a probability is associated. In tossing a coin twice the different values of head associated with corresponding probability could be presented in following way :

X (Number of Heads)	Frequency	Probability = P (X)
0	1	.25
1	2	.50
2	1	.25

Here the probability of getting one head is 0.50. Thus with the help of a random variable we put the sample space into precise form so that it becomes easy to extract information about the nature of a random experiment. The above presentation of the values of a random variable with their corresponding probabilities is called a probability distibution. In real life we may have probability distribution for different types of random variables like defective items produced by machines, demand for a good, sales of a firm, number of children in a family etc. If a random variable has a definite distance from any value to the next possible value, the variable is called a' discrete random variable' and the resulting distribution is a 'discrete probability distribution'. In our experiment of tossing the coin twice, number of head is a dicrete random variable. So it has discrete probability distribution. The number of children in a family, number of defective items produced by machine, number of T.V. set sold out etc. are the examples of discrete variable. However, if a random variable can take up any value in some interval of value then it is called a 'continuous random variable' and the resulting distribution is termed as a 'continuous probability distribution'. This types of random variable can assume an infinitely large number of values. Examples of continuous random variables are income of a person, demand of a commodity, height of a person etc. Few examples to illustrate the probability distribution are given below :

IIlustration 1 Suppose a random variable X denotes the number of tails appearing in an experiment of tossing a coin five times.Then x can take value 0, 1, 2, 3, 4, 5. Then the probability distribution of X is given below :

X	Probability
(no. of tails)	P(X)
0	.0312
1	.1562

2	.3125
3	.3125
4	.1525
5	.312

IIustration 2 : Suppose number of items produces by a machine are inspected per hour frequently. Then the resultant probability distribution of a random variable X (number of derfective items produced by a machine per hour) is can be given as below :

X	Probability
(no. of tails)	P(X)
0	.10
1	.35
2	.30
3	.15
4	.08
5	.02
6	.00

Thus with the help of a random variable we put the information regarding an experiment into precise form i.e. we put the sample space and events into probability distribution. But there is still scope for futher precision of information of a random experiment. In illustration 2, the probability distribution of defective items produced by a machine is no doubt an important piece of information for decision making but the production manager may be interested to know the average number of defectives produced by the machine per hour. The average of the random variable in a probability distribution is called expected value. The expected value of a discrete random variable can be

found by multiplying each value that the variable can assume by the probability of occurrence of that value and then summing up the products. Suppose a random variable X assumes values $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ with corresponding probabilities $P_1, P_2 \ldots\ldots\ldots P_n$ respectively. Then average or expected value of x is

$$E(x) = \sum_{i=1}^{n} xipi$$

$$E(x) = x_1 . p_1 + x_2 . p_2 + \ldots\ldots\ldots\ldots\ldots\ldots + x_n p_n$$

IIlustration 3, Find the expected defective items produced by a machine given in illustration 2.

Solution : The expected value of the random variable X (no. of defective items) will be computed as :

$$E(X) = X_1 P_1 + X_2 P_2 + X_3 P_3 + X_4 P_4 + X_5 P_5 + X_6 P_6$$

$$E(X) = 0x.10 + 1x.35 + 2x.30 + 3x.15 + 4x .08 + 5x .02 = 1.82$$

It means under the same condition of production the machine will produce 1.82 defective items per hour in the long run

The another value of random variable which may also provide a useful information regarding a random experiment is variance. The variances of a random variable x can be computed as

$$\text{Var}(x) = \sum_{i=1}^{n} [xi - E(x)]^2 pi$$

The possible values of x are $x_1, x_2 \ldots\ldots\ldots\ldots\ldots\ldots, x_n$ and the corresponding probabilities are p1, p2 $\ldots\ldots\ldots\ldots\ldots p_n$.

Then variance of x is

$$V(x) = [x_1 - E(x)]^2 . p_1 + [x_2 - E(x)]^2 . P_2 + \ldots\ldots\ldots\ldots + [x_n - E(x)]^2 . P_n$$

If x is number of defective items produced by a machine in the above example, then its variance is :

$$V(x) = (0-1.82)^2 \text{ x.10} + (1-1.82)^2 \text{ x.35} + (2-1.82)^2 \text{ x.30} + (3-1$$

$$.82)^2 \text{ x.15} + (4-1.82)^2 \text{ x.08} + (5-1.82)^2 \text{ x.02} = 0.0024$$

2. Probability Distribution

In the previous section we have seen that a representation of all possible values of a random variable together with their probabilities of occurrence is called a probability distribution. As random variables can be either discrete or continuous, so we have probability distribution that are either discrete or continuous. The theoretical probability distribution indicate the likely behaviour of a random variable under the given conditions. With the help of theoretical probability distribution we may list the probilities for different values of a random variable without performing an experiment. For instance in an experiment of tossing a coin, we may assign the probabilities for different values of head, a random variable, with the help of theoretical probability distribution. Now we shall discuss three types of theoretical probability distributions namely (i) Binomial (ii) Poisson, and (iii) Normal distribution.

2.1 Binomial Distribution.

It is one of the widely used probability distribution of a discrete random variable. Binomial distribution describes a variety of processes of real world. It describes the behaviour of a discrete random variable resulting from an experiment known as a Binomial process, named after the seventeenth century Swiss mathematician Jacob Bernoulli.

Bernoulli process is described as follows :

(i) An experiment is performed under the same conditions for a fixed and finite number of trials, say , n.

(ii) Each trial has two mutually exclusive possible outcomes : heads or tails, yes or no, good or defective, success or failure. Generally the outcomes of Bernoulli process are called success and failure.

(iii) The probability of success P, remains fixed from trial to trial. Thus the probability of failure (1-P) will also remain fixed.

(iv) The trials are statistically independent that is, the outcome of one trial does not afect the outcome of any other trial.

A simple example for illustrating the binomial distribution is cointossing experiment in which a coin is tossed n times and the number of heads occurring in n tosses is recorded. If the outcomes of all trials are independent of one another and we define the random variable x to be the number of heads appearing in n trials, then x is binomial random variable. Here each trial leads to two mutually exclusive outcomes either success (head) or failure (tail). The probability of success in a trial is denoted by p and, therefore, the probability of failure q = (1-p). Suppose we want to find what is probability of getting x number of successes (head) in n trials. Then the following binomial probability function will be used :

$$P(x) = {}_{n}C_{x}\, q^{n-x}\, p^{x},\ x = 0, 1, 2, \ldots\ldots\ldots\ldots n$$

$$\text{Where } {}_{n}C_{x} = \frac{n!}{x!\ n - x!}, \text{ and}$$

p is probability of getting success in each trial and q is probability of getting failure

$$q = (1-p).$$

IIlustration 4 : Suppose it is known that 10% of the students in a class wear glasses. If 8 students are selected at random what is probability that

(i) two students wear glasses.

(ii) none of those selected wear glasses

(iii) all those selected wear glasses.

Solution : Here, the probability of a student wearing glasses would be 0.10. Now, q = 0.90 and n = 8. The desired probabilities will be calculated as follows :

(i) Probability that 2 students wear glasses

$$P(x = 2) = 8_{C_2} (.90)^6 (.10)^2 = .0744$$

(ii) Probability that none of those selected wear glasses.

$$P(x = 0) = 8_{C_0} (.90)^8 (.10)^0 = 0.430$$

(iii) Probability that all those selected wear glasses.

$$P(x = 8) = 8_{C_8} (.90)^0 (.10)^8 = .00000001$$

2.1.1 The Mean and the variance of Binomial Distribution

The mean and the variance of binomial random variable can be computed by using the following methods.

Suppose x is a discrete random variable The mean and variance of distribution are generated. It means its probabilities are derived by using the Binomial rules as follows :

$P(x) = n_{c_x} q^{n-x} p^x$ then

$$\text{Mean of } x = E(x) = \sum_{x=0}^{n} x.p(x) = \sum_{x=0}^{n} x.n_{c_x} p^x q^{n-x} = np$$

$$\text{Variance of } x = V(x) = \sum_{x=0}^{n} [x - E(x)]^2 \, p(x) = npq$$

Thus the standard deviation of binomial variate x is given by $\sqrt{npq}$.

