Directorate of Distance Education

UNIVERSITY OF JAMMU
JAMMU

IInd Semester

Study Material

M.A. ECONOMICS	**Unit I to IV**
Course No. : ECO-203	**Lesson No. 1–21**

Course Coordinator :	Teacher Incharge :
Prof. Prakash C. Antahal	**Dr. Neelam Choudhary**

http://www.distanceeducationju.in

Printed and published on behalf of the Directorate of Distance Education, University of Jammu, Jammu by the Director, D.D.E., University of Jammu, Jammu

M. A. ECONOMICS SEMESTER II

Course Coordinators :

Prof. Jasbir Singh
Dr. Jasdeep Singh Toor

Editing & Proof Reading :

Ms. Satakshi Gupta
Research Scholar,
PG Department of Economics,
University of Jammu.

Printed by : Sethi Art Printers / 2019 / 500

Course No: 203 **Maximum Marks: 100**

Credits 4 **2nd Semester Examination 80**

Sessional Assessment 20

STATISTICAL ANALYSIS

Preamble: The objective of this course is to equip students in the use of various statistical techniques that can be used on empirical data which will enable them to analyse and quantify economic relationships, verify theories, measure variables that are pertinent to the study of Economics.

UNIT-I:

MEASURES OF CENTRAL TENDENCY, DISPERSION & CORRELATION AND REGRESSION

Statistics -nature and scope. Types of Statistics-descriptive and inferential

Measures of Central Tendency and Dispersion

Meaning, assumptions and limitations of Simple correlations and regression analysis; Pearson's Product Moment Formula and Spearman's Rank Correlations.

Concept of least squares and lines of regression; Methods of estimating non-linear equations e.g. parabolic equation, Standard Error of estimate

Partial and Multiple Correlations, Coefficient of determination

UNIT-II:

INDEX NUMBERS, TIME SERIES ANALYSIS, MEASURES OF INEQUALITY

Nature and Purpose of Index Numbers, Commonly Used Index Numbers, Laspeyres and Paasche's Index Numbers, Chain Base Index Numbers, Official Index Number, True Cost of Living Index, Fisher's test for Index Numbers

Nature and Decomposition of a Time Series, Analysis of Trend: Polynomial Trend, Non-linear Growth Curves; Moving Average Method, Seasonal Component, Cyclical and Random Component, Forecasts and their Accuracy

Measures of Inequality- Desirable properties of Measure of Inequality, Gini Coefficient, Lorenz Curve, Kuznet Ratio, Co-efficient of variation and Relative Range.

UNIT-III:

PROBABILITY AND DISTRIBUTION

Probability theory - different concepts and approaches, Laws and Axioms of Probability, conditional probability and concept of Interdependence, Baye's theorem and its applications.

Concept of random variable: probability, mass and density functions; Expectations, moments and moment generating functions

Probability distributions: Binomial, Poisson, and Normal.

UNIT-IV:

THEORY OF ESTIMATION AND HYPOTHESIS TESTING

Point Estimation, Concept of an estimator and its sampling distributions, Properties of a good estimator, Interval estimation, estimating means, proportions, variances of populations from samples

Testing of Hypothesis: Formulation of statistical Hypotheses-Null and Alternative Hypothesis Goodness of fit; confidence interval and level of significance, Hypothesis testing based on Z. t, F, and chi-square tests; Errors of types II and I.

CONTENTS

Unit III : Probability and Distributions

Script Writer : Prof. Jasbir Singh

Unit IV : Theory of Estimation and Hypothesis testing

Script Writer : Prof. Jasbir Singh

M.A. Economics Semester - IInd Lesson - 1
Course No. 203 Unit - I

STATISTICS - NATURE AND SCOPE, TYPES OF STATISTICS - DESCRIPTIVE AND INFERENTIAL

CHAPTER HIGHLIGHTS:

This chapter discuss about nature and scope of statistics. It also explains the descriptive and inferential statistics.

STRUCTURE

1.1 INTRODUCTION:

The subject of statistics, as it seems, is not a new discipline but it is as old as the human society itself. It has been used right from the existence of life on this earth, although the sphere of its utility was very much restricted.

1.2 ORIGIN AND DEVELOPMENT OF STATISTICS:

In the old ways, statistics was regarded as the 'Science of statecraft' and was the by product of the administrative activity of the state. The word statistics seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'Statistic' or the French word 'statistique' each of which means a political state. In the ancient times the scope of statistics was primarily limited to the collection of the following data by the governments for graming military and fiscal policies :

i. Age and sex-wise population of the country;

ii. Property and wealth of the country ;

The former enabling the government to have an idea of the manpower of the country (in order to safeguard itself against any outside aggression) and the latter providing it with information for the introduction of new taxes and levies.

Perhaps one of the earliest census of population and wealth was conducted by the Pharaoks (Emperors) of Egypt in connection with the construction of famous 'Pyramids'. Such censuses were later held in England, Germany and other Western countries in the middle ages. In India, an efficient system of collecting official and administrative statistics existed even 2000 years ago - in particular during the reign of Changragupta Maurya (324-300 B.C) Distincal evidences about the prevalence of a very good system of collecting vital statistics and registration of births and deaths even before 300 B.C. are available in Kautilya's Asthashashtia'. The records of land, agriculture and wealth statistics were maintained by Todermal, the land and revenue minister in the reign of Akbar (1556-1605 A.D.) A detailed account of the administrative and statistical survey conducted during Akbar's reign is available in the book 'Ain-e-Akbar' written by Abdul Fazl (1596-97), one of the nine gems of Akbar.

Sixteenth century saw the applications of statistics for the collection of the data relating to the movements of heavenly bodies - stars and planets to know about their positions and the prediction of eclipses. Whereas, seventeenth century witnessed the origin of vital statistics. In the mid-seventeenth century, theory of statistics was developed which is the backbone of modern theory. Modern Stalewarts in the development of the subject of statistics are Englishmen who did pioneering work in the application of statistics to different disciplines e.g., Francis Galton (Regression Analysis) and Karl Pearson (Correlation Analysis). Perhaps it was R.A. fisher who applies statistics to a variety of diversified fields such as genetics, biometry, psychology and education, agriculture etc. and who is rightly termed as the father of statistics. Indian statistican also did not lag behind in making significant contributions to the development of statistics. Indian statisticians also did not lag behind in making significant contributions to the development of statistics in various diversified fields. The valuable contributions of C.R. Rao (Statistical Inference); Parthasarathy (Theory of Probability) ; P.C. Mahalanobis and P.V. Sukhatme (Sample surveys); S.N. Roy (Multivariate Analysis); R.C. Bose, K.R. Nair, J.N. Srivastava (Design of Experiments) to mention only a few have placed India's name in the world map of statistics.

1.3 NATURE OF STATISTICS:

Statistics has been defined differently by different writers from time to time so much so that scholarly articles have collected together hundreds of definitions, emphasizing precisely the meaning, scope and limitation of the subject. The reasons for such a variety of definitions may be broadly classified as follows :

i. The field of utility of statistics have been increasing steadily and thus different people defined it differently according to the subject.

ii. The word statistics has been used to convey different meanings in singular and plural sense. When used as plural, statistics means numerical set of data and when used in singular sense it means the science of statistical methods embodying the theory and techniques used for collecting, analysing and drawing inferences from the numerical date.

1.3.1 Definitions of statistics:

i. “Statistics are the classified facts representing the conditions of the people in a state...specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement.” - Webster

ii. “Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.” - Bowley

iii. “By statistics we mean quantitative data affected to a marked extent by multiplicity of causes.” - Yule & Kendall.

Prof. Horace Secrist define statistics as :

i. **Aggregate of facts :** Simple or isolated items cannot be termed as statistics unless they are a part of aggregate of facts relating to any particular fields of enquiry.

ii. **Affected by Multiplicity of Causes :** Numerical figures should be affected by multiplicity of factors. In physical sciences, it is possible to isolate the effect of various factors on a single item but it is very difficult to do so in social sciences, particularly when the effect of some of the factors cannot be measured quantitatively. However, statistical techniques have been devised to study the joint effect of a number of a factors on a single item (Multiple correlation) or the isolated effect of a single factor on the given item (Partial correlation) provided the effect of each of the factors can be measured quantitatively.

iii. **Numerically expressed :** Only numerical data constitute statistics. Qualitative characteristics which cannot be measured quantitatively such as intelligence, beauty, honesty etc., cannot be termed as statistics unless they are numerically expressed by assigning particular scores as quantitative standards.

iv. **Enumerated or Estimated According to reasonable standard of Accuracy :** The numerical data pertaining to any field of enquiry can be obtained by completely enumerating the underlying population. In such a case data will be exact and accurate (but for the errors of measurement, personal bias. etc.) However, if complete enumeration of the underlying population is not possible then the data are estimated by using the powerful techniques of sampling and estimation theory. However, the estimated values will not be as precise and accurate as the actual values. The degree of accuracy of the estimated values largely depends on the nature and purpose of the enquriy.

v. **Collected in a systematic manner :** The data must be collected in a very systematic manner. Thus, for any socio-economic survey, a proper schedule depending on the object of enquriy should be prepared and trained personnel (investigation) should be used to collect the data by interviewing the persons. An attempt should be made to reduce the personal bias to the minimum. Obviously the data collected in haphazard way will not conform to the reasonable standards of accuracy and the conclusions based on them might lead to wrong or misleading decisions.

vi. **Collected for pre-determined purpose :** It is of utmost importance to define in clear and concrete terms the objectives or the purpose of the enquiry and the data should be collected keeping view view these objectives. An attempt should not be made to collect too many data some of which are never examined or analysed i.e., we should not waste time in collecting the information which is irrelevant for our enquiry. Also it should be ensured that no essential data are omitted.

vii. **Comparable:** From practical point of view, for statistical analysis the data should be compariable. They may be compared with respect to some unit, generally time (period) as place.

1.3.2 Statistical Method:

The large volume of numerical information gives rise to the need for systematic methods which can be used to organise, present, analysie and interpret the information effectively statistically methods are primarily developed to meet this need.

In this sense, statistics has been defined in numerous ways. According to

Bowley, "Statistics may rightly be called the science of averages."

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."

There are five stages in a statistical investigation :

i. **Collection:** Collection of data constitutes the first step in a statistical investigation. The data may be available from existing published as unpublished sources or else may be collected by the investigator himself.

ii. **Organisation:** Data collected from published sources are generally in organised form. The first step in organising a group of data is editing. The collected data must be edited very carefully so that the omissions, in consistencies, irrelevant answers and wrong computations in the returns form a survey may be corrected or adjusted. Next step is to classify i.e., to arrange the data according to some common characteristics possessed by the items constituting the data. The last step in organisation is tabulation i.e, to arrange the data in columns and rows so that there is absolute clarity in the data presented.

iii. **Presented :** After the data have been collected and organised they are ready for presentation. There are two different modes in which the collected data may be presented :

a) Diagrams

b) Graphs

iv. **Analysis :** The purpose of analysing data is to dig out information useful for decision-making. Various methods are used in analysing the presented data mostly in a tabular form. Most commonly used methods are : measures of central tendency, measures of variation, correlation, regression, etc.

v. **Interpretation :** It is drawing conclusions from the data collected and analysed. Only correct interpretation will lead to a valid conclusion of the study and thus and aid one in taking suitable decisions.

1.3.3 Statistics : Science or Art

Whether statistics is a science or an art is often a subject of debate. Science refers to a systematised body of knowledge. It studies cause and effect relationship and attempts to make generalisations in the form of scientific principles as laws. It describes objectively and avoids vague judgement as good or bad. Science, in short, is like a lighthouse that gives light to the ships to find out their own way but does not indicate the direction in which they should go. Art, on the other hand, refers to the skill of handling facts so as to achieve a given objective. It is concerned with ways and means of presenting and handling data, making inferences logically and drawing relevant conclusions.

While a century ago there were some misgivings among natural scientists as to whether statistics had the right to be recognised as a distinct science, now almost all sciences are statistical. What this suggest is that the design of scientific experiments and the evaluation of their results make use of principles and practices growing out of the science of statistics. However, statistics as a science is not similar to exact sciences like Physics, Chemistry, Zoology etc. This is because statistical phenomena are generally affected by a multiplicity of causes which cannot always be measured accurately. In other words, the science of statistics by its very nature is less precise than the natural sciences. It is science only in a limited sense, viz., as a specialized branch of knowledge. More appropriately, statistics may be regarded as a scientific method because it is really a tool which can be used in scientific studies. Wallis and Roberts have rightly remarked that “Statistics is not a body of substantive knowledge but a body of methods for obtaining knowledge”. If the science is knowledge, then art is action. Looking from this angle, statistics may also be regarded as an art. It involves the application of given method to obtain facts, derive results and finally to use them for appropriate action.

1.4 SCOPE OF STATISTICS:

In the ancient times statistics was regarded only as the science of statecraft and was used to collect information relating to crimes, military strength, population, wealth etc. for devising military and fiscal policies. But with the concept of welfare state taking roots almost all over the world, the scope of statistics has widened to social and economic phenomenon. Moreover, with the developments in the statistical techniques during the last few decades, today statistics is viewed not only as a more device for collecting numerical data but as a means of sound techniques for their handling, analysis and drawing valid inferences from them. Accordingly it is not merely a by product of the administrative set up of the state but it embraces all sciences - social, physical, and natural and is finding numerous applications in various diversified fields such as agriculture, industry, sociology, biometry, planning, economics business, management, psychometry, insurance, accountancy and auditing and so on. It is rather impossible to think of any sphere of human activity where statistics does not creep in. It will not be exaggeration to say that statistics has assumed unprecedented dimensions these days and statistical thinking is becoming more and more indispensable every day for an able citizenship. The importance of statistics is amply explained in the following words of Carrol D. Wright (1887), United States Commissioner of the Bureau of Labour :

> "To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected ... by... of those index numbers which describe the cost of living. It is impossible to understand psychology, sociology, economics, finance or a physical science without some general idea of the meaning of an average, of variation, of concomitance, of sampling, of how to interpret charts and tables."

Scope of Statistics

Let us us now discuss briefly the importance of statistics in some different disciplines.

1. **Statistics in Planning** : Statistics is indispensable in planning may it be in business, economics or government level. The modern age is termed as the 'age of planning' and almost all organisation in the government or business or management are resorting to planning for efficient working and for formulating policy decision. To achieve this end, the statistical data relating

to production, consumption, prices, investment, income, expenditure and so on and the advanced statistical techniques, such as index numbers, time series analysis demand analysis and forecasting techniques for handling such data are of paramount importance. For efficient planning, it must be based on a correct and sound analysis of complex statistical data. For instance, in formulating a fiv e year plan, the government must have an idea of the age and sex-wise break up of the population projections of the country for the next five years in order to develop its various sectors like agriculture, industry, textiles, education and so on. This is achieved through the powerful statistical tool of forecasting by making use of the population data for the previous years. Evenfor making decisions concerning the day to day policy of the country, an accurate statistical knowledge of the age and sex-wise composition of the population is imperative for the government. In India, the use of statistics in planning was well visualised long back and the National Sample Survey (NSS) was primarily set up in 1950 for the collection of statistical data for planning in India.

2. **Statistics in State :** As had already been pointed out, with the inception of the idea of the welfare state and its taking deep roots in almost all the countries, today statistical data relating to prices, production, consumption, income and expenditure, investments and profits, etc. and statistical tools of index numbers, time series analyses, demand analysis, forecasting, etc. are extensively used by the governments in formulating economic policies. It helps the government for planning future economic programmes. The study of population movemnent. i.e., population estimates, population projections and other allied studies together with birth and death statistics according to age and sex distribution provide any administration with fundamental tools which are indispensable for over all planning and evaluation of economic and social development programmes. The facts and figures relating to births, deaths and marriages are of extreme importance to various official agencies for a variety of administrative purposes.

Mortality (death) statistics serve as a guide to the health authorities for sanitary improvements, improved medical facilities and public cleanliness. Therefore, the use of such datas are of paramount importance to health authorities in taking appropriate remedial action to prevent or control the spread of the disease. The use of statistical data and techniques is so wide in government functioning that today, almost all ministries and the departments in the government have a separate statistical unit. The main statistical agencies in India are Central Statistical Organisation (C.S.O.); National Sample Survey Organisation (NSSO) and the Registrar General of India (RGI).

3. **Statistics in Mathematics :** Statistics is intimately related to an essential dependent upon mathematics. The interaction of mathematics and statistics started in the mid-seventeenth century with the development of the theory of probability. The developments of statistical techniques and theories for application to various sciences - social, physical and natural are based on fitting different mathematical models to the observed data under certain assumptions and the whole process of such assumptions, analysis and testing is basically mathematical in character, making wide applications of mathematical tools of integration, differentiations algebra, trigonemetry, matrix theory and so on. The main stalwarts in the theory of modern statistics namely Laplace, R.A. Fisher, S.N. Roy, R.C. Bose, H. Games to mention only a few of them were/are primarily skilled and talented mathematicians. Ever increasing role of mathematician into statistics has led to the development of a new branch of statistics called mathematical statistics. In the words of cornor, "Statistics is a branch of applied mathematics which specialises in data."

4. **Statistics in Economics** : The interaction between statistics and economics was first observed by William Petty (by end of 17th Century) in his book 'Political Arithmetic'. Statistics plays a very vital role in economics so much so that in 1926, Prof. R.A. Fisher complained of the painful misapprehension that statistics is a branch of economics."

Statistical data and advanced techniques of statistical analysis have proved immensely useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty etc. For example, the studies of consumption statistics reveal the pattern of the consumption of the various commodities by different sections of the society and also enable us to have some idea about their purchasing capacity and their standard of living. The studies of production statistics enable us to strike a balance between supply and demand which is provided by the laws of supply and demand. The income and wealth statistics, are mainly helpful in reducing the disparities of income. The statistics of prices are needed to study the price theories and the general problem of inflation through the construction of the cost of living and wholesale price index numbers. The statistics of market prices, costs and profits of different individual concerns are needed for the studies of competition and monopoly. Statistics pertaining to some macro-variable, like production, income, expenditure, saving, investments, etc., are used for the compilation of National Income Accounts which are indispensable for economic planning of a country. Statistical techniques have also been used in determining the measures of Gross National Product and Input-Output Analysis.

Use of statistics in economics has led to the formulation of many economics laws some of which are mentioned below for illustratin :

a) **Engel's law of consumption (1895) :** It is a detailed and systematic study of the family budget data which gives a detailed account of the family budgets showing expenditure on the main items of family consumption together with family structure and composition, family income and various other social, economic and demographic characteristics.

b) **Law of Distribution of Income (19th-20th Century):** It is propounded by Vilfredo Pareto by making an empirical study of the income data of various countries of the world at different times.

c) **Revealed Preference Analysis :** It is formulated by Prof. Samuelson by studying the data pertaining to the actual observation of buyers in the market. Time series analysis, index numbers, forecasting techniques and demand analysis are some of the very powerful statistical tools which are used immensely in the analysis of economic data and also for economic planning. The increasing interaction of mathematics and statistics with economics led to the development of a new discipine called econometrics. Therefore, econometric models based on sound statistical analysis are used for maximum exploitation of the available resources.

5. **Statistics in Business and Management :** According to Wallis and Roberts, "Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty". A refinement over this definition is provided by Prof. Ya-Lun-Chou as follows, "Statistics is a method of decision making in the face of uncertainity on the basis of numerical data and calculated risks." These definitions reflect the applications of statistics in business since modern business has its roots in the accuracy and precision of the estimates and statistical forecasting regarding the future demand for the product, market trends and so on. The manager and a team of management executives is imperative for the efficient handling of the various operations like sales, purchase, production, marketing, control, finance, etc. of the business house. Statistical tools are used widely by business enterprises for the promotions of new business. Before embarking upon any production process, the business house must have an idea about the quantum of the product to be manufactured, the amount of the raw material and labour needed for it, the quality of the finished product, marketing avenues for the product, the competitive products in the market and so on. Thus the formulation of a production plan is a must and this cannot be achieved without collecting the statistical information on the above items without resorting to the powerful technique of 'sample surveys'. Statistical techniques have also been used very widely by business organisations in :

a. Carrying out time and motion studies (which are a part of the scientific management)

b. Marketing decisions (based on the statistical analysis of consumer preference studies - demand analysis)

c. Investment (based on sound study of individual shares and debentures).

d. Personnel administration (for the study of statistical data relating to wages, cost of living, incentives plans, effect of labour disputes/unrest on the production, performance standards, etc.)

e. Credit policy

f. Inventory control (for co-ordination between production and sales).

g. Accounting (for evaluation of the assets of the business sales).

h. Sales control (through the statistical data pertaining to market studies, consumer preference studies, trade channel studies and readership surveys, etc) and so on.

6. Statistics in Accountancy and Auditing : Statistics has innumerable applications in accountancy and auditing. For example, the statistical data on some macro-variables like income, expenditure, investment, profits, productions, savings, etc. are used for the compilation of National Income Accounts which provide informatino on the value added by different sectors of economy and are very helpful in formulating economic policies. The statistical study (corelation analysis) of profit and dividend statistics enables one to predict the probable dividends for the future years. Further, in Accountancy the statistics of assets and liabilities, and income and expenditure are helpful to ascertainthe financial results of various operations. In auditing, sampling techniques are used widely for test checking. The business transactions and the volumes of the various items comprising balances in various accounts are so heavy that it is practially impossible to resort to 100% examination and analysis of the records because of limitations of time, money and staff at our disposal. Accordingly, sampling techniques based on sound statistical and scientific reasoning are used effectively to examine thoroughly ony a sample (fraction

2% or 5%) of the transactions or the items comprising a balance and drawing inferences about the whole lot (data) by using statistical techniques of estimation and inference.

7. **Statistics in Industry :** In industry, statistics is extensively used in 'Quality Control'. The main objective in any production process is to control the quality of the manufactured product so that it conforms to specifications. This is called process control and is achieved through the powerful technique of control charts and inspection plans.

8. **Statistics in Insurance** : Probability theory on which modern theory of statistics is based is the backbone of insurance. The idea of life insurance developed during the end of the seventeenth centry after the prepraration of the life table sby Edumund Hally in 1961. Life tables are indispensable for the solution of all questinos concerning the duration of human life. Life tables, which are based on to scientific use of statistical methods (probability and mathematical expectation), are the key stone or the pivot on which the whole science of life insurance hinges. Life tables forms the basis for determining the rates of premiums and annuity necessary to various amounts of life insurance. They provide the actuarial science with a sound foundation converting the insurance business from a mere gambling in human lives to the ability to offer well calculated safeguards in the event of death. The success of an insurance company largely depends on the accuracy of the statistical data which is used for the constructions of life tables.

9. **Statistics in Astronomy** : The principle of least square, one of the most important tools in statistical theory, was developed by Gauss who used it to obtain the equation of the famous 'Normal Law of Errors' in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies.

10. **Statistics in Physical Sciences :** In physical science, a large number of measurements are taken on the same item. There is bound to be variation in these measurements. In order to have an idea about the degree of accuracy

achieved, the statistial techniques (interval estimation confidence intervals and condifence limits) are used to assign certain limits within which the true value of the phenomenon may be expected to life Today, there is an increasing use of statistics in most of the physical sciences such as astronomy, geology, engineering, physics and meterology.

11. **Statistics in Social Sciences :** Every social phenomenon is affected to a marked extent by a multiplicity of factors which bring out the variation in observations from time to time, place to place and object to object. Statistical tools of regression and correlation analysis can be used to study and isolate the effect of each of these factors on the given observation. Sampling techniques and estimation theory are very powerful and indispensable tools for conducting any social survey, pertaining to any strata of society and then analysing the results and drawing valid inferences.

    According to Croxton and Cowden -

    "Without an adequate understanding of the statistical methods, the investigator in the social sciences may be like the blind man grouping in a dark room for a black cat that is into there. The methods of statistics are useful in an over-widening range of human activities in any field of thought in which numerical data may be had."

12. **Statistics in Biology and Medical Sciences :** Sir Francis Galton (1822-1911), a British Biometrician pionered the use of statistical methods with his work on 'Regression' in connection with the inheritance of statue. According to Prof. Karl Pearson (1857-1936) who pioneered the study of 'Correlation Analysis' the whole theory of heredity rests and statistical basis. In medical sciences also, the statistical tools for the collection, presentation and analysis of observed factual data relating to the causes and incidence of diseases are of paramount importance. For example, the factual data relating to pulse rate, body temperature, blood pressure, heart beats, weight etc. of the patients greatly help the doctor for the proper diagnosis of the disease; statistical papers are used to study heart beats through electro-cardiogram (ECG). Perhaps the most important

application of statistics in medical sciences lies in using the tests of significance for testing the efficiency of a manufacturing drug, injection a medicine for controlling/curving specific ailments.

13. **Statistics in Psychology and Education :** Statistics has been used very widely in education and psychology too e.g. in the scaling of mental tests and other psychological data; for measuring the reliability and validity of test scores; for determining the intelligence quotient (I.Q.); in them analysis and factor analysis. The vast applications of statistical data and statistical theories have given rise to a new discipline called 'Psychometry'.

14. **Statistics in War :** Not only in peace times but in war times also statistics can be used very effectively. It is practically impossible to face a war without the factual data concerning the military strength of the enemy in terms of manpower, military ranks, was aeroplanes, missiles, ammunition, etc. The theory of decision functions propounded by A. Wald can be of great assitance to the military with minimum effort. Moreover, a careful and intelligent study of the statistical data obtained after the war between any two countries night reveal some very useful information which can be used for planning future military strategies in the country.

Conclusion :

Therefore, it can be concluded that "All statistics are numerical statemnts of fact but all numerical statements of facts are not statistics." Statistics is both as a science and an art - science, since it provides tools laws for the analysis of the numerical information collected from the source of enquiry and art, since it undeniably has its basis upon numerical data collected with a view to maiintain a particular balance and consistency leading to perfect or nearly perfect conclusions. A statistician like an artist will fail in his job if he does not possess the requisite skill, experience and patience while using statistical tools for any problem.

Although statistics is indispensable to almost all sciences - social, physical and natural and is very widely used in almost all spheres of human activity it is not without limitations which restricts its scope and utility.

a. Statistics does not study qualitative phenomenon like beauty, honesty, welfare etc.

b. Statistics does not study individuals. Any isolated figure cannot be regarded as statistics unless it is a part of the aggregate of facts relating to any particular field of enquiry.

c. Statistical laws are not exact. Theys are probabilistic in nature.

d. Statistics is liable to be misused.

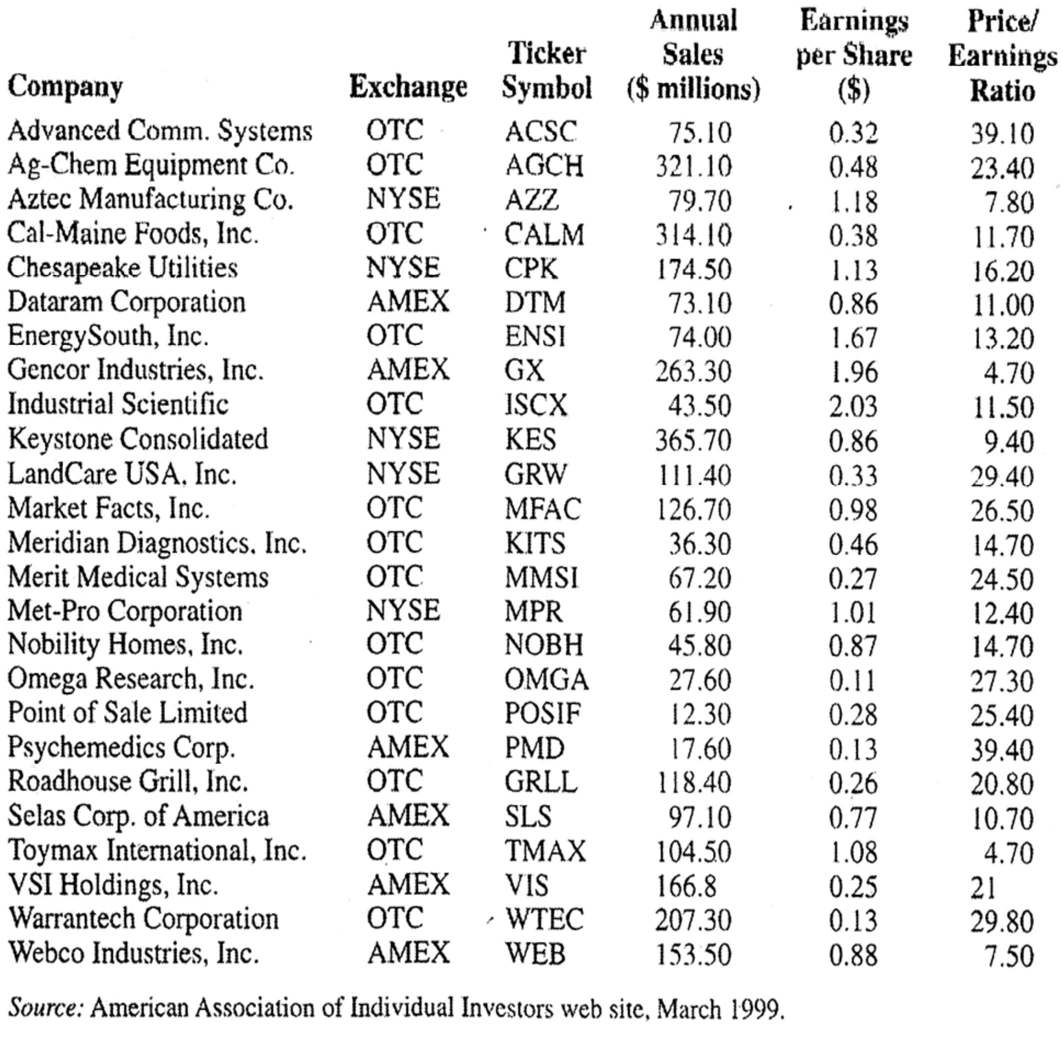
1.5 TYPES OF STATISTICS :

1.5.1 Descriptive Statistics :

Most of the statistical information in newspaper, magazines, company reports and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical or numerical are referred to as Descriptive Statistics.

For example: Table 1.1 where data on 25 shadow stocks are presented. Methods of descriptive statistics can be used to provide summaries of the information in this data set.

Table : A data set for 25 shadow stocks



Company	Exchange	Ticker Symbol	Annual Sales ($ millions)	Earnings per Share ($)	Price/ Earnings Ratio
Advanced Comm. Systems	OTC	ACSC	75.10	0.32	39.10
Ag-Chem Equipment Co.	OTC	AGCH	321.10	0.48	23.40
Aztec Manufacturing Co.	NYSE	AZZ	79.70	1.18	7.80
Cal-Maine Foods, Inc.	OTC	CALM	314.10	0.38	11.70
Chesapeake Utilities	NYSE	CPK	174.50	1.13	16.20
Dataram Corporation	AMEX	DTM	73.10	0.86	11.00
EnergySouth, Inc.	OTC	ENSI	74.00	1.67	13.20
Gencor Industries, Inc.	AMEX	GX	263.30	1.96	4.70
Industrial Scientific	OTC	ISCX	43.50	2.03	11.50
Keystone Consolidated	NYSE	KES	365.70	0.86	9.40
LandCare USA, Inc.	NYSE	GRW	111.40	0.33	29.40
Market Facts, Inc.	OTC	MFAC	126.70	0.98	26.50
Meridian Diagnostics, Inc.	OTC	KITS	36.30	0.46	14.70
Merit Medical Systems	OTC	MMSI	67.20	0.27	24.50
Met-Pro Corporation	NYSE	MPR	61.90	1.01	12.40
Nobility Homes, Inc.	OTC	NOBH	45.80	0.87	14.70
Omega Research, Inc.	OTC	OMGA	27.60	0.11	27.30
Point of Sale Limited	OTC	POSIF	12.30	0.28	25.40
Psychemedics Corp.	AMEX	PMD	17.60	0.13	39.40
Roadhouse Grill, Inc.	OTC	GRLL	118.40	0.26	20.80
Selas Corp. of America	AMEX	SLS	97.10	0.77	10.70
Toymax International, Inc.	OTC	TMAX	104.50	1.08	4.70
VSI Holdings, Inc.	AMEX	VIS	166.8	0.25	21
Warrantech Corporation	OTC	WTEC	207.30	0.13	29.80
Webco Industries, Inc.	AMEX	WEB	153.50	0.88	7.50

Source: American Association of Individual Investors web site, March 1999.

Figure : Bar graph of the exchange variable

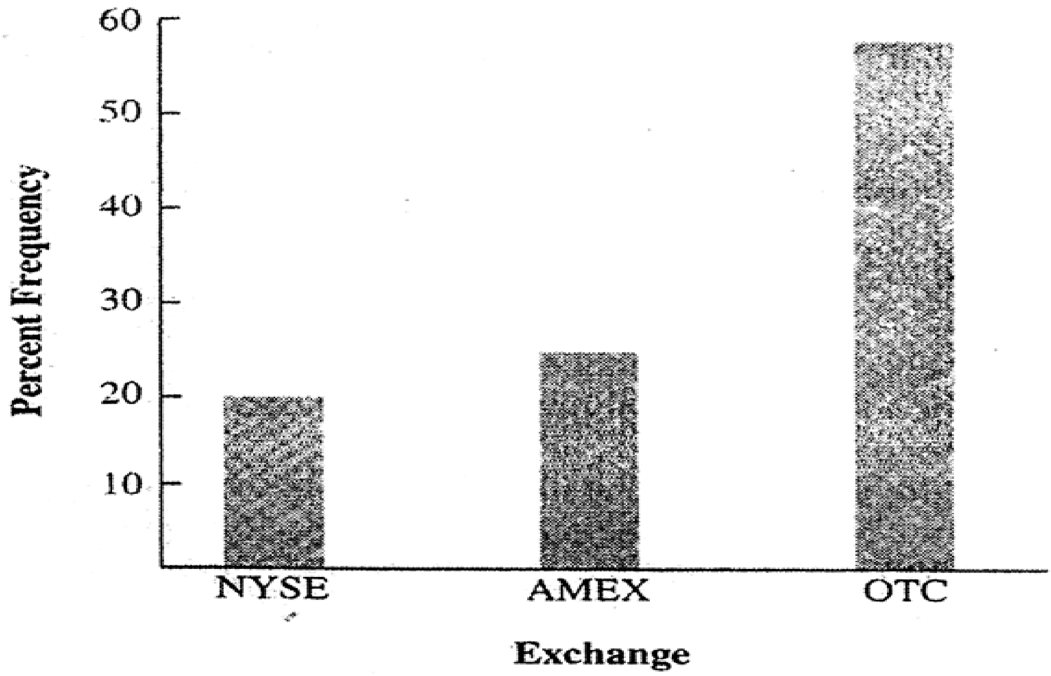


Table : Frequencies and percent freqnencies for the exchange variable

Exchange	Frequency	Percent Frequency
New York Stock Exchange (NYSE)	5	20
American Stock Exchange (AMEX)	6	24
Over-the-counter (OTC)	14	56
Totals	25	100

A tabular summary of the data for the qualitative variable exchange is shown in table 1.4 A graphical summary of the same data, called a bar graph, is shown in fig. 1.5. The purpose of these types of tabular and graphical summaries is to make the data easier to interpret. Referring to table 1.4 and figure 1.5, we can see easily that the majority of the stocks in the data set are traded over the counter.

In addition to tabular and graphical displays, numerical descriptive statistics are used to summarize data. The most common numerical descriptive statistic is the average or mean. Average is taken as a measure of the central tendency or central location of the data.

In recent years interest has grown in statistical methods that can be used for developing and presenting descriptive statistics.

1.5.2 Statistical Inference :

In many situations, data are sought for a large group of elements (individuals, companies, voters, households, products, customers, and so on). Because of time, cost, and other considerations, data are collected from only a small portion of the group. The larger group of elements in a particular study is called the population and the smaller group is called the sample. Formally, we use the following defintions.

Population : A population is the set of all elements of interest in a particular study.

Sample : A sample is a subset of the population.

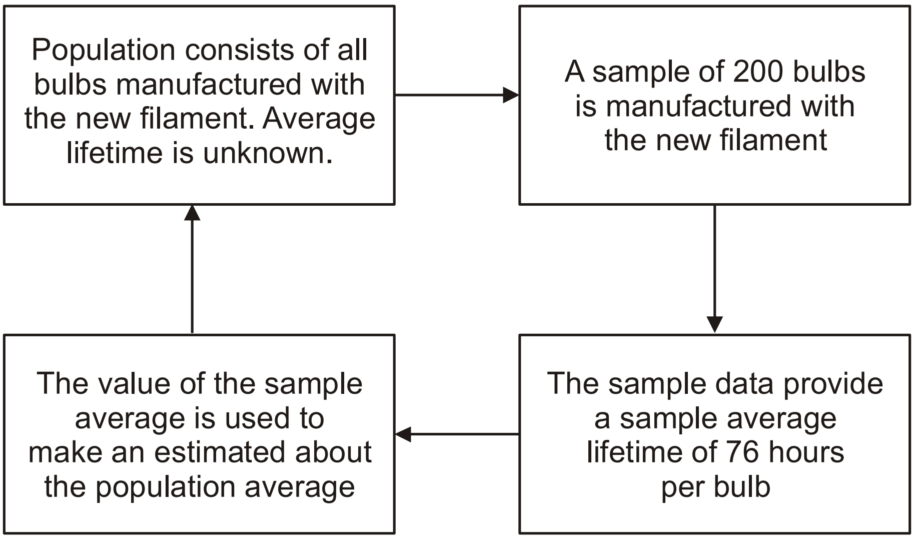
A major contribution of statistics is that data from a sample can be used to make estimates and test hypotheses about the characteristics of a population. This process is referred to as statistical inference. As an example of statistical inference, let us consider the study conducted by Norris Electronics. Norris manufactures a high-intensity light bulb used in a variety of electrical products. In an attempt to increase the useful life of the light bulb, the product design group has developed a new light bulb filament. In this case, the population is defined as all lightbulbs that could be produced with the new filament, 200 bulbs with the new filament were manufactured and tested. Data were collected on the number of hours each lightbulb operated before filament burnout. The data from this sample are reported in following table

Table : Hours until burnout for a sample of 200 lightbulbs for norris electronics

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

Suppose Norris is interested in using the sample data to make an inference about the average hours of useful life for the population of all lightbulbs that could be produced with the new filament. Adding the 200 values in above table and dividing the total by 200 provides the sample average lifetime for the lightbulbs : 76 hours. We can use this sample results to estimate that the average lifetime for the lightbulbs in the population is 76 hours. Figure 1.7 is a graphical summary of the statistical inference process for Norris electronics.