IIlustration 5 : Suppose that 20% of the items produced by a production are defective, and a sample of 25 items is selected from the produced items, at random. What is expected number of defective items in the selected 25 items ? What is the standard deviation of defective items in 25 selected items.

Solution : Assuming that binomial distribution is valid in this experiment, we have n=25 and p= .20. Therefore, the expected number of defective in a sample of 25 items is given by

$$E(x) = np = 25 \times .2 = 5 \text{ defective items.}$$

The variance of defective items is

$$V(x) = npq = np(1-p) = 25 \times .2 \times .8 = 4$$

therefore, the standard deviation is $\sqrt{4} = 2$ defective items.

2..1.2Fitting of Binomial Distribution

By fitting a binomial distribution, we mean to find out the theoretical or expected frequencies for all values of variate x = 0, 1, 2 ....., for fitting it to the given data the following procedure is adopted :

i) Determine the values of 'p' and 'q' keeping in mind that $X = n_1$ and q = (1-p)

ii) Find the probability for all possible values of the given random variable applying the binomial probability function

$$P(x) = {}_{n}C_{x} \, q^{n-x} \, p^{x}, \; x = 0, 1, 2, \ldots\ldots\ldots\ldots;n$$

iii) Compute the expected frequencies for all possible values of the random variable by multiplying N (total frequency) with corresponding probability as worked out in (ii) above.

iv) The expected frequency so calculated constitute the fitted binomial distribution to the given data.

IIIustration 6 : Four coins are tossed 208 times. Number of heads observed at each throw is recorded and the results are as follows :

Number of heads	0	1	2	3	4
Frequency	5	48	112	35	8

Fit a binomial distribution to the given data.

Solution : In the given data

N = 208, n =4, p = ‰, q = ‰

The random variable (x) is the number of heads at a throw which can assume values as 0, 1, 2, 3, and 4. For fitting the binomial distribution the following table is developed.

No. of heads	Frequency	Probability	Expected Frequency
0	5	$4_{C_0}\, p^0\, q^4 = 1/116$	$208 \bullet 1/16 = 13$
1	48	$4_{C_1}\, p^1\, q^3 = 4/116$	$208 \bullet 4/16 = 52$
2	112	$4_{C_2}\, p^2\, q^2 = 6/116$	$208 \bullet 6/16 = 78$
3	35	$4_{C_3}\, p^3\, q^1 = 4/116$	$208 \bullet 4/16 = 52$
4	18	$4_{C_4}\, p^4\, q^0 = 1/116$	$208 \bullet 1/16 = 13$

2.2 POISSON DISTRIBUTION:

It is a discrete probability distribution named after a French mathematician Simean Denis Poisson (1781-1840). In case of Binomial distribution, it is possible to count the number of times an event is observed, i.e. n was precisely known. But there are certain situations where this may not be possible. This may happen when events of an experiment are rare and casual, we can say that successful events in the total event space are few i.e. the events like accidents on a road, defects in a product, goal scored at a football match, arrivals of customers at a shop etc. In these, we know the number of times an event occur but not how many times it does not occur. Obviously the total number of trials in regard to a given experiment are not precisely known. The poisson distribution is very suitable in case of such rare events. Poisson distribution may be obtained as a limiting case of Binomial probability distribution under the following conditions :

(i) n, the number of trials is indefinitely large i.e.n- $\rightarrow \infty$

(ii) p, the fixed probability of success for each trial is indefinitely small i.e. P-$\longrightarrow$0

(iii) np = λ is finite. It is average number of occurrences of events per interval/ trial or space. For instance, in case of accidents, it may be average number of accidents occurred per hour for a day, on a particular road.

According to poisson probability distribution, the probability of getting x number of success when np = λ is small or finite , then

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Where e is a mathematical constant having value equal to 2.71828.

IIIustration 7 : The probability that a person will positively respond to an advertisement for a product to be 0.1. What is probability that if the advertisement is observed by 20 persons (i) three will respond., (ii) two or more respond (iii) no body will respond.

Solution : Here n =20, p = .1

So np = 20x .1 = 2

Then

(i) Probability that 2 will respond

$$P(x=2) = \frac{2^2 \cdot e^{-2}}{2!} = 0.2706$$

(ii) Probability that two or more will respond

$$P(x \geq 2) = 1-[p(x=0) + P(x=1]$$

$$= 1- [p(x=0) - P(x=1]$$

$$= 1- \frac{2^0 \cdot e^{-2}}{0!} + \frac{2^1 \cdot e^{-2}}{1!}$$

$$= 0.594$$

(iii) Probability that no body will respond

$$P(x = 0) = \frac{2^0 \cdot e^{-2}}{0!}$$

$$= .1353$$

IIIustration 8 : Defects in yarn manufactured by a local textile mill can be approximated by a poisson distribution with mean of 1.2 defects for every 6 meters of length. If the length of 6 meters are to be inspected, find the probability of less than 2 defects.

Solution : Here = $\lambda = 1.2$

Then the probability of getting less than 2 success (defects) is

$P(x<2) = p(x=0) + p(x=1)$

=.06626

2.2.1 The mean and the Variance

The mean and the variance of the random variable, following the Poisson probability distribution, are given by

Mean = E (x) = λ

And

Variance = V(x) = λ

Thus, both mean and variance of a poisson probability distribution are equal.

2.2.2 Fitting of Poisson Distribution :

When a poisson distribution is to be fitted to the given data, then following procedure is adopted.

i) Determine the value of λ , the mean of the distribution.

ii) Find the probabilities for the possible values of the given random variable using poisson probability function.

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

x = 0, 1, 2, ...

iii) Compute the expected frequencies as follows :

N.P. (x)

iv) The result of the (iii) above is the fitted poisson distribution to the given data.

IIIustration 9: Letters were received in an office on each 100 days. Assuming the following data to form a random sample from a poisson distribution fit the distribution and calculate the exected frequecies.

Numbers of Letters (x)	0	1	2	3	4	5	6	7	8	9	10
No. of days (Frequency)	21	4	15	22	21	20	8	6	2	0	1

Solution : Here $\lambda = \sum f_i x_i / \sum f_i$

$= \quad 400/100 = 4.$

The expected frequencies are computed by

$$\frac{100 \times e^{-4} \times 4^r}{r!}$$

Where r = 0, 1, 2, 3, ... 10

On calculation, the frequencies come out to be :

1.83, 7.32, 14.64, 9.22, 19.52, 15.62, 10.41, 5.95, 2.975, 1.322 and 5.29

2.3 The Normal Distribution :

It is the most widely used probability distribution for continuous random variables. Several mathematicians were instumental in its development including the eighteenth century mathematician Karlo Gauss. In honour of his work, the normal distribution is often called Gaussian distribution. The normal distribution has a prominent place in statistics because of two main reasons (i) it comes closely with factors such as observed frequency distribution of many natural and physical measurements, variability of human output, variability of machine output, household consumption pattern of edible oil, and demand for fish or meat etc. (ii) the normal distribution has some properties that make it applicable to a great many situations in which it is necessary to make inferences by taking samples.

2.3.1 Characteristics of the Normal Probability Distribution :

If we plot the continuous random variable (x) belonging to normal probability distribution, the diagram as shown in Figure 1 emerges.