Fig. : The process of statistical inference for the Norris Electronics example


Population consists of all bulbs manufactured with the new filament. Average lifetime is unknown.
A sample of 200 bulbs is manufactured with the new filament
The sample data provide a sample average lifetime of 76 hours per bulb
The value of the sample average is used to make an estimated about the population average

Whenever statisticians use a sample to estimate population characteristics of interest, they usually provide a statement of the quality, as precision, associated with the estimate. For the Norris example, the statistician might state that the estimate of the average lifetime for the population of new lightbulbs is 76 hours with a margin of error of ± 4 hours. Thus, an interval estimate of the average lifetime for all lightbulbs produced with the new filament is 75 hours to 80 hours. The statistician can also state how confident he or she is that the interval from 72 hours to 80 hours contains the population average.

1.6 LESSON END EXERCISE :

Q1. Explain critically a few of the definitions of statistics and state the one which you think to be the best.

Q2. Define the term 'Statistics' and discuss its use in business and trade.

Q3. Describe the term "Descriptive Statistics" and "Statistical Inference".

Q4. Discuss briefly the importance of statistics in the following disciplines:

i) Mathematics

ii) Economics

iii) Business Management

iv) Planning

v) Accountancy and Auditing

vi) Astronomy

vii) Insurance

viii) Industry

ix) Biology

x) Sociology

ix) Medical Sciences

Q5. Comment on "Statistics as Science as Art".

Q6. Indicate if the following statements are true (T) or false (F).

i. The subject of statistics is a century old.

ii. The word statistics seems to have been derived from Latin Word status.

iii. Statistics is no use to humanity.

iv. To a very striking degree, our culture has become a statistical culture.

v. Statistics can prove anything.

Ans. i) F; (ii) T; (iii) F; (iv) T; (v) F

****

M.A. Economics **Semester - IInd** **Lesson - 2**

Course No. 203 **Unit - I**

MEASURES OF CENTRAL TENDENCY

CHAPTER HIGHLIGHTS:

This lesson contains the information technology regarding to the various measures of central tendency.

CHAPTER OBJECTIVES:

2.1 Introduction

2.2 Objectives

2.3 Essentials or properties of a good average

2.4 Measures of central tendency

- 2.4.1 Arithmetic Mean
- 2.4.2 Median
- 2.4.3 Other Positional Measures
- 2.4.4 Mode
- 2.4.5 Geometric Mean
- 2.4.6 Harmonic Mean
- 2.4.7 Let us sum up
- 2.4.8 Lesson End Exercise

2.1 INTRODUCTION:

An important objective of the statistical analysis is to get a value out of the collected data which give us an idea about the basic features of the frequency distribution that we have tables, graphs or diagram do not give us an idea about the exact picture of that. So, we resort to the measures of central tendency for that purpose.

The measures of central tendency refers to those methods which gives us a value out of the collected observation around which all the values of the frequency distribution tend to centralise. That value is always in between the lowest and the highest observation. This value is also known as 'an average', which represents the whole set of date. It is always in that unit of measurement in which original data is given.

2.2 OBJECTIVES:

Following are the main objectives of 'an average' or the measures of central tendency.

i. to know about the features of the given means of data.

ii. to make meaningful comparison.

iii. to help in taking correct decisions.

2.3 ESSENTIALS OR PROPORTIONS OF A GOOD AVERAGE:

An average, which represents a group of values, should have following properties :

i. Easily understandable

ii. Simple to calculate

iii. Clearly defined

iv. Not affected by extreme values

v. Based on all observations

vi. Amenable to further algebric treatment

vii. Rigidly defined

2.4 MEASURES OF CENTRAL TENDENCY

2.4.1 Arithmetic Mean: Arithmetic mean, also known as 'Mean' or 'Common Average', is the number that we get after dividing the sum total of all observations with the number of observations. It is most commonly used measures of central tendency, which is very easy to compute.

Symbolically,

$$\text{Arithmetic Mean}: \overline{x} = \frac{x_1 + x_2 + x_3 + \ldots.. + x_n}{N} = \frac{\sum x_i}{N} \text{ or } \frac{\sum x}{N}$$

where i = 1, 2, 3,.....N

$$\text{i.e. } \overline{x} = \frac{\text{Sum total of the observations}}{\text{Number of the observations}}$$

a. Calculation of arithmetic mean in an individual series :

i. Direct method: Two steps are taken to calculate arithmetic mean in direct method:

Step I - Add-up all the observations

Step II: Divide the total by the number of observations.

$$\text{The formula } \overline{x} = \frac{\sum x}{N}$$

Example:

Calculate arithmetic mean from the following:

10, 20, 30, 40, 50

Solution:

$$\overline{x} = \frac{10 + 20 + 30 + 40 + 50}{5} = \frac{150}{5} = 30 \text{ Ans}$$

ii. Short Cut Method: Following steps are taken to calculate arithmetic mean is short cut method.

Step I : Assume any number as mean (A) among the observation.

Step II: Find deviations (dx) by subtracting that assumed mean from all observations an add these up.

Step III: Divide the sum of deviation by the number of observations and add the assumed mean to this.

The formula used is $\overline{x} = A + \frac{\sum dx}{N}$

Example:

Calculate arithmetic mean using short cut method:

10, 20, 30, 40, 50

Solution:

X	A = 30
	dx = X-A

10	- 20
20	- 10
30	0
40	+ 10
50	+ 20
	$\sum dx = 0$

$$\overline{x} = A + \frac{\sum dx}{N}$$

$$= 30 + \frac{0}{5}$$

$$= 30 + 0$$

$\therefore$ $\overline{x} = 30$ Ans

iii) Step Divation Method : Following steps are taken to find arithmetic using step deviation method.

Step I: Assume any number as mean (A) among the given observations.

Step II: Find deviations (dx) by subtracting the assumed mean (A) from all the observations.

Step III: Divide all the deviation (dx) by a common factor (c), to get step-deviations (d^1x) and add these up.

Step IV: Divide total of step deviation by number of observations multiply the quotient by the common factor and add the assumed mean to it.

The formula used is $\bar{x} = A + \left(\frac{\sum d^1x}{N} x\, C \right)$

Example :

X	A = 20
	dx = X-A
		$d^1x = \frac{dx}{c}$
		c = 10

10	- 10	- 1
20	0	0
30	+ 10	+ 1
40	+ 20	+ 2
50	+ 30	+ 3
		$\sum d^1x = +5$

$$\bar{x} = A + \left(\frac{\sum d^1x}{N} x\, C \right)$$

$$= 20 + \left(\frac{5}{5} x\, 10 \right)$$

$$= 20 + 10$$

$\therefore$ $\bar{x} = 30$ Ans

b) Calculations of arithmetic in a discrete series:

i) Direct Method: Following steps are taken to calculae arithmetic using direct method.

Step I: Multiply each observation (X value) with its respective frequency and add up these.

Step II: Divide this total by the sum of the frequencies.

The formula used is $\bar{x} = \frac{\sum fx}{\sum f}$

Example:

Calculate arithmetic mean from the following:

Marks:	10	20	30	40	50
No. of Students:	2	4	7	5	3

Solution:

Marks (x)	No. of students (f)	fx
10	2	20
20	4	80
30	7	210
40	5	200
50	3	150
	$\sum = 21$	$\sum fx = 660$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{660}{21}$$

$$\therefore \quad \bar{x} = 31-43$$

ii) Short cut method: Following steps are used to calculate arithmetic mean using short cut method.

Step I: Assume any number as mean (A) among the values of the variables/ observations.

Step II: Subtract that assumed mean from each of the observation/variable to get deviations (dx)

Step III: Multiply these deviations with their respective frequencies and add up these.

Step IV: Divide the sum of that with the sum of the frequencies and add assumed mean so that.

The formula used is $\overline{x} = A + \frac{\sum fdx}{\sum f}$

Example:

Marks (x)	No. of students (f)	dx = x-A
		A = 30
			fdx

10	2	-20	-40
20	4	-10	-40
30	7	0	0
40	5	+10	50
50	3	+20	60
	$\sum f = 21$		$\sum fdx = +30$

$$\overline{x} = A + \frac{\sum fdx}{\sum f}$$

$$= 30 + \frac{30}{21}$$

= 30 + 1.43

$\therefore \quad \overline{x} = 31.43 \text{Ans}$

iii) Step deviation method : Following steps are taken to calculate arithmetic using the step deviation method :

Step I: Assume any number as mean (A) among the values of the variable/ observations.

Step II: Subtract tha assumed mean (A) from each of the observation/ variable to get the deviations (dx)

Step III: Divide all the deviations by a common factor (c) to get step deviations (d^1x)

Step IV : Multiply these step-deviations (d^1x) with their respective frequncies and up these.

Step V: Divide the sum of that with the sum of the frequencies, multiply the quotient by the commond factor (c) and add assumed mean so that.

The formula used is $\bar{x} = A = \left(\frac{\sum fd^1x}{\sum f} x\ c \right)$

Example :

Marks (x)	No. of students (f)	DX = x - A
		A = 20
			$d^1x = \frac{dx}{c}$
			c = 10
				fd^1x

10	2	-10	-1	-2
20	4	0	0	0
30	7	+10	+1	7
40	5	+20	+2	10
50	3	+30	+3	9
	$\sum f = 21$			$\sum fd^1x = 24$

$$\bar{x} = A = \left(\frac{\sum fd^1x}{\sum f} x\ c \right)$$

$$= 20 + \left(\frac{24}{21} x10 \right)$$

=20 + (1.143 x 10)

= 20 + 11.43

$\therefore \quad \bar{x} = 31.43 \text{Ans}$

c) Calculation of arithmetic mean in a continuous series

i) Direct Method: Steps to find arithmetic mean same as in case of discrete series. Take x as mid points of the class intervals.

Example:

Find arithmetic mean from the following data:

Marks :	0 - 10	10-20	20-30	30-40	40-50
No. of Students :	2	4	7	5	3

Solution:

Marks	No. of students (f)	Mid points (x)	fx
0-10	2	5	10
10-20	4	15	60
20-60	7	25	175
30-40	5	35	175
40-50	3	45	135
	$\sum f = 21$		$\sum fx = 555$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{555}{21}$$

$\therefore \quad \bar{x} = 26.43$ Ans

ii) Short cut method : Steps to find arithmetic mean same as in case of discrete series. Take x as mid-poins of the class intervals.

Example:

Marks	No. of students (f)	Mid-points (x)	A= 25
			dx= x-A
				fdx

0-10	2	5	-20	-40
10-20	4	15	-10	-40
20-30	7	25	0	0
30-40	5	35	+10	50
40-50	3	45	+20	60
	$\sum f = 21$			$\sum fdx = 30$

$$\bar{x} = A + \frac{\sum fdx}{\sum f}$$

$$= 25 + \frac{30}{21}$$

$$= 25 + 1.43$$

$$\therefore \quad \bar{x} = 26.43 \text{ Ans}$$

iii) Steps deviations method : Steps to find arithmetic mean same as in case of discrete series. Take x as mid points of the class intervals.

Example:

Marks	No. of students (f)	Mid points (x)	A = 15 dx=x-A	$d^1x = \frac{dx}{c}$	fd^1x
0-10	2	5	-10	-1	-2
10-20	4	15	0	0	0
20-30	7	25	+10	+1	7
30-40	5	35	+20	+2	10
40-50	3	45	+30	+3	9
	$\sum f = 21$				$\sum fd^1x = 24$

$$\bar{x} = A + \left(\frac{\sum fd^1x}{\sum f} \text{ x c} \right)$$

$$= 15 + \left(\frac{24}{21} \text{ x } 10 \right)$$

$$= 15 + (1.143 \text{ x } 10)$$

$$= 15 + 11.43$$

$$\therefore \quad \bar{x} = 26.43 \text{ Ans}$$

Combined Mean: Combined mean of two or more series can be found without going back to original data if we have individual arithmetic mean and number of observations of each series, using the following formula:

Combined arithmetic mean $\overline{x}_c = \dfrac{N_1\overline{x}_1 + N_2\overline{x}_2 + ...+ N_n\overline{x}_n}{N_1 + N_1 + ...+ N_n}$

where $\overline{x}_1, \overline{x}_2......, \overline{x}$ = Arithmetic mean of individual series

N_1, N_2,.......N_n = Number of observations in each individual series.

Example:

The mean marks of 50 first year students are 30 and that of 40 second year students are 2.5 Find the average marks of all 90 students.

Soltuion:

$N_1 = 50, \overline{x}_1 = 30, N_2 = 40, \overline{x}_2 = 25$

Combined mean $\overline{x}_c = \dfrac{N_1\overline{x}_1 + N_2\overline{x}_2}{N_1 + N_2}$

$$= \frac{(50x30) + (40x25)}{50 + 40}$$

$$= \frac{1500 + 1000}{90}$$

$$= \frac{2500}{90} = 27.78$$

Thus average marks of all 90 students are 27.78

Connecting wrong arithmetic mean : Sometimes, mistanely or for any reason, some values are miscreat and wrongly included in the calculation of arithmetic mean. To correct that, following steps are taken:

Step I: Calculate incorrect sum of observations $(\sum x)$ by multiplying wrong arithmetic mean given with the number of observations.

Step II: Subtract wrong or miscread values and add correct values to get correct sum of observations ($\sum x$)

Step III: Divide correct sum of observations by the number of observations to calculate correct arithmetic mean.

Example:

Arithmetic mean of 40 items is 20. However, two items 3 and 16 were miscread as 13 and 26. Find the correct arithmetic mean if (i) it is an individual series (ii) it is a discrete series (iii) it is a continuous series with class intervals as 0-5, 5-10, 10-15 and so on.

Solution:

For (i) and (ii) In case of an individual and discrete series, the incorrect $\sum x = N\bar{x} = 40 \text{ x } 20 = 800$

The correct $\sum x = 800 - 13 - 26 + 3 + 16 = 780$

$\therefore$ Correct $\bar{x} = \dfrac{\text{Correct} \sum x}{N} = \dfrac{780}{40} = 19.5$ Ans

For (iii) in a continuous series, with class intervals as 0-5, 5-10, 10-15...., the wrong or miscread values 13 and 26 lie in the class intervals 10-15 and 25-30 respectively whose mid-values/mid-points are 12.5 and 27.5 respectively. On the other hand, the correct values 3 and 16 lie in the class intervals 0-5 and 15-20 respectively whose mid-points are 2.5 and 17.5 respectively. Now,

Incorrect $\sum x = N\bar{x} = 40 \text{ x } 20 = 800$

Correct $\sum x = 800 - 12.5 - 27.5 + 2.5 + 17.5 = 780$

$\therefore$ Correct $\bar{x} = \dfrac{\text{Correct} \sum x}{N} = \dfrac{780}{40} = 19.5$ Ans

Weighted Arithmetic Mean : Many a times, sometimes in a distribution carry more importance than others. Therefore, simple arithmetic mean unit be a true representative of average. Hence, weights are assigned to items as per their importance and weighted arithmetic mean is found.

Following steps are used to find weighted arithmetic mean:

Step I: Multiply weight (w) with the variable/observations

i.e. x and get $\sum wx$ by adding up these.

Step II: Divide this total by the sum of the weight $(\sum w)$

$\therefore$ Weighted arithmetic mean = $\frac{\sum wx}{\sum w}$

Example :

Following are marks of two candidates also applied for scholarship. If a candidate with highest marks is to be given the scholarship, such who would be awarded that ?

Subject	Weights (W)	Marks		wx$_1$	wx$_2$
		A (x_1)	B (x_2)		
Maths	1	70	52	70	52
Economics	4	63	65	252	260
Computer	3	65	70	195	210
Statistics	2	58	63	116	126
	$\sum w = 10$			$\sum wx_1 = 633$	$\sum wx_2 = 648$

Weighted marks of A = $\frac{633}{10}$ = 63.3, weighted marks of B = $\frac{648}{10}$ = 64.8

$\therefore$ Scholarship should be given to candidate B as be is having highest marks.

Advantage of arithmetic mean :

1. Simple to understand
2. Easy to calculate
3. Based upon all observations

4. Can be used further in future and analysis
5. Rigidly defined and have a definite value.
6. Satisfies most of the characteristics of a good average.

Disadvantages of arithmetic mean :

1. Unduly affected by extreme observations.
2. Difficult to calculate in case of open-end intervals.
3. Cannot be found graphically as like median and mode.
4. Cannot be calculated if even one value is missing.
5. May not correspond to any observation in the series.
6. May give meaning less results take we may get average number of children in a village as 2.7 which is meaningless.

2.4.2 Median

Median (m) is that value in the distribution which divides a series into two equal parts one above it and one below it when the observations are arranged in ascending or descending order. The number of items/observations above and below the median value are equal. It is a positional measure which refers to the place of value in a series i.e. the middle one.

Calculation of Median in an individual series:

Step I: Arrange the observations in ascending or descending order.

Step II: Find the middle term using the formula

Median = Size of the $\left(\frac{N+1}{2}\right)$ th item

The observation at the $\left(\frac{N+1}{2}\right)$ th serial number is median.

Step III: If size of $\left(\frac{N+1}{2}\right)$th item is infractions i.e. whose number of items N is even, then the average of $\left(\frac{N}{2}\right)$th item and $\left(\frac{N+2}{2}\right)$th item is the median.

Example :

Find median from the following :

15, 18, 16, 13, 22

Sr. No.	X
1	13
2	15
3	16
4	18
5	22
	N=5

M = Size of the $\left(\frac{N+1}{2}\right)$th item

M = Size of the $\left(\frac{5+1}{2}\right)$th item

M = Size of the $\frac{6}{2}$ th item

M = Size of the 3rd item

Now 3rd item = Observation at serial no. 3 = 16

$\therefore$ M = 16 Ans

Example :

Find median from the following:

15, 18, 16, 13, 22, 20

Sr. No.	x
1	13
2	15
3	16
4	18
5	20
6	22
N=6	

$$M = \text{Size of the } \left(\frac{N+1}{2}\right) \text{th item}$$

$$M = \text{Size of the } \left(\frac{6+1}{2}\right) \text{th item}$$

$$M = \text{Size of the } \frac{7}{2} \text{ th item}$$

$$M = \text{Size of the 3.5th item}$$

$$\text{Now 3.5th item} = \frac{\text{3rd item} + \text{4th item}}{2}$$

$$\therefore \quad M = \frac{16+18}{2} = \frac{34}{2} = 17$$

$\therefore$ M = 17 Ans

Median in a discrete series:

Step I: Find cummulative frequencies (cf) from the given frequencies (f)

Step II: Find the middle term using the formula

$$M = \text{Size of the } \left(\frac{N+1}{2}\right) \text{th item,}$$

Where N = sum of the frequencies

Step III: Find that cumulative frequency which is either equal to the 'size of the $\left(\frac{N+1}{2}\right)$th item' or next higher to that.

Step IV: The variabe (x) corresponding to that cumulative frequency is median.

Example :

(x)	No. of students (f)	c.f.
5	4	4
10	6	10
15	10	20
20	8	28
25	7	35
30	3	38
	N = 38	

M = Size of the $\left(\frac{N+1}{2}\right)$th item

M = Size of the $\left(\frac{38+1}{2}\right)$th item

M = Size of the $\left(\frac{39}{2}\right)$th item

M = Size of the 19.5th item

Now 20 is the cumulative frequency which is next higher to the 19.5th item and correponding variable marks are 15

$\therefore$ M = 15 Ans.

Median in a continuous series :

Step I : Find cumulative frequency (cf) from the given frequencies (f)

Step II: Find the Median class using the formula

M = Size of the $\frac{N}{2}$ th item

Step III: Find that cumulative frequency which is either equal to 'size of the $\frac{N}{2}$ th item' or just next higher to that.

Step IV: The class interval corresponding to that cumulative frequency is median class.