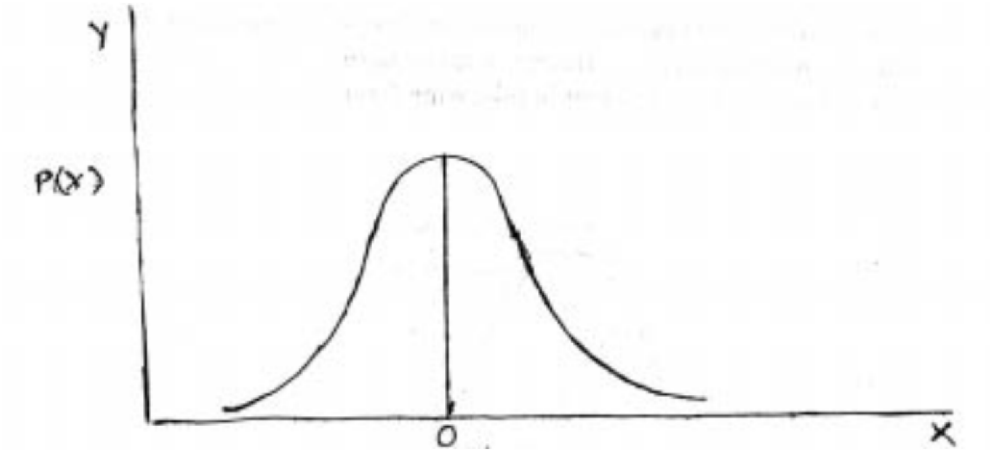


Fig. 1

In this figure we have taken the values of x on x-axis and probability of x on y-axis. From this diagram some important features of normal distribution emerge.

(i) This curve has a single peak, thus it is unimodal. It has the bell shape.

(ii) The mean of a normally distributed population lies at the center of its normal curve.

(iii) Because of the symmetry of the normal distribution the median and the mode of the distribution are also at the centre ; thus for a normal curve the mean, median and mode are the same value.

(iv) The two tails of the normal probability distribution extend indefinitely and never touch the horizental axis. The value of normal random variable lies between $-\infty$ to $+\infty$.

(v) In case of binomial distribution, to find the probability for its random variable we must have information on n (number of trials) and p (Probability of success). In case of poisson distribution it is λ .

The normal distribution requires information on the mean (μ) and the standard deviation (σ) which are also called its parameters, to find its probability distribution. So there is no single normal curve rather a family of normal curve. Different types of normal curves of different values of (μ) and σ are shown in following figures.

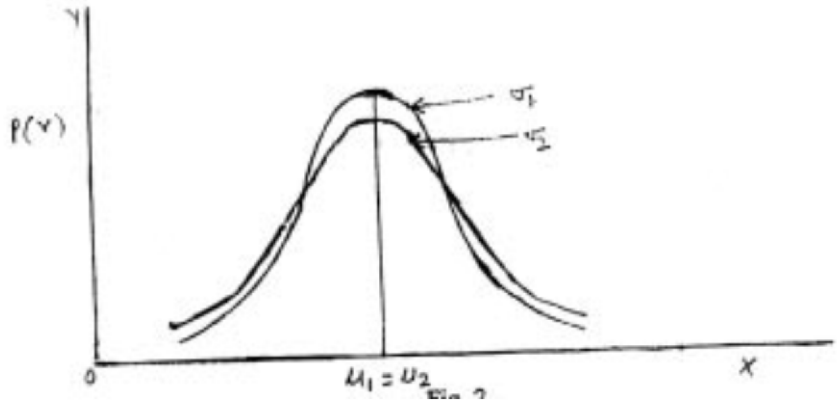


Fig. 2

Normal Distribution with identical mean but different standard deviations.

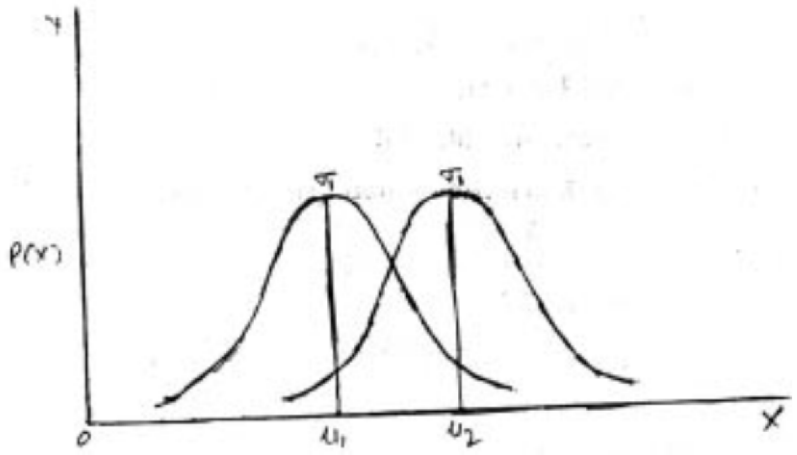


Fig. 3

Normal Distribution with different means but identical standard deviations.

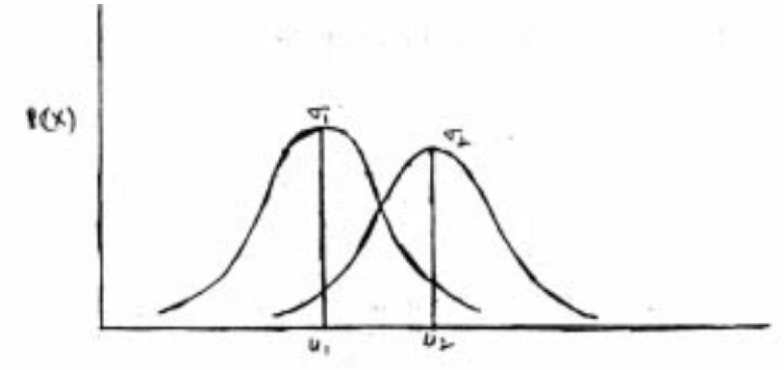


Fig. 4

Normal Distribution with different means and different standard deviations.

(vi) The normal probability distribution is continuous probability distribution, so we cannot find the probability for a single value of a random variable with the help of this distribution, instead we find the probability for an interval value of random variable.

(vii) No matter what the value of μ and σ are for a normal probability distribution, the total area under normal curve is 1.00. The normal distribution has the property that area between mean and standard deviation from the mean is same for all distributions, whatever be the value of the mean and the standard deviation. Thus $\mu \pm \sigma$ contains 68.3% of the area, $\mu \pm 2\sigma$ contains 95.5% of the area and $\mu \pm 3\sigma$ contains 99.7% of the area. It is shown in Fig. 5.

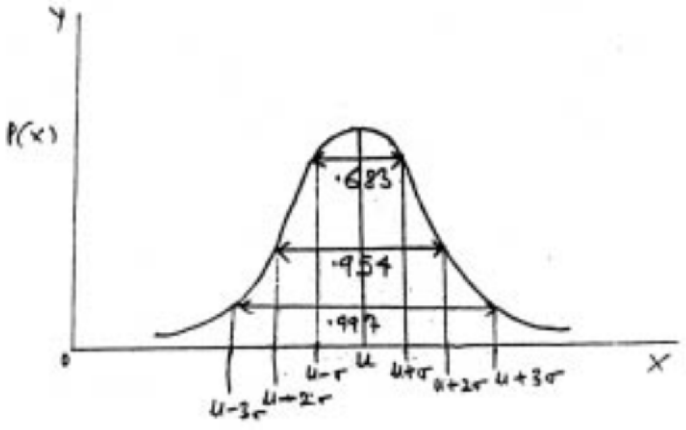


Fig. 5

IIIustration 10 : Suppose weekly income of worker in a locality is normally distributed with mean Rs. 500 and standard deviation Rs. 100. What is probability of falling the workers within one, two and three standard deviations from their mean income ?

Solution : Set x is a normal random variable having mean and standard deviation. Thus according to area properties.

(i) $P[\mu - \sigma \leq x \leq \mu + \sigma] = .6826$

(ii) $P[\mu - 2\sigma \leq x \leq \mu + 2\sigma] = .9545$

(iii) $P[\mu - 3\sigma \leq x \leq \mu + 3\sigma] = .9973$

In our example suppose x denotes the income of the worker, so that μ = Rs. 500 and σ = 100, then

a) Probability of falling the worker within one standard deviation of mean income means the probability of finding the worker having income between 400 i.e. (500-100) and 600 i.e. (500+100) is

$P[400 \leq x \leq 600) = .6826$

It means there are 68.26% chances that worker will have income between 400 and 600.