Step V: The median lies in the median class. Find the median using the formula

$$M = L_1 + \left(\frac{\frac{N}{2} - C.f.}{f} \text{ x } i \right)$$

Where L_1 = Lower limit of median class

N = Sum of frequencies

c.f. = Cumulative frequency of pre-median class

f = Frequency of Median class

i = Difference between upper and lower limit of median class interval

Example:

Find Median from the following:

Marks :	0-10	10-20	20-30	30-40	40-50
No. of students:	2	4	10	7	3

Marks	No. of Students	c.f.
0-10	2	2
10-20	4	6
20-30	10	16
30-40	7	23
40-50	3	26
	N = 26	

$M = \text{Size of the} \left(\frac{N}{2}\right) \text{th item}$

$M = \text{Size of the } \frac{26}{2} \text{ th item}$

M = Size of the 13th item

Now 16 is the cumulative frequency which is next higher to the 13th item and corresponding class interval is 20-30. Therefore median class 20-30

Now $M = L_1 + \left(\frac{\frac{N}{2} - C.f.}{f} \text{ x } i\right)$

$= 20 + \left(\frac{13-6}{10} \text{x} 10\right)$

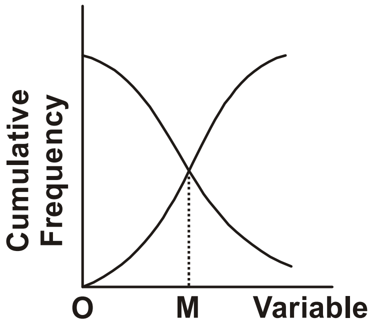
$= 20 + 7 = 27$

$\therefore$ Median marks = 27 Ans

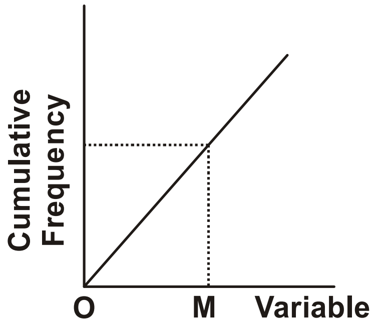
Finding the Median graphically:

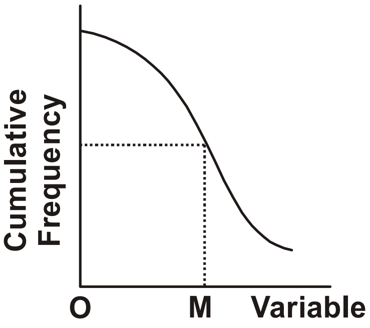
Graphically, medican can be found in the following ways:

a) The intersecting point of the 'less than O give' and 'Move than O give'. then draw perpendicular on X-axis. The value where that prependicular touches X-axis is the median.



b) Find the 'Size of the $\left(\frac{N}{2}\right)$th item'. Draw a horizontal line from y-axis from the point of '$\frac{N}{2}$th item' to the 'less than O give' or 'More than O give'. Then draw a perpendicular from that point on X-axis. The value where that perpendicular touches the x-axis is median.





Merit of Median :

1. Useful in case of open and intervals.
2. Not affected by extreme values.
3. Easy to calculate.
4. Simple to understand.
5. Can also be found graphically.
6. Based upon all observtions
7. Useful in care of markedly skewed destinations.

Demerits of Median :

1. Not anenable to algebric treatment like arithmetic mean.
2. Difficult to arrange large number of observations in ascending or descending order to find median.
3. Erratic if number of observations is small.

4. More affected by sampling fluctuations.

Median in case of unequal class intervals :

If we were given unequal class intervals, then we need not make these equal before finding median. It can be directly found from the class intervals as given.

2.4.3 Other Positional Measures

Quartiles : Quartiles are those values whch divide series into four equal parts. There are three quartiles first quartile (Q_1), Second Quartile (Q_2) and the third quartile (Q_3). Second Quartile (Q_2) is same as the median. Q_1 is that value in a series of which 25 percent observations are below and 75 percent observations are above Similarly, Q_3 is that value in a series of which 75 percent values are below and 25 percent values are above.

Deciles :

Deciles are those values which divide a series into ten equal parts. There are nine deciles in a series, devoted as D_1, D_2,...,D_a. D_1 is first decile in a series of which 10 percent observations are below and 90 percent observtions are above. Same pattern is for other deciles.

Percentiles:

Percentiles are those values which divide a series into 100 equal parts. There are 99 percentiles in a series, denoted as P_1, P_2....,P_{qq}. Now P_1 is a first percentile in a series of which one percent observation are below and 99 percent observation are above.

Calculation of quartiles, deciles and percentiles

Quartiles, deciles and percentiles are calculated just like we calculate the median. The steps are same and only values in the formula change.

Measure	Individual and Discrete series	Continuous series	Formula in continuous series
(Q_n)	Size of the	Size of the	
Quartiles	$\frac{n(N+1)}{4}$ th item	$\frac{nN}{4}$ th item	$L_1 + \left(\frac{\frac{nN}{4} - cf}{f} \text{ x i} \right)$
(Q_1, Q_2, Q_3)	where	n = 1, 2, 3	
(D_n)	Size of the	Size of the	
Deciles	$\frac{n(N+1)}{10}$ th item	$\frac{nN}{10}$ th item	$L_1 + \left(\frac{\frac{nN}{10} - cf}{f} \text{ x i} \right)$
$(D_1, D_2, ..., D_q)$	where	n = 1, 2, 3...9	
(P_n)	Size of the	Size of the	
Percentiles	$\frac{n(N+1)}{100}$ th item	$\frac{nN}{100}$ th item	$L_1 + \left(\frac{\frac{nN}{100} - cf}{f} \text{ x i} \right)$
$(P_1, P_2, ..., P_{qq})$	where	n = 1, 2, 3...9	

Graphically also, the quartiles, deciles and percentiles are found as the median but by drawing or using one Ogive only either the 'Less than Ogive' or 'More than Ogive'

2.4.4 MODE

Mode (z) is that value in a series which occurs most number of time i.e. repeated maximum number of times. The word 'Mode' has been derived from the french word 'la mode' which means the most popular phenomenon. It is very easy to calculate as the observation which occurs most number of times is mode.

Calculation or methods of finding mode :

Mode in an individual series : In case of an individual series, that observation is mode which is repeated maximum number of times. This is done by inspection method.

Example :

Find mode from the following :

2, 5, 3, 5, 6, 2, 5, 4, 5, 6, 1, 0, 2, 5, 3, 5, 1, 5

Solution :

By inspecting the observations, we find that the observation 5 has been repeated seven time. Hence mode is 5 or Z=5 ans.

Calculation of mode in a discrete series :

In a discrete series, that value of variable is mode whose frequency is highest i.e. which is repeated maximum number times. This is done by inspection method.

Example :

Find mode from the following :

Marks :	5	10	15	20	25	30
No. of students :	3	4	8	11	9	2

Solution :

Using inspection method, we see that maximum number of students i.e 11 students have got 20 marks : Z=20 Ans.

Calculation of mode in a continuous series :

In a continuous series, one need to follow the 'grouping method' to find the mode as it would be difficult, using inspection method to find the modal class interval and then made. Hence use of grouping method is recommended for that.

Grouping Method of Finding Mode :

In grouping method, a grouping table of six columns is made. The first column is the column of frequencies itself. Then frequencies are added in the groups of two frequencies in second and third column and in groups of three frequncies in the fourth, fifth and sixth columns. Following are main steps for that:

STEP I : The column of frequencies is the first column. Excircle the highest frequency.

STEP II : In second column, add up frequencies in groups of two frequencies, starting from the first one. Excircle the highest group total. Excircle two or more group totals if the totals are same.

STEP III : In third column, add up frequencies in groups of two frequencies, starting from second frequency and leaving the first one. Excircle the highest group total.

STEP IV : In fourth column, add up frequencies in the groups of three frequencies, starting from first frequency. Excircle the highest group total.

STEP V: In fifth column, add up frequencies in the groups of three frequencies each, starting from second frequency and leaving the first one. Excircle the highest group total.

STEP VI : One sixth column, add up frequencies in the groups of three frequencies each, starting from third frequency and leaving the first two frequencies. Excircle the highest group total.

In all the steps, if one or two frequencies are left in the end which cannot be added in groups of 2's and 3's, then leave these.

STEP VII : Now prepare another table of six columns and tick against the class intervals whose frequency or total is highest. Count the number of ticks against each class interval. That class interval is the 'Modal class' which has the highest number of ticks.

STEP VIII: Now find mode (z) using the formula :

$$Z = L_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \text{ x i} \right)$$

Where L_1 = Lower limit of modal class

f_0 = Frequency of pre-modal class

f_1 = Frequency of modal class

f_2 = Frequency of post-modal class

i = difference between the lower and upper limit of modal class.

OR

Another formula is

$$Z = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x i} \right)$$

Where L_1 = Lower limit of modal class

$\Delta_1 = |f_1 - f_0|$ = difference between/frequencies of modal and pre-modal classes, ignoging the signs.

$\Delta_2 = |f_1 - f_2|$ = difference between/frequencies of modal and post-modal classes, ignoring the signs.

i = difference between the lower and upper limit of modal class

Example :

Find mode from the following :

Marks :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No.of students:	20	24	32	28	20	16	34	10	8

Marks	No. of students (f)	II	III	IV	V	VI
0-10	20	44		76		
10-20	24 f_0		56		84	
20-30	32 f_1	60				80
30-40	28 f_2		48			
40-50	20	36		64		
50-60	16		50		70	
60-70	34	44		52		60
70-80	10		18			
80-90	8					

ANALYSIS TABLE

Marks	I	II	III	IV	V	VII	Total
0-10				√			1
10-20			√	√	√		3
20-30		√	√	√	√	√	5
30-40		√			√	√	3
40-50						√	1
50-60							0
60-70	√						1
70-80							0
80-90							0

The class interval 20-30 has highest number of ticks against and this is the modal class. Note that 60-70 class interval has the highest frequency but 20-30 class interval has turned out to be the modal class. This shows that inspection method cannot be fully relied upon in case of continuous series and it is essential to use the grouping method.

Now mode $(Z) = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \text{ x i}$

$$= 20 + \left(\frac{32 - 24}{(2 \text{ x } 32) - 24 - 28} \text{ x } 10 \right)$$

$$= 20 + \left(\frac{8}{12} \text{ x } 10 \right)$$

$$= 20 + \frac{80}{12}$$

$= 20 + 6.67$

$= 26.67$

$\therefore$ Z = 26.67 marks Ans.

When even group method is inclusive :

Example :

Find mode from the following :

GROUPING TABLE

Marks	No. of I students (f)	II	III	IV	V	VI
0-10	7	17		32		
10-20	10		25		38	
20-30	15	28				36
30-40	13		21			
40-50	8	22		35		
50-60	14		29		37	
60-70	15	26				40
70-80	11					

ANALYSIS TABLE

Marks	I	II	III	IV	V	VII	Total
0-10							0
10-20					√		1
20-30	√	√			√		3
30-40		√		√	√		3
40-50				√			1
50-60			√	√		√	3
60-70			√			√	2
70-80						√	1

Now in this assymmetrical distribution, each the grouping method is inconculsive in determining the modal class. Hence, the following formuala should be used to find mode.

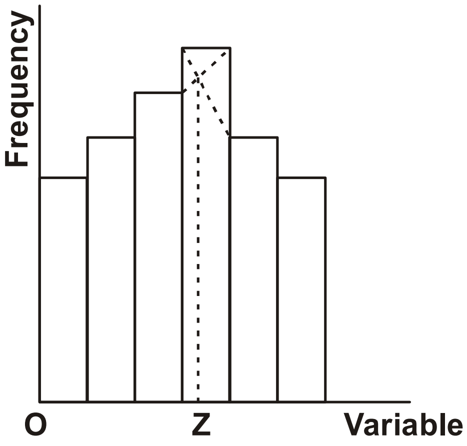
Mode = 3 Median - 2 Mean or $Z = 3m - 2\bar{x}$

Method to find mode graphically :

STEP I : Draw a histogram of the given data.

STEP II : Draw two lines diagonally inside the highest bar (that of modal class) from upper concerns of the bar to the upper corner of the adjoining bars.

STEP III: Draw a perpendicular line from the intersecting lines tothe X-axis. That value of the variable is mode where that particular touches the X-axis.



Mode in case of unequal class intervals:

If we are given unequal class intervals as 0-10, 10-20, 20-40, 40-70, 70-80 etc., then we need to make these unequal class interval as equal first. Then proceed to calculate the mode

Merits of Mode :

1. More authentic indicator of average
2. Easy to calculate
3. Easy to understand
4. Not affected by extreme values

5. Can be located graphically also.

Demerits of mode :

1. Not amenable to algebric treatment.
2. Not rigidly defined.
3. Not based on all items.
4. Cannot be determined in case of bi-modal series.

2.4.5 Geometric Mean (G.M.)

Geometric mean is defined as the '4th root' of the product of the x number of items in a series. Symbolically, the G.M. of X_1, X_2,.....,X_N is given by the expression.

$$\mathrm{GM} = \sqrt[n]{X_1, X_2, X_3......, X_n}$$

or $$\mathrm{GM} = \left(X_1, X_2, X_3......, X_n\right)^{\frac{1}{N}}$$

Where N= Number of items

Using logritamine method of solving this, we get

$$\mathrm{GM} = \text{antilog}\left[\frac{\log X_1 + \log X_2 + ... + \log X_n}{N}\right]$$

$$\therefore \quad \mathrm{GM} = \text{antilog}\left[\frac{\sum \log X}{N}\right]$$

Calculation of geometric mean in an individuals series

STEP I : Take log of all the observations of the variable X and add up these to get $\sum \log X$.

STEP II : Divide $\sum \log X$ by number of items N.

STEP III: Take antilog of the quotient to get GM.

Example :

Find the geometric mean from the following :

7, 38, 176, 500, 286, 79, 250

Solution :

X	logX
7	0.8451
38	1.5798
176	2.2455
500	2.6990
286	2.4564
79	1.8976
250	2.3979
N =7	$\sum \log X = 14.1213$

$$GM = \text{antilog}\left[\frac{\sum \log X}{N}\right]$$

$$= \text{antilog}\left[\frac{14.1213}{7}\right]$$

$$= \text{antilog}\ (2.0173)$$

$\therefore$ GM = 104.1 Ans.

Calculation of geometric mean in a discrete and continuous series

STEP I : Find the log of the variable X in case of a discrete series. Find the mid-points of the class intervals and take these as variable X, then find the log of these X values in care of continuous series.

STEP II : Multiply the log values with the respective frequencies and add up these to set $\sum(f.\log X)$.

Step III: Divide $\sum f.\log X$ by the sum of the frequencies i.e. the number total of observations (N or $\sum f$)

STEP IV : Take the antilog of the quotient to get GM.

Example :

Find the geometric mean from the following :

Marks :	5	15	25	35	45	55
No. of students :	3	8	15	20	10	4

Solution :

Marks (X)	No. of students (f)	log X	f.logX
5	3	0.6990	2.0970
15	8	1.1761	9.4088
25	15	1.3979	20.9685
35	20	1.5441	30.8820
45	10	1.6532	16.5320
55	4	1.7404	6.9616
	$\sum f = 60$		$\sum \text{flogX} = 86.8499$

$$\text{GM} = \text{antilog}\left(\frac{\sum \text{flogX}}{\text{N}}\right)$$

$$= \text{antilog}\left(\frac{86.8499}{60}\right)$$

$\therefore$ GM = 28.02 Ans.

Example :

Find GM from the following data :

Marks :	0-10	10-20	20-30	30-40	40-50	50-60
No. of students :	3	8	15	20	10	4

Solutions :

Marks	No. of students (f)	Mid points (X)	log X	flog X
0-10	3	5	0.6990	2.0970
10-20	8	15	1.1761	9.4088
20-30	15	25	1.3979	20.9685
30-40	20	35	1.5441	30.8820
40-50	10	45	1.6532	16.5320
50-60	4	55	1.7404	6.9616
	$\sum f = 60$			$\sum \text{flogX} = 86.8499$

$$\text{GM} = \text{antilog}\left(\frac{\sum \text{flogX}}{\sum f}\right)$$

$$= \text{antilog}\left(\frac{86.8499}{60}\right)$$

$$= \text{antilog}\ (1.4475)$$

$$\text{GM} = 28.02 \text{ Ans.}$$

Compound Interest Formula

Geometric mean is especially useful in determining the increase in geometric pattern of variable such as change in population or prices. The compound growth rate is calculated using the formula :

$$P_N = P_0\,(1 + r)^N$$

Where P_N = Value after Nth period

P_0 = Value at the start of Nth period

N = Number of years

r = Rate of growth

This formula translates into the following after solving for rate of growth.

$$r = \left[\text{antilog} \left(\frac{\log P_N \text{ - } \log P_o}{N} \right) \right] - 1$$

Example :

Population of India was 36 crores in 1951 and 121 crores in 2011. Find the annual rate of grwoth of population :

Solution :

The geometric rate of growth is given by

$P_N = P_0 (1 + r)^N$ or $121 = 36 (1 + r)^{60}$

or $$r = \left[\text{antilog} \left(\frac{\log 121 \text{ - } \log 60}{N} \right) \right] - 1$$

$$r = \left[\text{antilog} \left(\frac{2.0828 \text{ - } 1.7782}{60} \right) \right] - 1$$

$$r = \left[\text{antilog} \left(\frac{0.3046}{60} \right) \right] - 1$$

$$r = \text{antilog}(0.0051) - 1$$

$r = 1.012\text{-}1$

$\therefore$ $r = 0.012$

Therefore, annual rate of growth of population r = 0.012 x 100=1.2% ans.

2.4.6 Harmonic Mean (HM)

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of the individual values or observations. It is used in those situations when the work being performed is kept constant and the average rate is to be calculated.

Harmonic mean in an individual series

$$HM = \frac{N}{\Sigma \left(\frac{1}{x} \right)}$$

where N = Number of observations

STEP I : Take reciprocal of the observations of variable x and add up these to get $\Sigma\left(\frac{1}{x}\right)$

STEP II : Divide the number of observation N by $\Sigma\left(\frac{1}{x}\right)$ to get Harmonic Mean

Example :

Find Harmonic Mean of the following :

3, 5, 6, 8

Solution :

X	$\frac{1}{x}$
3	0.3333
5	0.2000
6	0.1667
8	0.1250
N = 4	$\Sigma\left(\frac{1}{x}\right)$ = 0.8250

$$H.M = \frac{N}{\Sigma\left(\frac{1}{x}\right)}$$

$$H.M = \frac{4}{0.8250}$$

H.M = 4.8485

∴ H.M = 4.8485 Ans.

Harmonic mean in a discrete and continuous series

$$H.M = \frac{N}{\sum\left(f.\frac{1}{x}\right)}$$

or $$H.M = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$$

where $N = \sum f$

STEP I : In case of a discrete series, take reciprocal of the observations of variable X and multiply these by their respective frequencies. We can also directly divide frequencies by the observations of variable X, to get $\sum\left(\frac{f}{x}\right)$. In case of a continuous series, take mid-point of the class intervals as X values and find $\sum\left(\frac{f}{x}\right)$.

STEP II : Divide the sum of the frequencies N or $\sum f$ by $\sum\left(\frac{f}{x}\right)$ to get H.M.