Similarly

b) The probability of finding a worker having income between two standard deviation is $p[500\text{-}2100 \leq x \leq 500 + 2 \times 100] = .9545$

or $p[300 \leq x \leq 700] = .9545$

c) The probability of finding a worker having income between three standard deviation is

$p[(500\text{-}3 \times 100) \leq x \leq (500 + 3 \times 100)] = .9973$

$p[200 \leq x \leq 800] = .9973$

2.3.2 Standard Normal Distribution :

In normal distribution , we find the probability for different intervals of random variables with the help of two methods, first is with the help of mathematical formula and second, by using statistical table. In business and economics, generally we use statistical table for finding the probabilites. But the mean μ and the standard deviation σ uniquely identifies a normal distribution and there is an infinitely large family of distributions, one for each possible combination of mean and standard deviation. Consequently, it would be necessary to develop a large number of tables to meet the needs. In order to avoid this complicated task of preparing table for use, a special form of normal distribution is used by making a simple transformation. This transformation, know as 'standard normal distribution' has features as any normal distribution.

A normal distribution with zero mean and unit standard deviation is called a 'standard normal distribution'. A normal random variable x with mean equal to zero and standard deviation equal to one is called a standardized normal random variable or standard normal variate and is denoted by z. The random variable z requires only a single probability table, as values of mean and standard deviation are fixed. Any normal random variable x having mean and standard deviation can be transformed into a standard normal variate by using the following simple transformation : $Z = \frac{x - \mu}{\sigma}$

The table given in the last of each book of statistics gives the probability of a standardized normal variate z for the intervals which lies between 0 and any positive value of z. As the normal distribution is symmetrical about its mean, and the total area under the curve is one half to left of zero and other half to the right of zero. The form of the standard normal distribution is given

in following figure.

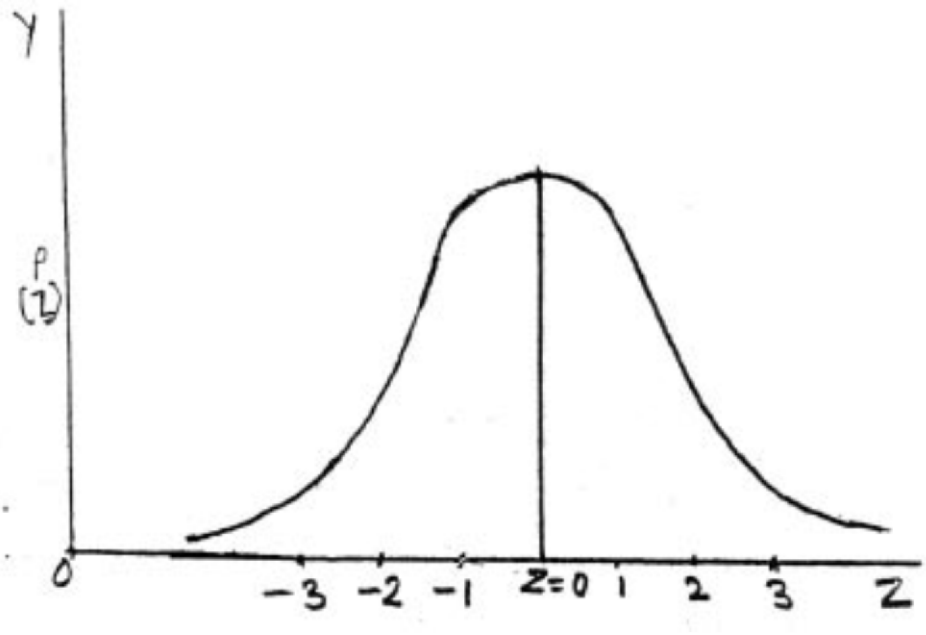


Fig. 6

IIlustration 11: Find the area under the standard normal curve for each of the following intervals.

(a) Between z = 0 and z = 1.75

(b) Between z = -1.45 and z = 0

(c) Between z = -.78 and z = 2.41

(d) Between z = 1.12 and z = 2.75

(e) Greater than z = 2.16

(f) Between z = -.55 and z = .55.

Solution : Table given in the end of this lesson is used to find this probability. The area between z = 0 and z = 1.75 is . 4599

It means any value of z falling between 0 and 1.75 has probability equal to 45.99%

It is shown in the figure 7.

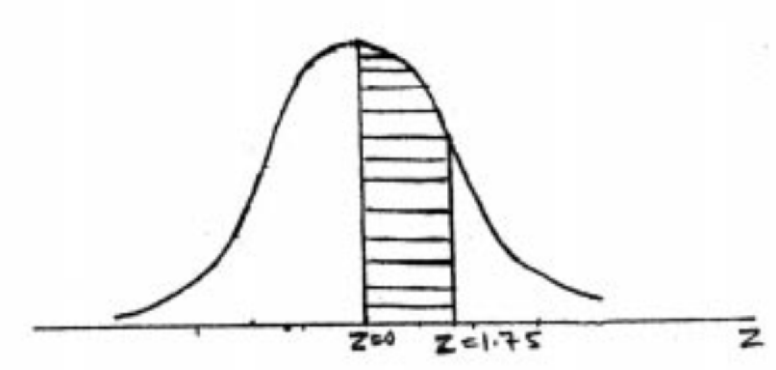


Fig. 7

(b) Since the normal distribution is symmetrical the area between -1.45 and 0 is equal to the area between 0 to 1.45. From statistical table the area between 0 to 1.45 is .4265. It means the probability of falling the z value between -1.45 to 0 is 42.65%. It is shown in the following figure (8).

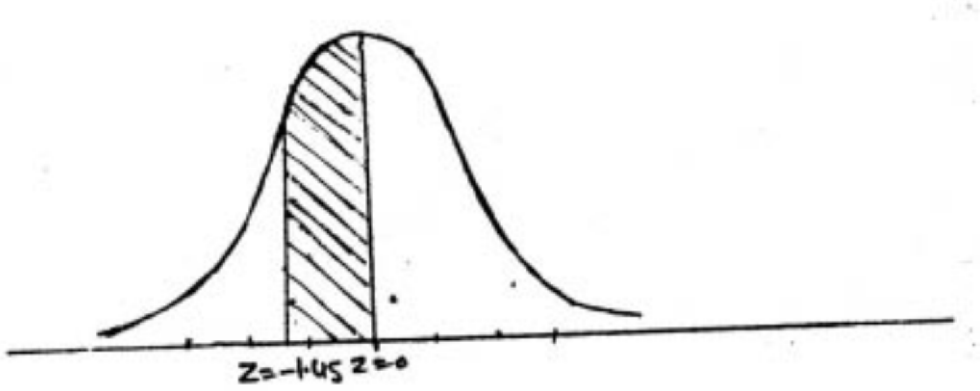


Fig. 8

(c) In this case, we will determine the area in two parts : Total area = (area between 0 and 2.41) + (area between - .78 and 0). The area between 0 and 2.41 is .4920 the area between - .78 and 0 is the same as area between 0 and .78; this area is equal .2518. Therefore, the combined area is 0.7438. it is shown in the figure (9).

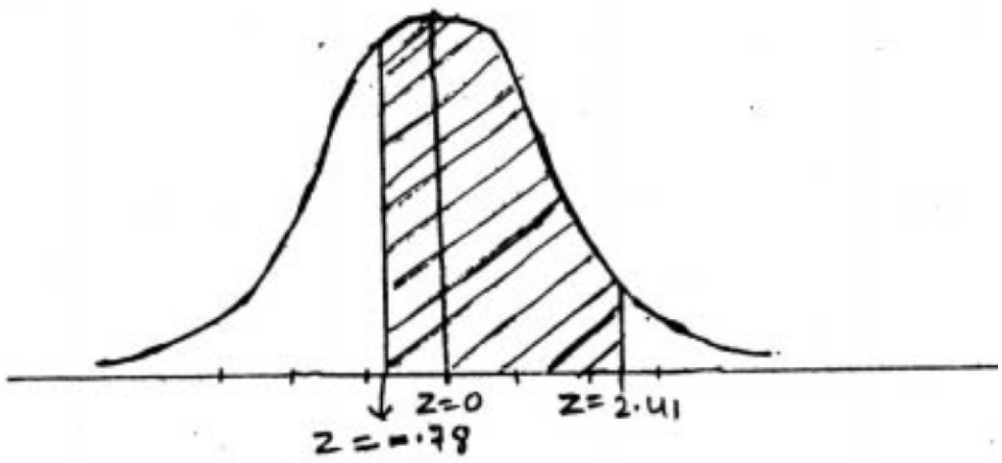


Fig. 9

(d) This area can be found by taking the difference between areas from 0 to 2.75 and from 0 to 1.12. The area between 0 to 2.75 is .4961 and the area between 0 and 1.12 is .3686. Thus, the area between 1.12 and 2.75 is .1279. It is shown in the following figure.

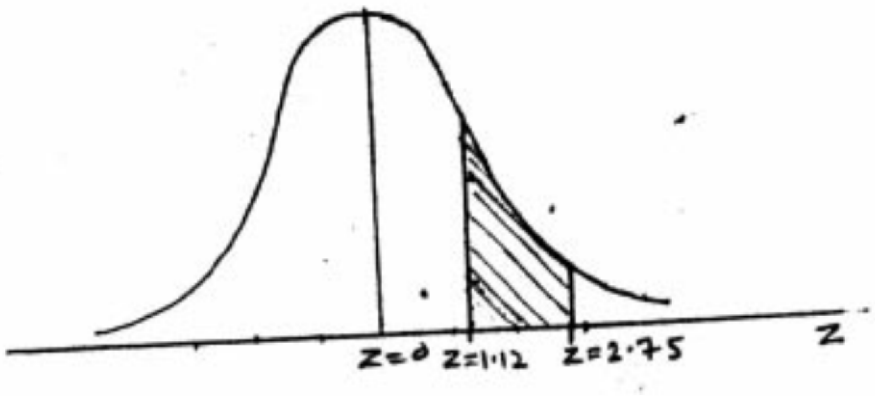


Fig. 10

(e) As the area to the left of 0 is 0.5, the beyond 2.16 is obtained by subtracting the area between 0 and 2.16 from 0.5. Thus the required area is 0.5 - .4846 = 0.0154. It is shown in the following figure.

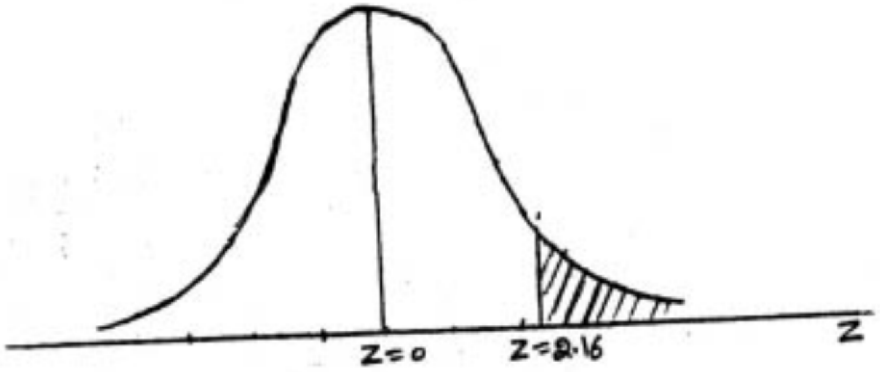


Fig. 11

(f) Beacause of the symmetry of the normal distribution, the area between -.55 and .55 is twice, the area between 0 and .55. Therefore, the required area is 2(.2088) = .4176.

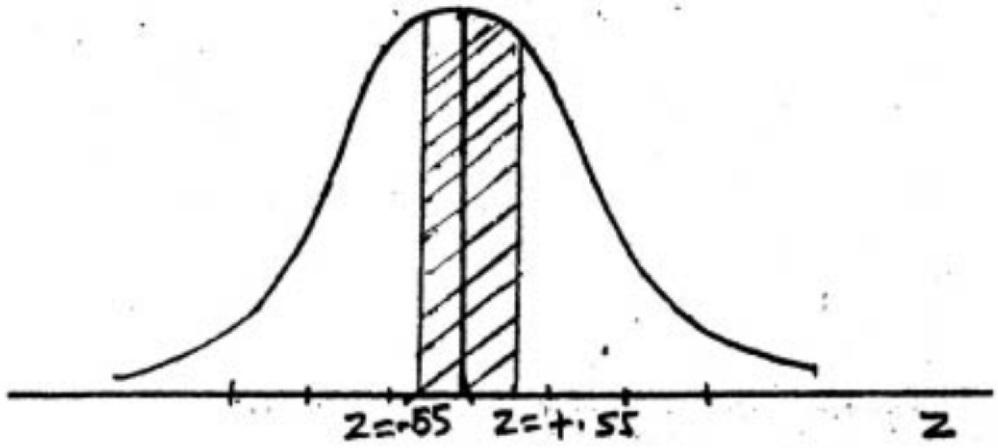


Fig. 12

IIlustration 12 : A machine is designed to produce parts having an average diameter of 2cm, and a standard deviation of 0.3 cm. If the distribution of the diameters of manufactured product is normal.

(a) What is the probability that a randomly selected part will have a diameter exceeding 2.05 cm ?

(b) What is the probability that a randomly selected part will be between 1.95 cm and 2.05 cm?

Solution

(a) The probability that a randomly selected product have diameter exceeding 2.05 cm is given by between x = 2.05 and + . This area can be determined by standardizing x:

$$z = \frac{2.05 - 2}{.3} = 1.6$$

Thus $p[x \geq 2.05] = p[z \geq 1.67]$

$= .5 - p[0 \leq z \leq 1.67]$

$= 0.5 - 0.4525 = .0475$

It means there is 4.75% probability that diameter will exceed 2.05cm.

(b) The probability that selected product has diameter between 1.95cm and 2.05cm is given by the area between $x_1 = 1.95$ and $x_2 = 2.05$. This area can be determined by standardizing x_1 and x_2 as z_1 and z_2

$$z_1 = \frac{1.95 - 2}{.3} = -1.67$$

$$z_2 = \frac{2.05 - 2}{.3} = +1.67$$

Thus

$p[1.95 \leq x \leq 2.05] = p[-1.67 \leq z \leq 1.67] = 2p[0 \leq z \leq 1.67] = 2 \times .4525 = .9050$

It means there is 90.50% probability that product has diameter between 1.95 cm and 2.05cm.

EXERCISES

1. Explain the term random variable and probability distribution of a random variable.
2. Explain what do you mean by the term mathematical expectation. Write down its uses.
3. Define Binomial distributon. Point out its chief characteristics and uses.
4. What is poisson distribution. Explain its main features.
5. Define normal distribution. What are its main characteristics.
6. A machine manufacturing screws is known to produce 5% defective. In a random sample of 15 screws what is probability that there are (i) exactly three defectives (ii) not more than three defectives.
7. An incidence of occupational disease in an industry is such that workers have a 20% chance of suffering from it. What is probability that out of six workers chosen at random, four or more will suffer from desease.
8. A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed on poission variate with mean 1.5. Calculate the proportion of days on which (i) neither car is used and (ii) same demand is refused.
9. Suppose that probability of dialing a wrong number is 0.04. What is probability of dialing not more than two wrong numbers in 50 dialings.
10. In an intelligence test administrated to 100 children, the average score is 42 and standard deviation is 24. Find the number of children (i) exceeding the score 60 and (ii) with score lying between 20 and 40
11. 5000 candidates appeared in a certain examination paper carrying a maximum of 100 marks. It was found that the marks were normally distributed with mean 39.5 and 12.5. Determine approximately the number of candidates who secured a first class for which a minimum of 60 is necessary.

Suggested Readings :

As given in the end of the lesson no. 10.

❄ ❄ ❄

Lesson : 12

SAMPLING : INTRODUCTION, ADVANTAGES AND METHODS OF SAMPLING

Author :
Dr. B. S. Bodla

Vetter:
Dr. R. K. Mittal

Introduction

In this lesson, we offer a variety of methods for selecting the sample, called sampling designs, which can be used to generate our sample data sets. Also, a bird eye view of the estimation process based on simple random sampling technique and method of determining sample size is discussed in it.