Example :

Find H.M. from the following :

No. of rooms :	1	2	3	4
No. of families :	5	4	3	2

Solution :

No. of rooms (X)	No. of families (f)	$\frac{f}{x}$
1	5	5
2	4	2
3	2	1
4	2	0.5
	$\sum f = N = 14$	$\sum\left(\frac{f}{x}\right) = 8.5$

$$H.M = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$$

$$H.M = \frac{14}{8.5}$$

H.M = 1.65

∴ H.M = 1.65 Ans

Example :

Find the Harmonic Mean of the following :

Marks :	0-40	10-20	20-30	30-40	40-50
No. of students :	20	45	125	140	90

Solution:

Marks	No. of students (f)	Mid points (x)	$\frac{f}{x}$
0-10	20	5	4
10-20	45	15	3
20-30	125	25	5
30-40	140	35	4
40-50	90	45	2
	$\sum f = N = 420$		$\sum\left(\frac{f}{x}\right) = 18$

$$H.M = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$$

$$H.M = \frac{14}{18}$$

$\therefore$ H.M = 23.33 Ans

Merits of H.M. :

1. Based upon all observations
2. Ragidly defined
3. Amenable to further algebric treatment
4. Measures average rates relating to time better.

Demerits of H.M.:

1. Difficult to understand
2. Difficult to calculate
3. Difficult to compute in case of zero and negative values.

Relationship between arithmetic, geometric and harmonicmean -

Always, AM > GM > HM

However, it all observations are came then AM = GM = HM

2.4.7 Let us sum UP

From the discussion of the merits and demerits of the various measures of central tendency. It is obvious that no single average is suitable for all practical problems. Each of the averages, has its own merits and demerits and consequently in own field of importance and utility. For example, arithmetic mean is not to be recommend while dealing with frequency distribution with extreme observations. Median and mode are the averages to be used dealing with open end classes. More is particularly used in business and geometric mean is to be used while dealing with rates and ratios. Harmonic mean is to be used for computing special types of average rates or ratios where time factor is variable and the act being performed e.g. distance is constant.

2.4.8 Lesson and Exercise :

Q1. Find the class intervals if the arithmetic mean of the following distribution is 33 and assumed mean 35 :

Step deviation	-3	-2	-1	0	+1	+2
Frequency	5	10	25	30	20	10

Ans : 5

Q2. Following is the distribution of marks in law obtained by 50 students :

Marks :	0	10	20	30	40	50
No. of students :	50	46	40	20	10	3

Calculate the median marks. If 60 percent of the students pass this test, find the minimum marks obtained by a pass candidate.

Ans. 25

Q3. What is the relationship between mean, median and mode ?

Find out the mode of the following data graphically and check the result through calculations :

Size :	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8
Frequency:	3	7	9	15	25	20	14	12
Size :	8-9	9-10	10-11					
Frequency:	8	6	2					

Ans. 4.67 marks

Q4. Find the geometric mean for the following distribution :

Marks :	0-10	10-20	20-30	30-40	40-50
No. of students :	5	7	15	25	8

Ans. 25.64 marks

Q5. If arithmetic mean and geometric mean of two values one 10 and 8 respectively find the values.

Ans. 16, 4

****

M.A. Economics **Semester - IInd** **Lesson - 3**

Course No. 203 **Unit - I**

MEASURES OF DISPERSION

CHAPTER HIGHLIGHTS:

The present lesson inculcates the information regarding the various measures of dispersion.

CHAPTER OBJECTIVES:

3.1 Introduction

3.2 Importance of Dispersion

3.3 Characteristics of a good measures of dispersion

3.4 Types of dispersion

3.4.1 Absolute measures

3.4.2 Relative measures

3.5 Different measures of dispersion

3.5.1 Range and coefficient of range

3.5.2 Inter-quartile range

3.5.3 Quartile deviation and co-efficient of quartile deviation

3.5.4 Mean deviation and co-efficient of mean deviation

3.5.5 Standard deviation and co-efficient of standard deviation

3.5.6 Variance and co-efficient of vriance

3.6 Let us sum up

3.7 Lesson end exercise

3.1 INTRODUCTION

The measures of central tendency give a value that represent the whole set of observations an average. Some of the observations are wear to the average value and some far from it. So the proper understanding of the variability of the observations around the average value it required to have proper knowledge of the characteristics of the statistical data.

The measures of dispersion are those which measure the scatterdness of data or variability of data around an average. These help us to know the important characteristics of a statistical distribution.

3.2 IMPORTANCE OF DISPERSION

1. To know about the structure of a distribution
2. To check the reliability of an average
3. To compare the variability of two or more series.
4. To help in the use of other statistical measures.
5. To help in putting a system of quality control in place.

3.3 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

1. Based upon all observations
2. Simple to understand
3. Simple to calculate
4. Rigidly defined
5. Not affected by extreme observations
6. Amerable to further algebric treatment
7. Not affected by sampling flunctuations.

3.4 TYPES OF DISPERSION

There are two types of dispersion:

a. **Absolute Dispersion :** When the dispersion is expressed in terms of some statistical units in the original data is expressed, then it is known as absolute dispersion. Now if original data is given in units of rupees, say income, then the variation in income in absolute terms will also be expressed or deduced in rupees. Absolute dispersion is not very helpful when we want to compare two or more different series, expressed in different units.

b. **Relative dispersion :** When the dispersion is expressed in terms of an abstract number or a ratio or percentage, it is called relative dispersion. It is always independent of the original units in which original data are given. It is also known 'co-efficient of dispersion'. Suppose, height of persons is given in centimetrers and their weight is given in kilograms and with help of measures of relative dispersion we come to know that variation in height is 15 percent as compared to 9 percent variation in weight, we can have sensible conclusion that variations are more in height than weight. Now, we could not have compared the two set of data, had we used original units of measurement of 'absolute dispersion' to compare the two different series. So, relative dispersion help us having useful comparisons.

3.5 DIFFERENT MEASURES OF DISPERSION

3.5.1 Absolute measures :

1. Range
2. Inter-quartile range
3. Quartile deviation
4. Mean deviation
5. Standard deviation
6. Variance

3.5.2 Relative dispersion :

1. Co-efficient of range
2. Co-efficient of quartiles deviation
3. Co-efficient of mean deviation
4. Co-efficient of standard deviation

5. Co-efficient of variation

3.5.3 Graphic method

a. Lorenz curve

3.5.1.

a. Range : Range is the simples measure of dispersion which is defined as the difference between the extreme observatios of a series. In other words, it is the difference between the highest and the lowest value or observation in a series of statistical data. Symbolically,

Range = Highest value - Lowest value

R = H - L

Co-efficinet of range $= \frac{H - L}{H + L}$

Range in an individual series :

In case of an individual series, range is calculated by subtracting the lowest value from the highest value.

Example :

Comute range and co-efficient of range from the following :

4, 7, 2, 9, 10, 8

Solution :

Range (R) = H - L

R = 10 - 2

$\therefore$ R = 8 Ans

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

$$= \frac{10 - 2}{10 + 2}$$

$$= \frac{8}{12}$$

Coefficient of range = 0.667 Ans.

Range in case of a discrete series :

In case of a discrete series, range is computer by subtracting the lowest value of variable X from its highest value.

Example :

Compute range fom the following :

Wages (in Rs.)	100	200	300	400	500
No. of workers :	5	7	9	8	6

Solution :

R = H - L

R = 500 - 100

R = 400

$\therefore$ Range = Rs. 400 Ans

$$\text{Co-efficient of range} = \frac{H-L}{H+L}$$

$$= \frac{500-100}{500+100}$$

$$= \frac{400}{600}$$

$\therefore$ Co-efficient of range = 0.67 Ans.

Range in case of continuus series

In case of a continuous series, there are two methods to complete range (a) by taking difference between the upper limit of highest class interval and lower limit of lowest class interval, and (b) by taking difference between the mid-poings of the highest and lowest class intervals. Although, both methods give different results, but both are considered corrrect.

Example :

Compute range from the following :

Marks :	10-20	20-30	30-40	40-50
No. of students :	2	4	7	5

Solution :

a) R = H - L

R = 50 - 10

R = 40

$\therefore$ Range = Rs. 40 Ans.

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

$$= \frac{50-10}{50+10}$$

$$= \frac{40}{60}$$

Co-efficient of range : 0.6 Ans.

b) R = H - L

R = 45 - 15

R = 30

$\therefore$ Range = Rs. 30 Ans.

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

$$= \frac{45-15}{45+15}$$

$$= \frac{30}{60}$$

Co-efficient of range : 0.50 Ans.

Merits of range :

1. Easy to understand
2. Easy to compute
3. Saving of time
4. Saving of labour
5. Helps in forecasting

Demerits of range

1. No based upon all items of a series
2. Affected by sampling fluctuations
3. Cannot be computed in case of open-end intervals
4. No amenable to further algebric treatment

Usefullness of range

1. In quality control of products
2. In studying variations in prices of goods.
3. In analysing variations in prices of stock and shares.
4. In examining the weather reports.

3.5.1. b. Inter Quartile range :

Inter-quartile range is the difference between the first quartile and the third quartile. It is a good measure of variations in the data as fifty percent of the observations are covered in its calculations.

Therefore inter-quartile range = $Q_3 - Q_1$

Merits of inter-quartile range

1. Simple to understand
2. Simple to calculate

3. Not affected by extreme observations

Demerits of inter-quartile range

1. Not based upon all observations
2. Affected by sampling fluctuations
3. Ignors composition of the series

3.5.1.c. Quartile deviation (QD)

Also known as the semi-inter quartile range, quartile deviation is that measure of dispersion, which measures how for the two quarties (Q_1 and Q_3) are from the median. In a symmetrical distribution, both the quartiles are at equal distance from the median while vice-versa holds true in case of symmetrical distributions. Symbolically, in symmetrical distributions,

$$Q_3 - M = M - Q_1 = QD.$$

or $M - QD = Q_1$ and $M + QD = Q_3$

While is asymmetrical distribution.

$$Q_3 - M \neq M - Q_1$$

or $M - QD \neq Q_1$ and $M + QD \neq Q_3$

The formula for computing deviation is

$$\text{Quartile Devition (Q.D.)} = \frac{Q_3 - Q_1}{2}$$

$$\text{Also, Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example:

Calculate the inter-quartile range, quartile deviation and co-efficient of quartile deviation of the marks of the students from the following :

Marks :	0-10	10-20	20-30	30-40	40-50
No. of students :	4	7	10	9	6

Solution :

Marks	No. of students (f)	Cumulative frequency (c.f.)
10-20	4	4
10-20	7	11
20-30	10	21
30-40	9	30
40-50	6	36
	$\sum f = 36$	

a. Q_1 = size of the $\frac{N}{4}$ th item

Q_1 = size of the $\frac{36}{4}$ th item

Q_1 = size of the 9th item

Therefore, the Q_1 class is 10-20, i.e. Q_1 lies in the 10-20 class interval

$$Q_1 = L_1 + \left(\frac{\frac{N}{4} - cf}{f} \text{ x i} \right)$$

$$Q_1 = 10 + \left(\frac{9-4}{7} \text{ x } 10 \right)$$

$$Q_1 = 10 + \left(\frac{5 \text{ x } 10}{7} \right)$$

$$Q_1 = 10 + \left(\frac{50}{7}\right)$$

$Q_1 = 10 + 7.14$

$\therefore$ $Q_1 = 17.14$ Ans.

b. $Q_3 = \text{size of the } \dfrac{3N}{4} \text{ th item}$

$Q_3 = \text{size of the } \dfrac{3 \text{ x } 36}{4} \text{ th item}$

Q_3 = size of the 27th item

$\therefore$ Q_3 lies in the 30-40 class interval or the Q_3 class is 30-40.

Now $$Q_3 = L_1 + \left(\frac{\frac{3N}{4} - cf}{f} \text{ x } i\right)$$

$$Q_3 = 30 + \left(\frac{27 - 21}{9} \text{ x } 10\right)$$

$$Q_3 = 30 + \left(\frac{6 \text{ x } 10}{9}\right)$$

$$Q_3 = 30 + \left(\frac{60}{9}\right)$$

$Q_3 = 30 + 6.67$

$\therefore$ $Q_3 = 36.67$ Ans.

Now, inter-quartile range = Q_3 - Q_1

= 36.67 - 17.14

= 19.53 Ans.

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{36.67 - 17.14}{2}$$

$$= \frac{19.53}{2}$$

= 9.765 Ans.

$$\text{Co-efficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{36.67 - 17.14}{36.67 + 17.14}$$

$$= \frac{19.53}{53.81}$$

= 0.363 Ans.

Similarly, we can calculate there is an individual series quartile deviation

Percentile Range : Percentile range is the difference between the 90th percentile and the 10th percentile. It covers 80 percent of the middle values of a series. Symbolically,

Percentile Range = P_{90} - P_{10}, Percentile Deviation $= \frac{P_{90} - P_{10}}{2}$ and

Co-efficient of percentile range $= \frac{P_{90} - P_{10}}{P_{90} + P_{10}}$

Merits of quartile deviation :

1. Easy to understand
2. Easy to calculate
3. Not affected by extreme observations
4. Can be computed in case of open-end intervals

Demerits of quartile deviation

1. Affected by sampling fluctuations
2. Not suitable for algebric treatment
3. Not based upon all observations
4. Does not show scatterdness around an average but distance on a scale only.

3.5.1.d. Mean Deviation (MD)

Also known as average deviation, mean deviation is the arithmetic mean of the deviations taken from a measure of central tendency, ignoring the negative signs and taking all deviations as positive. Although the deviations can be taken from any measure of central tendency, but it is avoided to take deviations from mode which is ill-defined in many series. Though, deviations can be taken from mean and median, yet median is preferred for computing mean deviation because as compared to mean, the sum of deviations taken from median is minimum when we ignore the negative signs and take all signs of deviations as positive. Symbolically,

$$\text{M.D.} = \frac{\Sigma|X-M|}{N} \text{ or } = \frac{\Sigma|X-\overline{X}|}{N} \text{ or } = \frac{\Sigma|X-Z|}{N}$$

where X = observations of a given variable

M = Median

$\overline{X}$ = Arithmetic Mean

Z = Mode

$N = \sum f$ = Sum of frequencies or Number of observation

Also co-efficient of Mean Deviation $= \frac{\text{M.D.M}}{\text{M}}$ if deviations are taken from median.

Co-efficient of M.D. $= \frac{\text{M.D.}\overline{\text{X}}}{\overline{\text{X}}}$ if deviations are taken from arithmetic mean.

Co-efficient of M.D. = $= \frac{\text{M.D.Z}}{\text{Z}}$ if deviations are taken from mode.

Mean deviaton in an individual series

STEP I : Find mean or median or mode from the given series.

STEP II : Take deviations from that measure of central tendency by subtracting mean or median or mode from all observations, ignore the minus signs and add up these to get $\Sigma|X-\overline{X}|$ or $\Sigma|X-M|$ or $\Sigma|X-Z|$

STEP III : Divide that total by the number of observations N to get M.D.

STEP IV : Divide M.D. by that measure of central tendency from which the deviations have been taken, to compute co-efficient of M.D.

Example :

Calculate M.D. and is co-efficien from the following :

4, 7, 6, 9, 12, 10, 14

Solution :

X	X - M	$\|X-M\|$
4	-5	5
6	-3	3
7	-2	2
9	0	0
10	1	1
12	3	3
14	5	5
N = 7		$\Sigma\|X-M\| = 19$

$$M = \text{Size of the } = \frac{N+1}{2} \text{ th item}$$

$$M = \text{Size of the } \left(\frac{7+1}{2}\right) \text{th item}$$

$$M = \text{Size of the } \frac{8}{2} \text{ th item}$$

M = size of the 4th item

$\therefore$ M = 9

Now $\quad M.D._{m} = \frac{\Sigma|X - M|}{N}$

$$= \frac{19}{7}$$

$$M.D._{m} = 2.71 \text{ Ans.}$$

and co-efficient of $M.D_{m} = \frac{MDM}{M}$

$$= \frac{2.71}{9}$$

$$= 0.301 \text{ Ans.}$$

Mean deviation in a discrete and a continuous series.

STEP I: Find mean or median or mode from the given series.

STEP II : Take deviations from that measure of central tendency by subtracting mean or median or mode from the observations of variable X (in case of a discrete series) and from the mid-points of class intervals (in case of a continuous series). Ignore the minus signs of the negative deviations.

STEP III : Multiply these deviations with their respective frequencies and add up these to get M.D= $\Sigma f.|X - \overline{X}|$ or $\Sigma f.|X - M|$ or $\Sigma f.|X - Z|$.

STEP IV : Divide this total by that number of observations $\left(N = \sum f\right)$ or sum of the frequencies to get $\frac{\sum f\left|X - \overline{X}\right|}{\sum f}$ or $\frac{\sum f\left|X - M\right|}{\sum f}$ or $\frac{\sum f\left|X - Z\right|}{\sum f}$

STEP V : Divide M.D. by that measure of central tendency from which the deviations have been taken, to compute co-efficient of M.D.

Where co-efficient of $M.D. = \frac{MD}{\overline{X} \text{ or } M \text{ or } Z}$

Example :

Find M.D. and Co-efficient of M.D. from the following :

Marks :	5	10	15	20	25	30
No. of students :	4	6	10	8	7	3

Solution :

a) Marks	No. of students (f)	Cumulative frequency (cf)	$\left\|X - M\right\|$	$f.\left\|X - M\right\|$
5	4	4	10	40
10	6	10	5	30
15	10	20	0	0
20	8	28	5	40
25	7	35	10	70
30	3	38	15	45
	$N = \sum f = 38$			$\sum f.\left\|X - M\right\| = 225$

M = Size of the $= \left(\frac{N+1}{2}\right)$th item

M = Size of the $\left(\frac{38+1}{2}\right)$th item

$$M = \text{Size of the } \frac{39}{2} \text{th item}$$

$$M = \text{size of the 19.5th item}$$

$$\therefore \; M = 15$$

Now $$M.D._{m} = \frac{\Sigma(f|X-M|)}{\Sigma f}$$

$$= \frac{225}{38}$$

$$M.D._{m} = 5.92 \text{ Ans.}$$

and co-efficient of $$M.D_{m} = \frac{MDM}{M}$$

$$= \frac{5.92}{15}$$

= 0.395 Ans.

b.

Marks	No. of students (f)	c.f.	Mid point (X)	$\|X-M\|$	$f.\|X-M\|$
0-10	2	2	5	12	44
10-20	4	6	15	12	48
20-30	10	16	25	2	20
30-40	7	23	35	8	56
40-50	3	26	45	18	54
	$N = \Sigma f = 26$				$\Sigma f.\|X-M\| = 222$

$$M = \text{Size of the } \frac{N}{2} \text{th item}$$

$$\text{M} = \text{Size of the } \frac{26}{2} \text{ th item}$$

$$\text{M} = \text{size of the 13th item}$$

$\therefore$ Median class = 20-30

Now $$\text{M} = \text{L}_1 + \left(\frac{\frac{\text{N}}{2} - \text{cf}}{\text{f}} \text{x i} \right)$$

$$\text{M} = 20 + \left(\frac{13 - 6}{10} \text{x } 10 \right)$$

$$\text{M} = 20 + 7$$

$\therefore$ M = 27 Ans.

Therefore, $$\text{MD}_\text{m} = \frac{\sum \text{f} |\text{X} - \text{M}|}{\sum \text{f}}$$

$$\text{MD}_\text{m} = \frac{222}{26}$$

$$\text{MD}_\text{m} = 8.54 \text{ Ans.}$$

and co-eff. $$\text{MD}_\text{m} = \frac{\text{MD}_\text{m}}{\text{M}}$$

$$\text{MD}_\text{m} = \frac{8.54}{27}$$

$$\therefore \quad \text{MD}_\text{m} = 0.316 \text{ Ans.}$$

Merits of mean deviation :

1. Simple to understand
2. Easy to calculate
3. Based upon all observation of a series

4. Not influenced by extreme observations
5. Can be used for comparing two or more series as it is based upon a central value.

Demerits of mean deviation :

1. Not amenable to further algebric treatment
2. Mathematically wrong as signs are ignored
3. Not reliable if calculate from mode
4. Requires more labour and time to calculate

3.5.1.e. Standard deviation (SD or σ):

Standard deviation is the under root of the arithmetic mean of the square of the deviations taken from arithmetic mean of the data given. It is considered as the most superior method of measuring dispersion as it does not ignore the negative signs like mean deviation. Also, it used arithmetic mean in computation which is amenable to algebric treatment. The sum of square of deviations taken from mean is also minimum as compared to find from median and mode. Symbolically,

$\text{SD or } \sigma = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{N}}$ in case of an individual series where N = no. of observations and $\overline{X}$ = Arithmetic mean of the series

and $\sigma = \sqrt{\frac{\Sigma f\left(X - \overline{X}\right)^2}{N \text{ or } \Sigma f}}$ in case of a discrete and a continuous series where N=Σf = Sum of observations.

Also co-efficient of SD $= \frac{\sigma}{\overline{X}} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}}$

Standard deviation in an individual series

STEP I : Find mean ($\overline{X}$) of the series.

STEP II: Take deviations from arithmetic mean by subtracting $\overline{X}$ from all observations of variable X.

STEP III : Square the deviations and add up these to get $\Sigma(X - \overline{X})^2$.

STEP IV : Divide this total by the number of observations and take underroot of that to get standard deviation.

STEP V : Divide SD of the series by the arithmetic mean of the series to get the co-efficient of SD.

Example :

Find standard deviation and its co-efficient from the data :

14, 26, 38, 42, 50

Solution :

X	$X - \overline{X}$	$(X - \overline{X})^2$
14	-20	400
26	-8	64
38	4	16
42	8	64
50	16	256
Σx N = 5 $\overline{X}$ = 34		$\Sigma(X - \overline{X})^2$ = 800

$$\overline{X} = \frac{\Sigma x}{N} = \frac{170}{5} = 34$$

Now $\text{SD or } \sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$

$$\text{SD or } \sigma = \sqrt{\frac{800}{5}}$$

$$\text{SD or } \sigma = \sqrt{160}$$

or $\sigma = 12.65$ Ans.

Also co-eff. of $SD = \frac{\sigma}{\overline{X}}$

$$SD = \frac{12.65}{34}$$

SD = 0.370 Ans.

Standard deviation in a discrete and a continuous series :

STEP I : Find the arithmetic mean ($\overline{X}$) of the series.

STEP II : Take deviations from arithmetic mean by subtracting the $\overline{X}$ from all observations of variable X (in case of a discrete series) and by subtracting $\overline{X}$ from mid-points of class interval (in case of a continuous series)

STEP III : Square the deviations

STEP IV : Multiply the squared deviations with the respective frequencies and add up these to get $\sum\left[f\left(X-\overline{X}\right)^2\right]$

STEP V : Divide this total by the number of observation i.e. the sum of the frequencies $\left(N = \sum f\right)$ and take square root or under root of that to get standard deviation.

STEP VI : Divide SD of the series by the arithmetic mean of the series to get the co-efficient of the standard deviation.

Example :

Find the standard deviation and the co-efficient of SD from the following data :

a)	*Marks :*	10	20	30	40	50	60	70
	No of students:	7	9	25	22	18	11	8

Solution :

Marks (X)	No. of Students (f)	fx	$X - \overline{X}$	$(X - \overline{X})^2$	$f(X - \overline{X})^2$
10	7	70	-30	900	6300
20	9	180	-20	400	3600
30	25	750	-10	100	2500
40	22	880	0	0	0
50	18	900	10	100	1800
60	11	660	20	400	4400
70	8	560	30	900	7200
	$\sum f = N = 100$	$\sum fx = 4000$			$\sum f(X - \overline{X})^2 = 25800$

Now $\sigma = \sqrt{\dfrac{\sum(X - \overline{X})^2}{\sum f}}$

$$\sigma = \sqrt{\frac{25800}{100}}$$

$$\sigma = \sqrt{258}$$

$\therefore$ $\sigma =$ 16.06 Ans.

$$\overline{X} = \frac{\sum fx}{\sum f}$$

$$\overline{X} = \frac{4000}{100}$$

$$\overline{X} = 40$$

and co-eff of SD $= \dfrac{\sigma}{\overline{X}}$

$$SD = \frac{16.06}{40}$$

$\therefore$ SD = 0.402 Ans.

co-eff. of variation $= \dfrac{\sigma}{\overline{X}} \times 100$

$$= \frac{22.605}{53} \text{x} \, 100$$

$$= 42.65$$

i.e. 42.65 percent Ans.

Merits of standard deviation :

1. Amenable to algebric deviation
2. Based upon all observations
3. Less affected by sampling fluctuations
4. Can be used to find combined standard deviation
5. Extremly useful in comparing the variability of two or more series.

Demerits of standard deviation :

1. Cannot be computed in open-end intervals
2. Difficult to compute and understand
3. Gives more importance to values near mean and less to value away from mean
4. Does not give any idea about the precision of measurement.

Combined standard deviation :

It is computed by using the formula :

$$\sigma_{12} = \sqrt{\frac{N_1\left(\sigma_1^2 + d_2^2\right) + N_2\left(\sigma_2^2 + d_2^2\right)}{N_1 + N_2}}$$

Where σ_{12} = combined standard deviation

σ_1 = standard deviation of Ist series

σ_1 = standard deviation of 2nd series

$d_1 = \overline{X_1} - \overline{X_{12}}, d_2 = \overline{X_2} - \overline{X_{12}}$

$\overline{X_{12}}$ = combined mean of two series

N_1, N_2 = No. of items in series

b) Marks	No. of Students (f)	Mid-points (X)	fx	$X - \overline{X}$	$(X - \overline{X})^2$	$f(X - \overline{X})^2$
0-20	8	10	80	-43	1849	14792
20-40	12	30	360	-23	529	6348
40-60	30	50	1500	-2	9	270
60-80	20	70	1400	17	289	5780
80-100	10	90	900	37	1369	13690
	$N = \Sigma f = 80$		$\Sigma fx = 4240$			$\Sigma f(X - \overline{X})^2$ =40880

$$\overline{X} = \frac{\Sigma fx}{\Sigma f} = \frac{4240}{80} = 53$$

Now $\sigma = \sqrt{\frac{40880}{80}} = \sqrt{511}$ =22.605 Ans.

and co-eff. of $= \frac{\sigma}{\overline{X}} = \frac{22.605}{53} = 0.427$ Ans.

3.5.1.f. a) Variance

Variance of the given observatios of a series is the square of the standard deviation. Symbolically,

$$\text{Variance}\,\sigma^2 = \frac{\Sigma(X - \overline{X})^2}{N} \quad \text{or} \quad \frac{\Sigma f(X - \overline{X})^2}{\Sigma f}$$

b) Co-efficient of variation :

Developed by Karl Pearson, the co-efficient of variation is that measure of dispersion which gives the degree of variation in a series in percentage terms. Symbolically, Co-eff. of variation $\frac{\sigma}{\overline{X}}$ x 100

Example :

From above example, compute variance and co-efficient of variation

Solution :

Variance $\sigma^2 = (22.605)^2$

$= 510.986$ or 511 Ans.

3.6 LET US SUM UP

The term dispersion is used to indicate the facts that within a given group, the items differ fromone another in size or in other words, there is lack of uniformity in their sizes. Thus the measures of central tendency must be supported and supplemented by some other measures. One such measure in dispersion.

3.7 LESSON AND EXERCISE

Q1. Compute the coefficient of quantile deviation of the following data.

Size	Frequency	Size	Frequency
4-8	6	24-28	12
8-12	10	28-32	10
12-16	18	32-36	6
16-20	30	36-40	2
20-24	15		

Ans. $Q_1 = 14.5$, $Q_3 = 24.92$ coefficient of Q.D. = 0.2643

Q2. Calculat mean deviation from the median for the following data :

Marks less than :	80	70	60	50	40	30	20	10
No. of students :	100	90	80	60	32	20	13	5

Ans. 14.29

Q3. Find the median deviation of the following data :

Size :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	7	12	18	25	16	14	8

Ans. Median = 35.2, M.D., = 13.148

Q4. Calculate the standard deviation from the following data :

Value :	90-99	80-89	70-79	60-69	50-59	40-49	30-39
Frequency :	2	12	22	20	14	4	1

Ans. S.D. = 12.505

Q5. What is meant by dispersion ? Discuss four important measures of spread indicating their uses ?

****

M.A. Economics **Semester - IInd** **Lesson - 4**

Course No. 203 **Unit - I**

CORRELATION

CHAPTER HIGHLIGHTS :

The present lesson contains the information regarding the meaning, uses of correlation along with types, degree and its various methods.

CHAPTER OUTLINES:

4.1 INTRODUCTION:

Correlation is that statistical device which studies the correlation between the two or more variables, when the moment in one is accompained by the moment in another. The degree of association between the two variables is measured through the correlation analysis. Correlation studies the pattern as well as the closeness of the relationship between two or more variables and the tool used for this is called co-efficient of correlation (r).

4.2 USES OR SIGNIFICANCE OF STUDYING CORRELATION:

1. to study the degree of relationship between two or more variables.
2. to estimate the value of one variable on basis of value of another variable (using regression) after establishing existence and closeness of relationship between the varibales.
3. To understand the behaviour of the economic variables.
4. to study the significance of relationship between the variables.

4.3 CORRELATION AND CAUSATION:

Correlation only measures the co-variance between the two or more variables but does not tell anything about the cause and effect relationship between the variables. Existence of some degree of correlation between the variables does not mean that the cause of variation in one variable is due to the variation in another variable. The variation in variables may be due to any of the following reaons :

1. The existence of correlation between the variable may be due to the fact that both the variables mutually influence each other in such a way that it is difficult to designate one as cause and another as result or effect. For example, mutual influencing of demand and price.
2. The two variables showing some degree of correlation may be due to the fact that both are related to same third variable. For example, increasing demand of one commodity may be related to the falling demand of another commodity which is ultimately due to the change in third variable the all or rise in the price of any of these.

3. The correlation between two variables may be due to pure chance, especially in small samples and may be due to sampling fluctuations. For example, the correlation between the height and income of a person.

4. The interdependence of variables may also show correlation. For example, high prices induce to produce more, which may affect cost of production, that may ultimately affect the production and the price level.

4.4 TYPES OF CORRELATION :

a. ***From change in direction point of view*** :-

1. **Positive or Direct :-** When the two variables move or change in same direction, the correlation is known as positive or direct. For example, income and demand.

2. **Negative or Inverse :-** When the variables move or change in opposite direction, the correlation is known as negative or inverse. For example, price and demand.

b. From number of variable under study point of view-

1. **Simple correlation :-** When the relationship between two variables only is studied, it is known as simple correlation. For example relation between income and demand only is studied.

2. **Partial Correlation :-** When the relationship between one dependent and one independent variable is studied while keeping the effect of all other independent variables is kept as constant. For example, as like law of demand where we study the relation between demand and price only while keeping all other factors affecting demand as constant.

3. **Multiple correlation :-** When the relationship between all the variables, dependent and independent, is studied semultaneosuly it, is known as multiple correlation. For example, the relation between output of wheat and all the inputs likeirrigation, seed, pesticide etc. being studied.

c. ***From the proportion change point of view -***

1. **Linear Correlation :-** When the demand or rate of change in variables is same, it is known as linear correlation. For example,

X:	5	10	15	20	25
Y:	100	200	300	400	500

2. **Non-Linear Correlation :-** When the amount or rate of change in the variables is not same, it is known as non-linear correlation. For example,

X:	5	10	15	20	25
Y:	100	130	170	250	300

4.5 DEGREE OF CORRELATION :

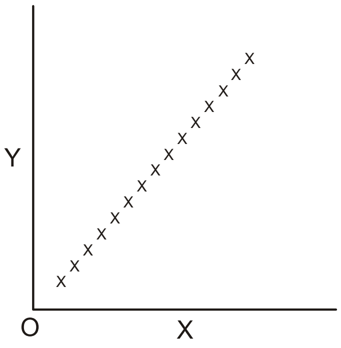
The co-efficient of correlation(r) measures the degree of relation between the variables and always varies between -1 (minus one) and +1 (plus one). i.e. $-1 \leq r \leq +1$ always. The negative value of r shows negative or inverse relation between the variables and the positive value of r shows the positive or direct relation between the variables. If the value of r is zero, then it shows that there is no relation between the variables.

Value of correlation	Degree of correlation
$r = +1$	Perfectly postive
$+0.75 \leq r < +1$	Highly positive
$+0.50 \leq r < +0.75$	Moderately positive
$0 < r < +0.50$	Low positive
$r = 0$	No correlation
$-0.50 < r < 0$	Low negative
$-0.75 < r \leq -0.50$	Moderately negative
$-1 < r \leq -0.75$	Highly negative
$r = -1$	perfectly negative

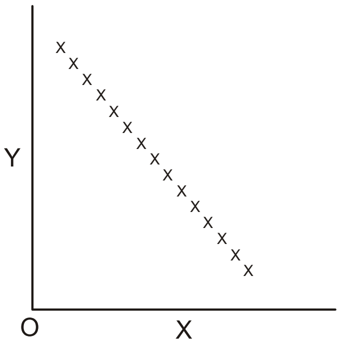
4.6 METHODS OF CORRELATION :

4.6.1 Graphic Methods :-

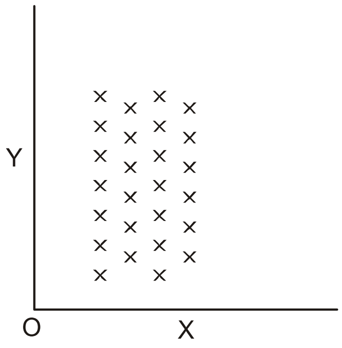
i. **Scatter Diagram Method :-** It is the simplemost method of studying correlation in which the pair of value of two variables (X_1, Y_1), (X_2, Y_2)........(X_n, Y_n) are plotted on a diagram and a rough idea about the relation between the variables is taken. For example.



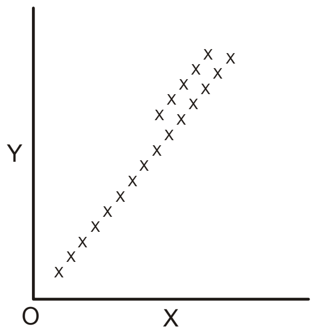
Perfect Positive Correlation



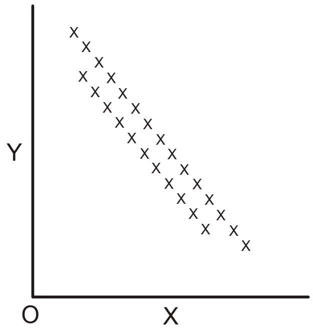
Perfect Negative Correlation



No or Zero correlation

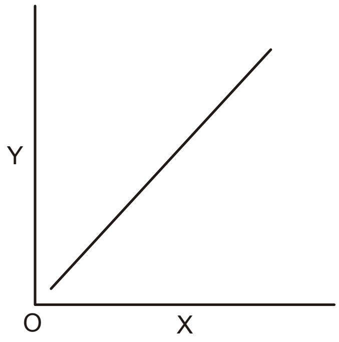


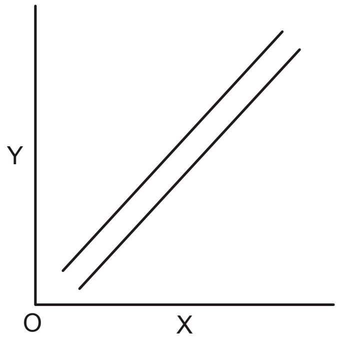
Postitive Correlation

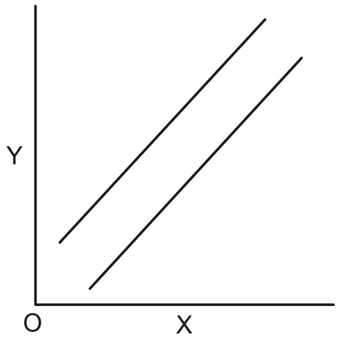


Negative Correlation

ii. **Correlation Graph method :-** In this method we plot the values of two variables on a graph and join these to get two lines representing the given two variables. Now, by examing the closeness and derection of these two lines, we get an idea about the relation between the variables.



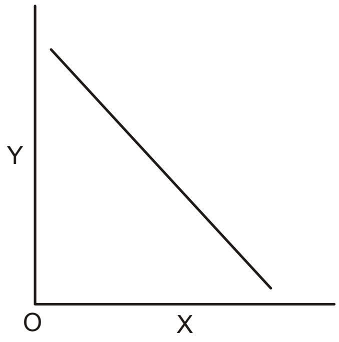


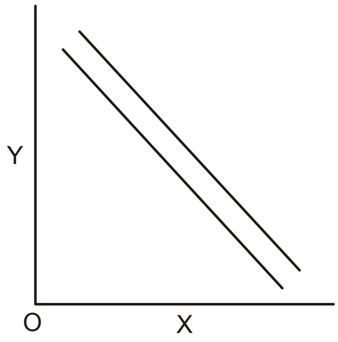


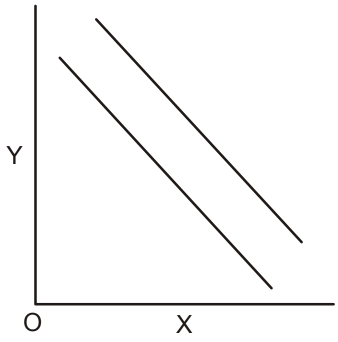
Perfect positive Correlation if lines coincide

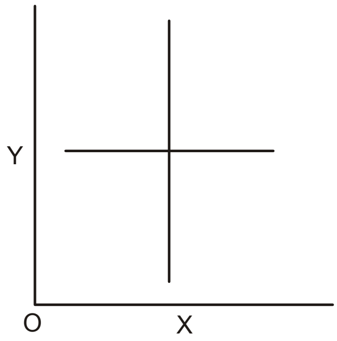
Highly Positive Correlation

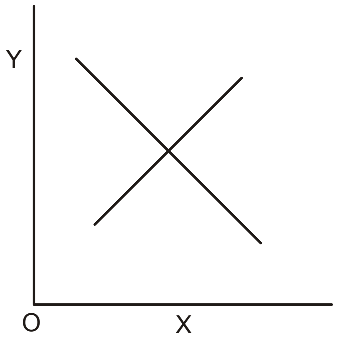
Low Positive Correlation











No or zero correlation

Perfectly Negative Correlation

4.6.2. Mathematical on Algebric Methods -

i. **Karl Pearson Co-efficient of Correlation** - This method was given by a famous statistician Karl Pearson which is mostly used in practice.

a. When deviation from actual mean are taken

$$r = \frac{\text{Covariance of X and Y}}{\left(\sqrt{\text{Variance of X}}\right)\left(\sqrt{\text{Variance of Y}}\right)} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}_X}.\sqrt{\text{Var}_Y}}$$

Where Covariance of X and Y = $\frac{\sum xy}{N}$

Variance of X = σ_x^2

Variance of Y = σ_Y^2

Or

$$r = \frac{\sum xy}{N.\sigma_X.\sigma_Y}$$

Where N = No. of pairs of observations

σ_X = Standard Deviation of X variable = $\sqrt{\frac{\sum X^2}{N}}$

σ_Y = Standard Deviation of Y variable = $\sqrt{\frac{\sum Y^2}{N}}$

$x = X - \overline{X}$

$y = Y - \overline{Y}$

Or

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$ Where $x = X - \overline{X}$, $y = Y - \overline{Y}$

b. When actual values are used direct or the Product Moment method -

$$r = \frac{N.\sum XY - \sum X.\sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}.\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

Where N = No. of pairs of observations

$\sum XY$ = Sum of product of corresponding observations of variables X and **Y**

$\sum X$, $\sum Y$ = Sum of observations of X and Y variables.