Statisticians commonly separate the statistical techniques into two broad categories-descriptive and inferential. The descriptive statistics deals with collecting, summarising and simplifying data which are otherwise too complicated and unwidely. Descriptive statistics facilitates understanding and makes the reporting and discusson of data much easier.

Inferential statistics consists of methods that are used for drawing inferences about the totality of observations on the basis of knowledge about a part of the totality. The totality of observations about which inferences are drawn or generalisations are made is called a population or universe. A part of the totality on which information is generally collected and analysed for the purpose of understanding any aspect of the population is called a sample.

Census or Sample

Sometimes, it is possible and practical to examine every person or item in the population we wish to describe. We call this a complete enumeration, or census. For example, the wage figure could be obtained from each and every worker working in the sugar industry and by dividing the total wages which all these workers receive by the number of workers working in that industry, we can get the figure of average

wage. The census method is not very popularly used in practice. Since the effort, money and time required for carrying out complete enumeration will generally be extremely large and in many cases cost may be so prohibitive that the idea of collecting data may have to be dropped.

In sampling theory, the concept of sample is much more than being merely a part of the population. The sampling units comprising a population may be viewed either as elementary sampling units or primary sampling units. The items contained in a population which are to be ultimately measured or counted with respect to any of their characteristics are known as elementary sampling units. In sugar industry the workers whose wages may have to be measured are the elementary sampling units. The primary sampling units, on the other hand, may be viewed as groups or elementary units such as the number of industry workers working in different departments. We use sampling when it is not possible to count or measure every item in the population. The sample must be representative of the population which is sampled with reference to a given characterstic of interest. That is, a chunk of the sampled population selected on account of convenience in reaching to sampling units cannot be treated as a sample.

Advantages of Sampling

The sampling technique has the following merits over census method.

1) Facilitating Timely Results

Since only a part of the population is to be inspected and examined, the sample method results in considerable amount of saving in time and labour. There is saving in time because (i) a sample usually takes less time in investigation than complete enumeration of the population, and (ii) the time required in editing, coding, and tabulating sample data is much less as compared to that required in the case of census data.

2) More Accurate Results

Although the sampling technique involves certain inaccuracies owing to sampling errors, the results obtained are generally more reliable than that obtained from a complete enumeration. This is mainly because the survey

results, whether based on a sample method or census method, are subject to certain errors. These errors arise because of factors such as poor planning, ineffective execution, and lack of proper control over the various activities concerning with conducting the survey. But the effect of these errors is bound to be much less in the case of sampling simply because the magnitude of sampling operations are much smaller.

3) Less Cost

The sample method is much more economical than a complete enumeration. This is because of the fact that in sampling, we study only a part of population and the total expenses of collecting data is less than that required when the census method is adopted.

4) Destructive Testing

If in the course of inspection, the units are destroyed or affected adversely, then we are left with no option than to resort to sampling. For example, to test the quality of explosives, crackers, shells, etc. sampling is used.

5) Used in Certain Cases

In many cases census method may not be physically possible. This is particularly the case where the population to be investigated are either infinite in terms of numbers or otherwise constantly changing. In such situations sampling technique is the only method of obtaining desired information about the population.

The merits of sample surveys over census method can be realised only if

a) the sample is drawn in a scientific manner,

b) the appropriate sampling design is used, and

c) the sample size is adequate,

There are some situations in which conducting a census is better than taking a sample. For example

i) When the population size is very small the cost and time required for complete enumeration will be less.

ii) When the population is very heterogeneous.

iii) In cases like totalling of cash, income, expenditure, etc the accuracy can be achieved only through census.

Some Definitions :

1. Definition of Sampling Design :

The sampling design or survey design specifies the method of collecting the sample. The design does not specify a method of collecting or measuring the actual data. It only specifies a method for collecting the objects that contain the required information. These objects are called elements.

2. Definition of Element :

An element is an object on which a measurement is taken. The elements may occur individually or in groups in the population. A group of elements, like a household of community residents or a carton of light bulbs, is called a sampling unit.

3. Definition of Sampling Unit :

Sampling units are non-overlaping collection of elements from the population. In some cases a sampling unit is an indivdual element.

4. Definition of Frame :

To select a random sample of sampling units, we need a list of all sampling units contained in the population. Such a list is called a frame.

5. Population or Unvierse

In statistical investigation we usually study the various characteristics relating to items or individuals belonging to a particular group. This group of individuals under study is known as the population or universe. Examples are- incomes of all people in a certain country during a specific time period or all outcomes of repeatedly tossing a coin. A population containing a finite number of objects or items is known as finite population. On the contrary, populations consisting of unlimited number of units are known as finite population. This population generally refer to processes operating under a given set of well-

defined conditions. The number of units produced by a machine as long as the machine continues to operate under given set of conditions represents an infinite population.

From the point of view of sampling, a clearly defined and identified population which we intend to reach and sample is known as the target population. Many times all the elementary units comprising the target population may not be located or may not be accessible for necessary observation by the investigator. This problem indicates need to prepare a list of elementary sampling units of the target population that are actually within our reach. Such a list is known as the "frame", and constitutes "sampled population".

Theoretical Basis of Sampling

Sampling theory helps in predicting and generalising the behaviour of mass phenomena. This is possible because the variability in the measure of the elementary units is a universal property of all population, but the actual variations are not without limits. For example, wheat varies to a limited extent in colour, protein content, weight etc., but it can always be identified as wheat. Now we shall discuss some important laws which form the basis of the sampling theory.

Law of Statistical Regularity

According to King, "The Law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group." In other words, this Law points out that if a sample is taken at random from a population, it is likely to possess almost the same characteristics as that of the population. The law of statistical regularity impresses upon the one very important point, that is, the sample should be selected at random from the population.

By random selection we mean a selection in which each and every unit in the universe has an equal chance of being selected or rejected in the sample. A sample selected randomly would be representative of the universe. If our sample is random, then it will depict the true characteristics of the population fairly and accurately and can be used for drawing valid inferences about the population.

Law of Inertia of Large Numbers

This law is an immediate deduction from the law of statistical regularity. It states, "Other things being equal, as the sample size increases, the results tend to be more reliable and accurate". This is based on the fact that the large numbers are more stable as compared to small ones. This is because a number of force operate on the given phenomenon and if the units are large then the typical odd variations in one part of the universe in one direction will get neutralised by the variations in equally bigger part of the universe in the other direction. For example, if a coin is tossed 10 times we should expect equal number of heads and tails, i. e., 5 each. But, on account of small number of tosses it is likely that we may not get exactly 5 heads and 5 tails. If this experiment is carried out 1,000 times, the chance of 500 heads and 500 tails would be very high. The basic reason of such outcome is that the experiment has been carried out in sufficiently large number of times and the possibility of variations in one direction compensating for others in a different direction is greater.

Bias and Error in Sampling

In statistics, the difference between the true value and the estimated value is called 'error'. In other words 'error' refers to the difference between the true value of a population parameter and its estimate provided by an appropriate sample statistic computed by some statistical device. There are many causes for such deviations between two results. The errors in any statistical investigation may be broadly classified as sampling and non sampling errors.

Sampling Errors

The results of a sample survey are bound to differ from the census results since only a small portion of the population is studied in the sample method. These errors are associated only with sample surveys and tend to disappear in census surveys. These errors are caused primarily due to faulty selection of a sample, faulty demarcation of sampling units, improper use of the estimation techniques and the variability or heterogeneity of the population to be sampled.

A measure of the sampling error is provided by the standard error of the estimate. The reliability of a sampling plan is determined by the reciprocal of the standard error and is called the precision of the estimate. It has been

observed that standard error of the estimate is inversely proportional to the square root of the sample size.