$\sum X^2$, $\sum Y^2$ = Sum of squares of observations of the X and Y Variables.

c. When deviations from assumed mean are taken :

$$r = \frac{N.\sum dxdy - \sum dx.\sum dy}{\sqrt{N\sum d^2x - \left(\sum dx\right)^2}.\sqrt{N\sum d^2y - \left(\sum dy\right)^2}}$$

Where N = No. of pairs of observations

$\sum dxdy$ = Sum of product of deviations taken from assumed mean in observations of X and Y variables.

$\sum dx \sum dy$ = Sum of deviation taken from assumed mean in observations of x variable and Y variable.

$\sum d^2x, \sum d^2y$ = Sum of squares of deviation taken from assumed mean in observations of X variables and Y variable.

d. When two values of Actual and Assumed means of X and Y variables and Standard Deviation of X and Y variables are given -

$$r = \frac{\sum dxdy - N\left(\overline{X} - A_X\right)\left(\overline{Y} - A_Y\right)}{N.\sigma_X.\sigma_Y}$$

Where N = No. of pairs of observations

$\overline{X}, \overline{Y}$ = Actual means of X and Y series.

A_X, A_Y = Asumed means of X and Y series.

σ_X, σ_Y = Standard Deviation of X and Y series.

$\sum dxdy$ = Sum of product of deviations taken from assumed means of X and Y series.

Example - Find Karl Pearsonian Co-efficient of Correlation from the following data :-

X:	30	50	70	40	10
Y:	40	30	10	20	50

Solution :

X	Y	$x = X-\overline{X}$	X^2	$Y=Y-\overline{Y}$	Y^2	XY
30	40	-10	100	+10	100	-100
50	30	+10	100	0	0	0
70	10	+30	900	-20	400	-600
40	20	0	0	-10	100	0
10	50	-30	900	+20	400	-600
$\sum x=200$ $\overline{X}=40$	$\sum y=150$ $\overline{Y}=30$		$\sum x^2=2000$		$\sum y^2=1000$	$\sum xy = -1300$

Now $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-1300}{\sqrt{2000 \times 1000}} = -\frac{1300}{1414.214} = -0.919$$

b. Using product moment Method :-

x	y	x^2	y^2	xy
30	40	900	1600	1200
50	30	2500	900	1500
70	10	4900	100	700
40	20	1600	400	800
10	50	100	2500	500
$\sum x= 200$	$\sum y= 150$	$\sum x^2=10,000$	$\sum y^2= 5500$	$\sum xy=4700$

N = 5

$$r = \frac{N\sum xy - \sum x.\sum y}{\sqrt{N\sum x^2-(\sum x)^2}\sqrt{N\sum y^2-(\sum y)^2}}$$

$$= \frac{(5\times 4700)-(200\times 150)}{\sqrt{(5\times 10000)-(200)^2}\sqrt{(5\times 5500)-(150)^2}}$$

$$= \frac{23500-30000}{\sqrt{50000-40000}\sqrt{27500-22500}}$$

$$= \frac{-6500}{\sqrt{10000}\sqrt{5000}}$$

$$= -\frac{6500}{100\times 70.711}$$

$$= -\frac{6500}{7071.1}$$

$$\therefore r = -0.919$$

c.

X	Y	Ax=50 X-A_x=d_x	Ay=40 Y-A_y =d_y	d^2x	d^2y	dxdy
30	40	-20	0	400	0	0
50	30	0	-10	0	100	0
70	10	+20	-30	400	900	-600
40	20	-10	-20	100	400	+200
10	50	-40	+10	1600	100	-400
		$\sum dx=-50$	$\sum dy=-50$	$\sum d^2x=2500$	$\sum d^2y=1500$	$\sum dxdy=-800$

N = 5

$$r = \frac{N\sum dxdy - \sum dx.\sum dy}{\sqrt{N\sum d^2x-(\sum dx)^2}.\sqrt{N\sum d^2y-(\sum dy)^2}}$$

$$=\frac{[5\times(-800)]-(-50).(-50)}{\sqrt{(5\times2500)-(-50)^2}.\sqrt{(5\times1500)-(-50)^2}}$$

$$=\frac{-4000-2500}{\sqrt{12500-2500}.\sqrt{7500-2500}}$$

$$=\frac{-6500}{\sqrt{10000}.\sqrt{5000}}$$

$$=-\frac{6500}{100\times70.711}$$

$$=-\frac{6500}{7071.1}$$

$$\therefore r=-0.919$$

Example :- Find coefficient of correlation from the following Number of pairs of observation of X and Y series = 9

Arithmetic Mean of X series = 74, Assumed Mean of X series = 9

Arithmetic Mean of Y series = 125, Assumed Mean of Y series = 69

Standard Deviation of X series = 13.1, Standard deviation of Y series = 15.9

Sum of product of deviation of X and Y series = 2170

Sol. N = 9, $\overline{X}$=74, A_x=69 , $\sum$ dxdy=2170, $\overline{Y}$=125, A_y=112,

σ_x=13.1, σ_y=15.9

$$r=\frac{\sum dxdy-N(\overline{x}-Ax)(\overline{y}-Ay)}{N.\sigma_x.\sigma_y}$$

$$r=\frac{2170-9(74-69)(125-112)}{9\times13.1\times15.9}$$

$$=\frac{2170-(9\times5\times13)}{1874.61}$$

$$=\frac{2170-585}{1874.61}$$

$$=\frac{1585}{1874.61}$$

$$\therefore r = 0.846$$

- **Assumptions of Karl Pearson's Co-efficient of Correlation-**

i. Variables under study are influnced by a number of variables and form a normal distribution.

ii. The relationship between the variables is linear.

iii. There exist a cause and effect relationship between the variables.

- **Merits of Karl Pearson's Co-efficient of Correlation**

i. Indicates the magnitude of relationship between the variables.

ii. Gives the direction of relationship between the variables.

iii. Practical method of measuring correlations.

iv. Helpful in estimation of the value of dependent variable on the basis of given value of interdependent variable.

- **Limitations of Karl Pearson's Co-efficient of Correlation**

i. Unscientific assumption of linear relationship between the variables.

ii. Unduly affected by the extreme observations.

iii. Cannot be completed in case of qualitative variables like intelligence, beauty etc.

iv. More time consuming and labourious.

v. Does not give any idea about cause and effect relationship between the variables.

- **Properties Karl Pearson's Co-efficient of Correlation**

i. Co-efficient of correlation is independent of the change of origin and change of scale.

ii. Its value change lies between -1 and +1

i.e. $-1 \leq r \leq +1$

iii. It is independent of the units of measurment of the original data or variables.

iv. The degree of relationship between the variables is symmetric i.e. $r_{xy} = r_{yx}$

$$\text{or } \frac{\Sigma xy}{N.\sigma_x.\sigma_y} = \frac{\Sigma yx}{N.\sigma_y.\sigma_x}$$

v. The co-efficient of correlation is the geometric mean of the regression co-efficients.

$$\text{i.e. } r = \pm\sqrt{b_{xy} \times b_{yx}}$$

vi. Co-efficient of correlation will have same sign as that of both regression co-efficient.

2. **Spearman's Rank Correlation** :-

Developed by Charles Edward Spearman and denoted by r_k or ρ (Rho), this method of computing correlation is used when the data given is of qualitative nature and also when the distribution of data given is not normal. Spearman's rank correlation co-efficient possesses the same properties as that of Pearson's correlation co-efficient spearman's rank correlation co-efficient is defined as

$$r_k \text{ or } \rho = 1 - \frac{6\Sigma d^2}{N^3 - N}$$

Where N = No. of pairs of observations

d = R_x-R_y or R_y-R_x = difference between the ranks of the variables X and Y

➢ **Methods of computing Spearman's rank correlation co-efficient -**

A. ***When Ranks are given*** : When the ranks of the variables are given, we just need to find the difference between the ranks of the given variables (d = R_x-R_y), square these differences (d^2) and add up these to get Σd^2. Then use the formula : r_k or $\rho = 1 - \frac{6\Sigma d^2}{N^3 - N}$

Example :- Seven contestants were judged in a beauty contest as below. Find Spearman's rank correlation co-efficient

Ranks by the first judge :	1	3	6	2	7	5	4
Ranks by the second judge :	6	4	1	3	5	2	7

Solution

Rx	Ry	d=Rx-Ry	d^2
1	6	-5	25
3	4	-1	1
6	1	5	25
2	3	-1	1
7	5	2	4
5	2	3	9
4	7	-3	9
N = 7			$\sum d^2 = 74$

Now $\rho = 1 - \frac{6\sum d^2}{N^3 - N} = 1 - \frac{6 \times 74}{7^3 - 7} = 1 - \frac{444}{343 - 7}$

$$= 1 - \frac{444}{336} = 1 - 1.321$$

$\therefore \rho = -0.321$

b. ***When Ranks are not given*** - When the values of variables are given, then rank the variables is either in ascending or descending order. Then proceed as above.

Example :- Seven contestants were judged in a beauty contest and marks were awarded to them. Compute spearman's rank correlation co-efficient.

Marks awarded by the First judge	90	78	55	84	50	60	73
Marks awarded by the Second judge	52	65	80	70	60	78	45

Solution :

(X) Marks by first judge	(Y) Marks by second judge	Rx	Ry	d=Rx-Ry	d^2
90	52	1	6	5	25
78	65	3	4	1	1
55	80	6	1	-5	25
84	70	2	3	1	1
50	60	7	5	-2	4
60	78	5	2	-3	9
73	45	4	7	3	9
N = 7					$\sum d^2 = 74$

Now $\rho = 1 - \frac{6\sum d^2}{N^3 - N} = 1 - \frac{6 \times 74}{7^3 - 7} = 1 - \frac{444}{343-7}$

$$= 1 - \frac{444}{336} = 1 - 1.321$$

$$\therefore \rho = -0.321$$

C. ***When the ranks are repeated*** :- In case the ranks are repeated or when the variables have same values repeated once on twice or more, then these repeated values are asigned the average of ranks that they would have got, had they not been repeated. For example, if the two repeated values are at no.3 and 4, then these would get $\frac{3+4}{2} = \frac{7}{2} = 3.5$ Rank each. Similarly, three repeated values get $\frac{3+4+5}{3} = \frac{12}{3} = 4$ rank each. the adjustment factor $\left(\frac{m^3 - m}{12}\right)$ is added to $\sum d^2$ that number of times which is equal to number of observations getting repeated. M is number of times a of particular value has been repeated.

The formula for this is

$$\rho = 1 - \frac{6\left[\sum d^2 + \left(\frac{m^3 - m}{12}\right) + \left(\frac{m^3 - m}{12}\right) + ...\right]}{N^3 - N}$$

Example :- Seven contestants were judged in a beauty contest and marks were awarded to them. Compute spearman's rank correlation co-efficient.

Marks by First judge	90	78	50	84	50	60	73
Marks by Second judge	52	65	80	65	65	78	45

Solution:

(X) Marks by first judge	(Y) Marks by second judge	Rx	Ry	d=Rx-Ry	d^2
90	52	1	6	-5	25
78	65	3	4	-1	1
60	80	5.5	1	4.5	20.25
84	65	2	4	-2	4
50	65	7	4	3	9
60	78	5.5	2	3.5	12.25
73	45	4	7	-3	9
N = 7					$\sum d^2 = 80.5$

$$\rho = 1 - \frac{6\left[\sum d^2 + \left(\frac{m^3 - m}{12}\right) + \left(\frac{m^3 - m}{12}\right) + ...\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[80.5 + \left(\frac{2^3 - 2}{12}\right) + \left(\frac{3^3 - 3}{12}\right)\right]}{7^3 - 7}$$

$$= 1 - \frac{6\left[80.5 + \frac{6}{12} + \frac{24}{12}\right]}{343 - 7}$$

$$= 1 - \frac{6(80.5 + 0.5 + 2)}{343 - 7} = 1 - \frac{6 \times 83}{336} = 1 - \frac{498}{336} = 1 - 1.482$$

$\therefore\ \rho = -0.482$

- ***Merits of Spearman's rank correlation co-efficient -***

i. Easy to understand

ii. Easy to compute

iii. Best for data of qualitative nature

iv. Useful when ranks, and not original data are given

v. Ignores the effect of extreme observations.

- ***Demerits or Limitation of Spearman's rank correlation co-efficient :-***

i. More time consuming and labourious in case of large no of pairs of observations.

ii. Cannot be used in case of grouped frequency distribution.

3. **Concurrent Deviations Method** :-

This is the simplest method of studying correlation. It is used to find out the direction of change of the variables. Only the direction of change of varibles is taken into account and not the magnitude of the variables. The formula used for this is

$$r_c = \pm\sqrt{\frac{\pm(2c - n)}{n}},$$

+ sign if 2c-n = +ve and - sign if 2c-n = -ve

Where n = N-1 = No. of pairs of observation less once.

c = No. of positive signs obtained after multiplying the dx and dy.

dx, dy = sign of change of observations over previous observtion or direction of change of observations of variables X and Y.

If 2c-x is positive, then plus sign is taken both inside and outside the square root. On the other hand, if 2c-x is negative, then minus sign is taken inside and outside the square root. The co-efficient of concurrent deviation also lies between +1 and -1.

Example : Find the co-efficieint of concurrent derivations from the following.

X:	90	78	55	84	50	60	73
Y:	52	65	80	70	60	78	45

Solution:

X	Y	dx	dy	dx.dy
90	52	x	x	x
78	65	-	+	-
55	80	-	+	-
84	70	+	-	-
50	60	-	-	+
60	78	+	+	+
73	45	+	-	-
				c = 2

x = N-1 = 7 -1 = 6

Now 2c - x = (2 x 2) - 6 = 4 -6 = -ve

$$r_c = \pm\sqrt{\frac{\pm(2c-n)}{n}} = -\sqrt{\frac{-[(2x2)-6]}{6}} = -\sqrt{\frac{-[4-6]}{6}}$$

$$= -\sqrt{\frac{-[-2]}{6}} = -\sqrt{\frac{2}{6}} = -\sqrt{0.333} = -0.577$$

4.7 PROBABLE ERROR, STANDARD ERROR AND CO-EFFICIENT OF CORRELATIONS

- ***Problem Error*** :- The relationship between the variables is often studied on the basis of sample data rather than using all the observations of the given two variables in a bivariate universe sampling fluctuations may lead to the inclusion of unwanted observations and exclusion of desirable observations from the data under analysis, leading to different values of correlation co-efficient from different samples taken from the same bivariate distributions. So, the statictical tool, used to check the reliability of the correlation co-efficient and also to determine the limits within which the different correlation co-efficients calculated from different samples taken from same bivariate distribution are expected to lie, is known as probable Error. Denoted by $P.E_r$, its formula is

$$P.E_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

Where 0.6745 a constant number

r = correlation co-efficient

N= No of pairs of observations.

- **Interpretation of 'r' -**

i. If the value of correlation co-efficient (r) is less than probable error ($P.E_r$), then there is no evidence of correlation between the two variables. Also, the correlation co-efficient is not significant.

ii. If the value of correlation co-efficient (r) is not more than six times the probable error ($P.E_r$), then correlation may exist between the two variables and it is not significant, if it exists.

iii. If the value of correlation co-efficient (r) is more than six times the probable error ($P.E_r$), then correlation certainly exists between the two variables it is significant.

Note :- The sign of the correlation co-efficient (minus sign) shoud be ignored while doing the above excercise.

- ***Determination of Limit*** :-

The limits within which the correlation co-efficient calculated from different samples taken from same bivariate distribution are expected to lie, are determined by adding and subtracting the value of probable error to and from the value of correlation co-efficient. i.e. r $\pm$ P.E$_r$.

The Upper limit is given by r + P. E$_r$ and the lower limit is given by r - P.E$_r$.

- **Conditions or limitations or the use of Probable Error-**

i. The data must be near to a normal distribution curve.

ii. The sample used to calucate the correlation co-efficient must be a random sample.

iii. The sample must be fairly large.

- **Standard Error** - It is co-efficient or the constant number 0.6745 is removed or omitted from the formula of Probable Error, than what we have $\left(\frac{1-r^2}{\sqrt{N}}\right)$ is standard Error of correlation co-efficiant. There fore,

$$S.E_r = \frac{1-r^2}{\sqrt{N}}$$

Example:

Given r = 0.90 and N =10, Calculate probable error and determine limits for population correlation co-efficient.

Solution: $P.E_r = 0.6745.\ \frac{1-r^2}{\sqrt{N}} = \frac{0.6745\left[1-(0.9)^2\right]}{\sqrt{10}}$

$$= \frac{0.6745[1-0.81]}{3.162} = \frac{0.6745\times 0.19}{3.162} = \frac{0.128}{3.162} = 0.040$$

Upper limit = r + P.E$_r$ = 0.90 + 0.040 = 0.940

Lower limit = r - P.E$_r$ = 0.90 - 0.040 = 0.860

Also, 6 x P.E$_r$ = 6 x 0.040 = 0.24

Therefore, as r > 6. P.Er or (0.90 > 0.24), the correlation co-efficient is highly significant.

4.8 CO-EFFICIENT OF DETERMINATION :

The measure, which tells how much variation in one variable (the dependent variable) is the result of the variationo in another variable (the independent variable), is known as Co-efficient of Determination. It shows the percentage of variability of dependent variable on the independent variable. It is denoted by r^2 and is found by squaring the co-efficient of correlation. The value of r = 0.86 and that of r^2 = 0.74 means that 74 percent of variation in dependent variable is explaned by or due to the independent variable. Dividing total variation into explained variation variability due to independent variable and unexplained variation (variability not attributed to independent variable), we can compute the co-efficient of determination as follows :

$$\text{Co-efficient of Determination} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

or Co-efficient of Determination

$$= 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} = 1 - \frac{\Sigma(ui - \hat{ui})^2}{\Sigma(ui - \bar{ui})^2}$$

The value of co-efficient of determination varies between 0 and 1. If it is zero, then none of the variation in dependent variable is due to the independent variable under study. If it is one, then whole of the variation in the dependent variable is attributed to the independent variable.

E.g. 1 Estimator the demand curve and find the predicted demand, if p = ` 4

Quantity	1	3	5	8	11	14
Price	10	9	8	7	6	5

Sol.

Quantity	Price	Qi-$\overline{Q}$	Pi-$\overline{P}$	i i	i2
(Q)	(P)	(ui)	(vi)		
1	10	-6	5	-30	36
3	9	-4	3	-12	16
5	8	-2	1	-2	4
8	7	1	-1	-1	1
11	6	4	-3	-12	16
14	5	7	-5	-35	49
42	45	0	0	-92	122

Where, $\mu i = Qi - \overline{Q}$, $\nu i = \frac{Pi - \overline{P}}{0.5}$

$\mu i = Qi - 7$ .... I , $\nu i = \frac{pi - 7.5}{0.5}$ II

∴ The normal equations are

$$\Sigma \nu i = \propto n + \beta \Sigma \mu i$$

$$\&,\ \Sigma \mu i \nu i = \propto \Sigma \nu i + \beta \Sigma \mu i^2$$

then, $6\propto + 0 = 0$ $\Rightarrow \propto = 0$

and, $-92 = 0 + \beta\ (122)$

$\beta = \frac{-92}{122} = -0.75$

∴ The predicted = n is

$\nu i = \propto + \beta \mu i$

$\nu i = 0 + \beta \mu i$

$\nu i = -0.75 \mu i$ ………….III

Now, $(\frac{p-75}{0.5}) = -0.75[Qi-7]$ ………….. [Putting I & II in III]

$P = -0.375Q + 10.125$ ……………IV

Now, $R^2 = 1 - \frac{\Sigma(\mu i - \hat{\mu} i)^2}{\Sigma(\mu i - \bar{\mu} i)^2}$

At P =4 in = n .... IV we have

Then, $4 - 10.125 = -0.375Q$

$\Rightarrow \quad \frac{-6.125}{0.375} = Q$

$\Rightarrow \quad 16.34 = Q$

The predicted demand at price (`4) is 16.34 units.