Non sampling Errors

Non sampling errors are not attributed to chance and are a consequence of certain factors which are within human control. These arise in all surveys, whether it is a sample survey or a census survey. Such errors can arise due to a number of causes such as defective methods of data collection and tabulation, faulty definition, incomplete coverage of the population or sample etc. More specifically, non sampling errors may arise from one or more of the following factors :

1) Faulty planning, including vague and faulty definitions of the population to be sampled.
2) Inaccurate methods of interview, observation with inadequate or ambiguous schedules or instructions.
3) Lack of trained and experienced investigators.
4) Personal bais of the investigator.
5) Failure of respondents' memory to recall the events or happenings in the past.
6) Errors committed in data processing operations such as coding, punching, verification etc.
7) Errors may arise during presentation and printing of tabulated results.

The above list is not exhaustive, but are some of the important causes of biasness and mistakes. The non sampling errors can be controlled by the methods including employment of qualified and trained staff, using sophisticated statistical techniques, pretesting or conducting a pilot survey, more effective follow up of non-response cases, through editing and scrutiny of the results.

It must be noted that the researcher can only minimize the chances of errors due to random variation by selecting a proper sampling design. The researcher can have a much more direct effect on errors due to nonresponse. Continued efforts can be made to reach the nonrespondents. To minimize the chances of incurring errors due to wrong specification, a very careful statement

of the survey objectives can be made in advance of the study, thus providing a clear image of the elements that comprise the population.

Sampling Methods

There are two methods of selecting samples from population non-random or judgement sampling, and random or probability sampling. In probability sampling, all the items in the population have a chance for being chosen in the sample. In non random or judgment sampling, personal knowledge and opinions are used to identify those items from the population that are to be included in the sample. A sample selected by judgment sampling is based on someone's expertise about the population. Sometimes a judgement sample is used as a pilot to decide how to take a random sample later.

In the following pages of this lesson we shall discuss some of the important sampling methods.

Simple Random Sampling

Simple random sampling selects samples by methods that allow each possible sample to have an equal probability of being selected and each item in the entire population to have an equal chance of being included in the sample. Selection of sample units may be with or without replacement. A very important and interesting feature of simple random sampling without replacement is that, the probability of selecting specified unit of population at any given draw is equal to the probability of its being selected at the first draw. "This implies that in simple random sampling with replacement the probability of each unit to be included in the sample is I/N at each draw. In case of sampling without replacement particular unit selected in any draw cannot be selected again in any other draw, and the probability of selection of each of the a remaining units in the population is I/N at the first draw, I/N-1 at the second draw, I/N-2 at the third draw, 1/[N-(n-1)] at the nth draw.

How to Select a Random Sample :

A random sample may be selected by

(i) Lottery Method

(ii) Use of Random Numbers

Lottery Method

This is a very popular and simple method of taking a random sample. Under this method, every member or unit of the population are numbered or named on separate slips of paper of identical size and shape. Then these slips are put in a bag and thoroughly shuffled and then as many slips as units needed in the sample are drawn one by one, the slips being thoroughly shuffled after each draw. For example, let us suppose that we want to take a sample of 10 persons out of the population of 100, the procedure is to write the names of all the 100 persons on separate slips of paper, fold these slips, mix them thoroughly and then make blindfold selection of 10 slips. This method is quite frequently used in the random draw of prizes, in the Tambola games and so on.

Use of Table of Random Digits

The best way to ensure that we are employing random sampling is to use table of random numbers. The random numbers are usually generated by some mechanism which ensures approximately equal frequencies for the numbers from 0 to 9 and also proper frequencies for the combination of number such as 00, 01, ..., 99 etc.

The method of drawing a random sample comprises the following steps :

(1) Identify N units of the population with the numbers, 1 to N.

(2) A random number table of two or more pages is obtained with closed eye, a pencil edge is put anywhere on any number in the table which becomes the starting point for the selection of units to be included in the sample.

(3) Then from the starting point we make a move on to the next number either way we like-vertically, horizontally, or diagonally. The process of recording of random numbers continues till we have obtained the random numbers equal to our sample size.

Several standard tables of random numbers are available. Tippett's Table of random numbers is most popularly used in practice. The first forty sets from

Tippetts Table have been reproduced below.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	2183	3408	2762	3563	1089	6913	6591
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Example 12.1 :

Draw a random sample (without replacement) of 20 students from a class of 400 students.

Solution

At the outset, identify the 400 students with numbers from 1 to 400. Now put your pencil edge with closed eyes on any number and then we pick out one by one on the three digited numbers less than or equal to 400, till 20 numbers from 400 are obtained. In this process the numbers over 400 are discarded and the repeated numbers, if any, are taken only once.

The above numbers grouped in three's will be as below :

295	266	413	992	979	279	795	911	317	056	244
167	952	415	451	396	720	353	561	300	269	323
707	483	340	827	623	563	108	969	137	691	056
052	461	112	610	760	088	126	423	387	762	754
914	314	059								

The desired sample of 20 students will be constituted by the students corresponding to the ramdom numbers given below :

295	266	279	056	244	167	396	300	269	323
340	108	137	052	112	088	126	387	314	059

The use of random numbers in simple random sampling is quite

practicable in case the size of population is limited and the sampling units not scattered over a large area. However, when the size of sampled population is quite large, listing of items itself and then using the random numbers would become a very time consuming and costly job.

Estimation Based on a Simple Random Sample :

Once the sample has been drawn, the next aspect of sampling is to estimate the population parameters on the basis of observations on sampling units. The population parameters of common interest are the population mean, population total and population proportion.

Let Y be the character of interest and $Y_1, Y_2, \ldots\ldots\ldots\ldots Y_n$ be the values of the character on N units of the population. Further, let $y_1, y_2, \ldots\ldots\ldots\ldots, y_n$ be the sample of size n, selected by simple random sampling. The sample mean is an unbiased estimate of the population mean μ.

$$\overline{Y} = \frac{\Sigma y_i}{n}$$

Variance estimator is given by

$$V(\bar{y}) = \frac{N-n}{N}\,\frac{s^2}{n} \qquad \text{where } s^2 = \sum_{i=1}^{n} (y_i-\bar{y})^2/n-1$$

And bounds on the error of estimation is given by

$$\bar{y} - 2\sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}}$$

This bound is taken to imply that at least 75% and most likely 95%, of the estimates will deviate from the mean by less than two deviations.

When the survey objective is to use simple random sampling to estimate the population total 1, we have the following formulae :

$\hat{\tilde{1}} = N\text{-}y$

$$\hat{V}(\hat{\tilde{I}}) = N^2 V(y) = \frac{N(N-n)\ s^2}{n}$$

Bounds on the error of estimation :

$$N\bar{y} - 2\sqrt{\frac{N(N-n)\ s^2}{n}}$$

The finite population correction (fpc) factor, $\frac{N-n}{N}$, is to adjust for sampling from a finite population. When n is small relative to the population size N, the fpc, $\frac{N-n}{N}$, is close to 1.

Example 12.2

The effective management of cash flows by business organization is necessary for the proper budgeting and control of their present and future resources. In an analysis of the cash position of a departmental store, an accounting firm decides to select a simple random sample of n=15 monthly retail accounts receivable from among the N=1000 current monthly retail accounts of the department store in order to estimate the total amount due to all outstanding accounts receivable.

The simple random sample of n=15 accounts yielded the following :

Rs. 14.50	Rs. 23.40	Rs. 42.00
30.20	15.50	13.30
17.80	27.50	23.70
10.00	6.90	18.40
8.50	19.50	12.10

Solutions :

y_i	14.50	30.20	17.80	10.00	8.50	23.40
y_1^2	210.25	912.04	316.84	100.00	72.25	547.56
y_1	15.50	27.50	6.90	19.50	42.00	13.30

y_1^2	240.25	756.25	47.61	380.25	1764.00	176.89
y_1	23.70	18.40	12.10	$\Sigma y_i = 283.30$		
y_i^2	561.69	338.56	146.41	$\Sigma y^2_i = 6570.85$		

a. Our estimate of the mean account balance μ is

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{283.30}{15} = \text{Rs. } 18.89$$

To find a bound on the error of estimation of μ, we must first compute

$$s^2 = \frac{\Sigma (y_i - \bar{y})^2}{n-1} = \frac{\Sigma y_i^2 - (\Sigma y_i)^2/15}{14}$$

$$= \frac{1}{14}\left[6570.85 - \frac{(283.30)^2}{15}\right]$$

$$= \frac{1}{14}[6570.85 - 5350.59] = 87.16$$

The estimated variance of y is therefore

$$V(y) = \frac{N-n}{N}\frac{s^2}{n} = \left(\frac{1000-15}{1000}\right)\left(\frac{87.16}{15}\right) = 5.72$$

An estimate of the mean account balance μ, with a bound on the error of estimation, is Rs. 18.89–$2\sqrt{5.72}$ = Rs. 18.89–Rs. 4.78

(b) An estimate of the total amount due on outstanding accounts receivable is provided by N(y) = 1,000 (Rs. 18.89) = Rs. 18,890.

Hence, an estimate of the total due on all N = 1000 accounts, with a bound on the error of estimation, is

$Ny - 2N\sqrt{\hat{V}(y)} = Rs.\ 18,890 - 2\ (1000)\ \sqrt{5.72}$

=Rs. 18,890–Rs. 4,783.

Stratified Random Sampling

This design is recommended when the population consists of a set of heterogeneous groups. To use this method, we divide the population into relative homogeneous groups, called strata. Then we use one of two approaches. Either we select at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole, or we draw an equal number of elements from each stratum and give weight to the results according to the stratum's proportion of total population. The stratified sampling guarantees that every element in the population has a chance of being selected irrespective of the approach used.

The overall sample size "n" depends on our available budget for sampling and on how precise and accurate we wish to estimate. The sample size is allocated among the K strata so that,

$n = n_1 + n_2 + \ldots\ldots\ldots\ldots + n_i,$

Where each n_i is given by the formula :

$n_i = n(N_i/N) \qquad i=1, 2, \ldots\ldots\ldots, k$

Where N_i is the number of elements in stratum i and

N is the size of the population.

From the information obtained from the sample elements, we can compute the estimated mean Y_i of the observations within each stratum by using the formula.

$$\overline{Y}_i = \sum_{i=1}^{n} y_{ij}/n_i$$

Where y_{ij} is the jth observation in stratum i.

Y_i is the mean of the ith stratum in the population. Then the estimate Y_{st} of the population mean u, based on stratified random sampling is obtained as

$$Y_{st} = 1/N \sum_{i=1}^{k} N_i y_i$$

Example 12.3 : You are given the following data of the number of male and female students in a University :

Faculty	*Male*	*Female*	*Total*
Arts	860	540	1400
Science	240	160	400
Business Studies	100	100	200
Total	1200	800	2000

Work out how many male and female students would be selected from each category if we follow stratified proportionate sampling method and take 10% of the universe equivalent to the sample size.

Solution

The sample size is 10% of the universe hence 200 students woud be selected in the sample. Since 6 strata are formed and we want to follow proportionate stratified sampling method, we will take 10% from each stratum. The number of students selected shall be as follows :

Faculty	*Male*	*Female*	*Total*
Arts	86	54	140
Science	24	16	40
Business Studies	10	10	20
Total	120	80	200

The advantage of stratified samples is that when they are properly designed, they more accurately reflect characteristics of the population from which they were chosen than do other kinds of sampling.

Cluster Sampling

In cluster sampling, we divide the population into groups, or clusters, and then select a random sample of these clusters. It is assumed that these individual clusters are representative of the population as a whole. If the selection of a sample passes through more than two stages of sampling, such a sampling method would be known as multi-stage sampling.

Suppose we want to take a sample of 5,000 households from the State of Haryana. At the first stage, the state may be divided into a number of districts and a few districts selected at random. At the second stage, each district may be subdivided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at the second stage.

A well-designed cluster sampling procedure can produce a more precise sample at considerably less cost than that of simple random sampling.

Systematic Sampling

A design that avoids the cumbersome data collection requirements of simple random sampling is systematic sampling. A systematic sample is obtained by randomly selecting one element from the first 'K' elements in the frame and then selecting every Kth element thereafter, since it is easier and less time-consuming to perform this, than simple random sampling, systematic sampling can provide more information per sampling rupee. However, hidden periodicities exist in the population, the 1-in K systematic sample may bias results by introducing sampling error resulting from the periodic influence. Also, it should be avoided when the population size is unknown.

Determining Sample Size

One of the first questions asked by some one undertaking a sample survey is, "How many sample elements should I select ?" Since sampling is time-consuming and costly, our objective in selecting a sample is to obtain a specified amount of information about a population parameter at a minimum cost. We can accomplish this objective by first deciding on a bound the error of estimation and then applying an appropriate sample size estimation formula.

The objective behind the selection of a sampling design and selection of the sample size are the same-to obtain a specified amount of information at a minimum cost. Sampling design decisions are made according to the "law of the land", that is, how the elements group contained by those elements. Sample size decisions are made according to the inherent variability in the population of measurements and how accurate the experimenter wishes the estimate to be. These two criteria are, of course, inversely related. To obtain greater accuracy, and hence more information about a population, we must select a larger sample size; the greater the inherent variability in the population, the larger is the sample size required to maintain a fixed degree of accuracy in estimation.

Simple Random Sampling

When using simple random sampling, the sample size required to estimate the population mean μ, with a bound B on the error of estimation is given by

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad D = \frac{B^2}{4}$$

where σ^2 is the population variance, N is the number of elements in the population, and B is the bound on the error of estimation.

When N is very large, the sample size formula reduces to

$$n = \frac{4\sigma^2}{B^2}$$

When the objective is to estimate the population total 1, with a bound B on the error of estimation, we must substitute $D = B^2/4N^2$ into the sample size formula given above.

One may notice a dilemma in the guidelines provided for finding n. To find n we must know the population varince, but to estimate σ^2 we must have a set of sample measurements from the population. The variance can be estimated by s^2 obtained from a previous sample or by knowledge of the range of measurements, giving the estimate

$$\sigma^2 = \frac{1}{16} (\text{range})^2$$

The range approximation procedure is derived from the Empirical Rule and provides a very rough approximation to σ^2.

Judgment Sampling

In this method of sampling the choice of sample items depends exclusively on the judgment of the investigator. For example, if a sample of ten students is to be selected from a class of sixty for analysing the spending habits of students, the investigator would select ten students who, in his opinion, are representative of the class. This method, though simple, if not scientific because the population units to be sampled may be affected by the personal prejudice or bias of the investigator. Hence, the success of this method depends upon the excellence in judgment.

Quota Sampling

In this method, the investigator is told in advance the number of sample units he is to examine or enumerate from the stratum assigned to him. The sampling quotas may be fixed according to some specified characteristics such as income group, sex, occupation, religious affiliations, etc. The choice of the particular units of individuals for investigation is left to the investigators themselves. Quite often the investigator does not make a random selection of the sample units. He usually applies his judgment in the choice of the sample. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is not widely used in practical work.

Do yourself :

7.1 Distinguish between a population and a sample.

7.2 Point out the importance of sampling in solving business problems. What are the basic principles on which sampling theory results ?

7.3 Differentiate between the following :

a) Target and sampled population

b) Sampling and non-sampling errors

c) Stratified sampling and cluster sampling

7.4 Enumerate the various methods of sampling and describe two of them mentioning the situations where each one is to be used.

7.5 What is systematic sampling ? How does it differs from simple random sampling ?

7.6 Discuss the simple random sampling and draw a sample of size 15 from a hypothetical population of your choice by using ranndom number tables.

7.7 The manager of the credit division of a commercial bank would like to know the average amount of credit purchases placed on Bank card each month by the customers who hold Bank cards issued by the bank. Since there are currently 20,000 open Bank card accounts at the bank, time and expense prohibit a complete review of every account. The manager thus proposes selecting a simple random sample of open Bank card accounts to estimate the average monthly account balance μ, with a bound on the error of estimation of B = Rs. 10. Although no prior information is available to estimate the variance σ^2 of monthly Bank card account levels, it is known that most account levels lie within the range from Rs. 50 to Rs. 450. Find the sample size necessary to achieve the stated bound.

(Answer n = 400)

❋ ❋ ❋

Printed by : Competent Printing Press, Hisar-125001
Mobile : 98960-68720, 92156-25100