Q2 Determine the R^2 from the given information.

P	1	3	5	8	11	14
Q	10	9	8	7	6	5

Sol.

P	Q	μi	υi	μiυi	μi^2	υi^2
1	10	-6	2.5	-15	36	6.25
3	9	-4	1.5	-6	16	2.25
5	8	-2	0.5	-1	4	0.25
8	7	1	-0.5	-0.5	1	0.25
11	6	4	-1.5	-6	16	2.25
14	5	7	-2.5	-17.5	49	6.25
42	45	0	0	-46	122	17.5

And, $\sigma^2 \text{P} = \frac{1}{n}\Sigma(\text{Pi-}\overline{\text{P}})^2$

$= \frac{1}{n} (ui)^2 = \frac{122}{6} = 20.34$

Also, $\sigma^2 q = \frac{1}{n} (qi - q)^2$

$= \frac{1}{n} (ui)^2 = \frac{17.5}{6} = 2.92$

Then, $\sigma^2 p\ \sigma^2 q = 2.34 \text{ x } 2.92$

$= 59.39$

$\Rightarrow (\sigma p\ \ \sigma q)^2 = 59.39$

$\Rightarrow \sigma p\ \ \sigma q = \sqrt{59.39}$

$\Rightarrow \sigma p\ \ \sigma q = 7.71$

$\therefore \gamma p,q = \frac{\text{Cov(p,q)}}{\sigma p.\sigma q}$

$\gamma = \frac{-7.67}{7.71}$

$\gamma = -0.99$

Now, Co-efficient of Determination, $r^2 = (-0.99)^2 = 0.98$

4.9 PARTIAL AND MULTIPLE CORRELATION :

Simple correlation analyses the relation between two variables, where movement in one is accompained by the moment in another variable. One dependent variable and one independent variable. However, in economics, we have often seen that the movement in one variable is not due to the movement of another one, but is affected by movements of some other variables also. It is very rare that we find the care of one-to-one relation between two variable only. More often, we find several variable affecting the movement of one variable at the same time. The change in demand is due to the change in price, income, climate, fashion, habbits etc. Similarly, output of wheat is determined by the irrigation, quality of seed, fertilisers, pesticides, herbirdes etc Therefore, it becomes important to study how one variable is affected by many other variables. The partial and multiple correlation analysis helps in this context.

- ***Partial correlation*** - When the relationship between two variables one dependent and one independent variable, is studied, while keeping the effect of all other independent variable is kept as constant, it is known as partial correlation. For example, in law of demand, we study the effect of price only on demand, while keeping the effect of all other variables affecting demand like income, fashion, climate etc as constant. The correlation co-efficient $r_{12.3}$ represents the partial correlation between variable 1 (dependent variable) and variable 2 (independent variable), while the affect of variable 3 is kept constant. In the suffix 12.3, the variables to the left of the dot or decimal are those whose relationship is being studied and the variable to the right of the dot or decimal are those whose effect is kept constant. In case, we have three variables case, there will be three co-efficient of partial correlation i.e. $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$. Note that the property of symmetry hold in case of partial correlation co-efficient also. Therefore, $r_{12.3} = r_{23.1}$, $r_{13.2} = r_{31.2}$ and $r_{23.1} = r_{32.1}$ as was true in case of sample correlation ($r_{12}=r_{21}$).

- ***Order of the partial correlation co-efficient*** - The simple correlation co-efficient (r_{12} on r_{xy}) is known as the correlation co-efficient of zero order as no varibale is held constant. The partial correlation co-efficient, when one variable is kept constant, i.e $r_{12.3}$ is correlation co-efficient of secured order. When two variables are kept constant, then it is known as the correlation coefficient of second order ($r_{12.34}$). Similarly, we can have partial correlation co-efficients of third, fourth....upto nth order, depending upon the number of independent variables kept constant.

- ***Partial correlation co-efficient (Two independent variable)***- If we have X_1, X_2 and X_3 variables, then the partial correlation co-efficient between X_1 and X_2, keeping X_3 as constant, is given by

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1-r_{13}^2}.\sqrt{1-r_{23}^2}}$$

Similarly, $$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1-r_{12}^2}.\sqrt{1-r_{32}^2}}$$

and $$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1-r_{21}^2}.\sqrt{1-r_{31}^2}}$$

- ***Partial correlation co-efficient (Three independent variables)***- In case of one dependent (X_1) and three independent variables (X_2, X_3 and X_4), the partial correlation co-efficient will be

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}.r_{24.3}}{\sqrt{1-r_{14.3}^2}.\sqrt{1-r_{24.3}^2}}$$

Alternative formula giving same result is

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}.r_{23.4}}{\sqrt{1-r_{13.4}^2}.\sqrt{1-r_{23.4}^2}}$$ and so on.

Example:

Given $r_{12} = 0.69$, r_{13}=0.45 and $r_{23} = 0.58$, determine the partial correlation co-efficients.

Solution :

a) $$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1-r_{13}^2}.\sqrt{1-r_{23}^2}}$$

putting the given value, we have

$$r_{12.3} = \frac{0.69-(0.45\times 0.58)}{\sqrt{1-(0.45)^2}.\sqrt{1-(0.58)^2}}$$

$$= \frac{0.69-0.261}{\sqrt{1-0.2025}.\sqrt{1-0.3364}} = \frac{0.429}{\sqrt{0.7975}.\sqrt{0.6636}}$$

$$= \frac{0.429}{0.893\times 0.815} = \frac{0.429}{0.728}$$

$$\therefore r_{12.3} = 0.589$$

b. $$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1-r_{12}^2}.\sqrt{1-r_{32}^2}}$$

$$= \frac{0.45-(0.69\times 0.58)}{\sqrt{1-(0.69)^2}\sqrt{1-(0.58)^2}}$$

$$= \frac{0.45-0.4002}{\sqrt{1-0.4761}\sqrt{1-0.3364}} = \frac{0.0498}{\sqrt{0.5239}\sqrt{0.6636}}$$

$$= \frac{0.0498}{0.7238\times 0.8146} = \frac{0.0498}{0.5896}$$

$$\therefore r_{13.2} = 0.084$$

c. $$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1-r_{21}^2}.\sqrt{1-r_{31}^2}}$$

$$= \frac{0.58-(0.69\times 0.45)}{\sqrt{1-(0.69)^2}\sqrt{1-(0.45)^2}}$$

$$= \frac{0.58-0.3105}{\sqrt{1-0.4761}\sqrt{1-0.2025}} = \frac{0.2695}{\sqrt{0.5239}.\sqrt{7975}}$$

$$= \frac{0.2695}{0.7238 \times 0.8930} = \frac{0.2695}{0.6464}$$

$\therefore r_{23.1} = 0.417$

- ***Properties of Partial Correlation Co-efficient*** :-

1. It is a pure number which is independent of the units of measurement of original data.
2. It is independent of change of origin and change of scale.
3. Its value also lies between -1 and +1, i.e. $-1 \leq r_{xy.z} \leq +1$.
4. It is the geometric mean of the partial regression co-efficient .
5. The square of the partial correlation co-efficient gives as co-efficient of partial determination whose value lies between zero and one.
6. The demand of relationship between the variables is symmetric. ie. $r_{12.3} = r_{21.3}$, $r_{13.2} = r_{31.2}$ and $r_{23.1}=r_{32.1}$

- ***Merits of Partial correlation co-efficient*** :-

1. Useful in analysis of inter-related series.
2. Expressed in a few well defined co-efficient
3. Useful in testing reliability of small amount of data.

- ***Limitations of partial correlation co-efficieint*** :-

1. All zero order co-efficient must have linear regression.
2. The effects of independent variables must be additively and not jointly related.
3. The computation of partial correlation co-efficients beyond first order is labourious and time consuming.
4. It involves difficult interpretations of result.

- **Multiple Correlation** :-

When the relationship between all the variables, one dependent and all independent variables, is studied simultaneously without keeping any variable constant, it is known as multiple correlation. One variable is made dependent, as per the choice of statistician and the requirement of the study, and the effect of all other independent variables on that dependent variables is studied simultaneously. So, the tool that helps in studying or measuring the degree of relationship between all the variables simultaneously is known as multiple correlation. For example, the case of studying the effect of irrigation, seeds, fertilizers, pesticides, climate etc. together on output of wheat relates to multiple correlation analysis.

- ***Co-efficient of Multiple Correlation*** :-

The multiple correlation co-efficieint, denoted by R, helps in studying the degree of relation between the dependent variable (X_1) and all the independent variables (X_2, X_3, X_4....X_n) on it simultaneously. $R_{1.23}$ denotes the multiple correlation co-efficieint, with suffix, 1.23 depicting the multiple correlation between the variables. The subscript to the left of the decimal or dot always represent the dependent variable and those to the right of to dot is decimal represent the independent variables.

The formula for computation is

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}}$$

Similarly, $$r_{2.31} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{31}^2}}$$

and $$r_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{12}^2}}$$

The alternative methods for these are

$$r_{1.23} = \sqrt{r_{12}^2 + r_{13.2}^2\left(1 - r_{12}^2\right)}$$ and so on.

Example:

Given r_{12} = 0.69, r_{23}=0.58 and r_{13}=0.45, determine the co-efficient of multiple correlation

Solution :

a) $$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}}$$

putting the given values, we have

$$r_{1.23} = \sqrt{\frac{(0.69)^2 + (0.45)^2 - 2(0.69)(0.58)(0.45)}{1-(0.58)^2}}$$

$$= \sqrt{\frac{0.4761 + 0.2025 - 0.3602}{1 - 0.3364}}$$

$$= \sqrt{\frac{0.3184}{0.6636}} = \sqrt{0.4798}$$

$$\therefore r_{1.23} = 0.693$$

b. $$r_{2.31} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{31}^2}}$$

$$= \sqrt{\frac{(0.58)^2 + (0.69)^2 - 2(0.69)(0.58)(0.45)}{1-(0.45)^2}}$$

$$= \sqrt{\frac{0.3364 + 0.4761 - 0.3602}{1 - 0.2025}}$$

$$= \sqrt{\frac{0.4523}{0.7975}} = \sqrt{0.5671}$$

$$\therefore r_{2.31} = 0.753$$

c. $$r_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.45)^2 + (0.58)^2 - 2(0.69)(0.58)(0.45)}{1-(0.69)^2}}$$

$$= \sqrt{\frac{0.2025 + 0.3364 - 0.3602}{1 - 0.4761}}$$

$$= \sqrt{\frac{0.1787}{0.5239}} = \sqrt{0.3411}$$

$$\therefore r_{3.12} = 0.584$$

- ***Properties of Multiple correlation co-efficient*** -

1. Multiple correlation co-efficient is a pure number, which is independent of units of measurement of original data.
2. The value of multiple correlation always lies between zero and one, i.e. $0 \leq R_{x.yz} \leq +1$
3. Multiple correlation co-efficient can never be less than the simple and partial correlation co-efficients.
4. If value of $R_{1.23}$ is zero, then r_{12} and r_{13} must be zero.
5. The square of multiple correlation co-efficient is the co-efficient of multiple determination.

- ***Merits of Multiple correlation co-efficient*** -

1. Helps in studying assocation between all variables simultaneously, without having to ignore or to keep any variable constant.
2. Series as a measure of goodness of fit of the regression line.
3. Expresses the type and degree of relation in a few concise coefficieint.

- ***Limitations of Multiple correlation co-efficient*** -

1. Assumption of linear relationship between variables limits applicability of this.
2. Quite labourious and difficult as well as time consuming.
3. Assumption of separate, distinct and addetive effect of independent variables on dependent variable is wrong.

4.10 LETS US SUM UP:

Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spreed and suggest to him the paths through which stabilising forces may become effective. Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

4.11 LESSON END EXERCISE:

Q1. Define correlation. Explain various types of correlation with suitable examples.

Q2. From the following table calculate the coefficieint of correlation by Karl Pearson method.

X	:	6	2	10	4	8
Y	:	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively. (Ans 5.)

Q3. Calculate correlation co-efficient r (x, y) from the following data :

n = 10, $\sum$x=140, $\sum$y=150, $\sum(x-10)^2=180$, $\sum(y-15)^2=215$, $\sum(x-10)(y-15)=60$ (Ans. 0.91)

Q4. Given r = 0.5, $\sum xy = 60$, $\sigma y=2.5$ and $\sum x^2=90$

Find the number of items. Here x and y are deviations from respective means. (Ans. n = 10)

Q5. What is spearman's rank correlation coefficient ? Discuss its usefulness ?

****

M.A. Economics **Semester - IInd** **Lesson - 5**
Course No. 203 **Unit - I**

REGRESSION

CHAPTER HIGHLIGHTS :-

This lesson inculcates the ideas about the regression and various methods for computing regression lines.

CHAPTER OUTLINES :

5.1 Introduction

5.2 Type of Regression

5.3 Difference between correlation & Regression

5.4 Regression lines

5.5 Nature of Regression lines

5.6 Methods of computing Regression lines

- 5.6.1 Scatter diagram method
- 5.6.2 Least square method
- 5.6.3 Regression equations through normal equations
- 5.6.4 Alternative method
- 5.6.5 Regression equation through actual mean.
- 5.6.6 Regression equation through actual mean standard deviation
- 5.6.7 Regression equation through assumed given mean
- 5.6.8 Correlation coefficient from the given regression equation

5.1 INTRODUCTION :

The term 'Regression' originally meaning 'to regress' on 'the tendency to go back' was first used by Sir Francis Galton in 1877 in his study of relationship between heights of fathers and sons. His study revealed an unteresting relationship that tall fathers tend to have tall sons and short fathers tend to have short sons but the average heights of tall sons was less than average hight of tall fathers and that of short sons was more than that of short fathers. Thus, there was tendency in the average height of tall 2 short sons towards the average height of the general population. The line, describing this relationship, was called regression live.

In the modern statistical world, after establishing the existence correlation between the dependent and the independent variable, the statistician may be interested in estimating the most probable value of the dependent variable on the basis of the given value of the independent variable. The statistical device used for this purpose is known as regression.

5.2 TYPES OF REGRESSION :-

i. ***Simple and Multiple Regression*** :- When one independent variable is used to estimate the most likely value of the dependent variable, it is known as simple regression analysis. On the other hand, when two or more independent variables are used at the sometime to estimate the most profable value of the dependent variable, it is known as multiple regression analysis.

ii. ***Linear and Non-Linear Regression*** :- When the value of the dependent variable change by a constant amount due to a unit change in the independent variable and the plotting of these on a graph give a straight line, it is known as linear regression. On the other hand, if the value of the dependent variable do not change by a constant amount due to a unit change in the independent variable and the plotting of these points on a graph do not give a straight line but a curve, then it is known as non-linear regression.

5.3 DIFFERENCE BETWEEN CORRELATION AND REGRESSION :-

Correlation

1. Correlation studies the covariability between two variables.
2. Correlation analysis the degree of relation between the variables.
3. Correlation studies the covariation between the variables without any further objective.
4. Existence of correlation need not imply cause and effect relationship between the variables under study.
5. Correlation co-efficients have the property of symmetry, i.e. $r_{xy} = r_{yx}$.
6. correlation co-efficient is a pure number, independent of the units of measurement of the original data.
7. Correlation co-efficieint can never be greater then unity.
8. Correlation co-efficieint is independent of change of origin and change of scale.
9. Correlation can be spuriour or non -sense due to chance factor.
10. Correlation analysis is applied in those cases where the direction of dependency is not clear between the variables.

Regression

1. Regression studies dependence of one variable on the other variable.
2. Regression analysis the nature of relation between the variables.
3. Regression studies the dependence among the variables with a further objective to estimate the value of dependent variable on basis of given value of independent variable.

4. Regression clearly indicates the cause and effect relationship between the variables under study.

5. Regression co-efficieints of not have the property of symmety, i.e. $b_{xy} \neq b_{yx}$, in gereral.

6. Regression co-efficieint always have units of measurement of the original data.

7. Regression co-efficient can be greater than unity.

8. Regression co-efficieint is independent of change of origin, but not change of scale.

9. Regression can never be spurior or non-sense

10. Regression analysis is applied in those cases where the direction of dependency is clear between the variables.

5.4. REGRESSION LINES :

The line, which describe the functional relationship between the dependent and the independent variables, is called the regression line. There will be two regression lines in case of two variables being studied one dependent variable and one independent variable in a bi-variate distribution.

The two regression lines are

i. Y on X, i.e. Y = a +bX where Y is the dependent and X is the independent variable.

ii. X on Y, i.e. X = a + bY where X is the dependent and Y is the independent variable.

Where 'a' represent the intercept and 'b' represent the slope of line. Since, there is no third variable in a bi-variate distribution, hence there cannot be more than two regression lines.

5.5 NATURE OF REGRESSION LINES :-

The nature of the regression lines depends upon the nature and extent of correlation between the variables being studied.

i. If correlation is positive, then the regression lines will be sloping upward from left to right. Low positive correlation, further the regression lines. High positive correlation, closer the regression lines. The regression lines will coincide in case of perfect positive correlation.

ii. If correlation is negative, then the regression lines will be sloping downward from left to right. Low negative correlation, farther the regression lines. High negative correlation, closer the regression lines. The regression lines will coincide in case of perfect negative correlation.

iii. If correlation is zero, the regression lines will intersect each other at right angle at the 'mean values' of both X and Y.

5.6 METHODS OF COMPUTING REGRESSION LINES :-

5.6.1 Scatter Diagram Method :-

This is the simplest method of constructing the regression lines. In this, the given values of two variables are plotted on a graph in form of dots. Then, a free hand curve or line is drawn going through the centre of the scatter points. The line to obtained is the regression line, depicting the behaviour of the variables. This method is not used very often in practice as it requires a high level of accuracy to predict the behaviour of the variables.

5.6.2 The Least Squares Method or the Algebric Method :

This is the best method to construct the regression lines. In this method, the regression line is fitted to the different points in such a manner that the sum of the squared deviations of the observed values from the fitted line is minimum than from any other straight line. That is why this method is called the 'least squares Method'. Also, the sum of the positive deviations is equal to the sum of the negative deviations. Therefore, the total of the positive and negative deviations is zero. That is why the line fitted through this method is also called the line of best fit. Following are various methods to construct the regression lines.

5.6.3 Regression Equations through normal equations (When actual values are used directly) :

The two regression lines or equations are ‘Y on X’ or Y = a + bX and ‘X on Y’ or X = a +bY. The regression line Y = a +bX is used to estimate the most likely value of dependent variable Y, given the value of independent variable X. Similarly, the regression line X = a +bY is used to estimate the most probable value of dependent variable X, given the value of independent variable Y. ‘a’ and ‘b’ are the constants or numerical constants whose value do not change. ‘a’ is the intercept, i.e., that value at which the regression line meets the Y-axis. It is also known as ‘Y-intercept’. ‘b’ is the slope of the regression line, showing change in the dependent variable for a unit change in the independent variable. The regression equation or line ‘Y on X’ or $Y = a + b_{yx}X$ is obtained using the two normal equations.

$$\sum Y = Na + b\sum X$$

$$\sum Xy = a\sum X + b\ \sum X^2$$

Similarly, the normal equation for the regression line

‘X on Y’ or $X = a + b_{xy}Y$ are

$$\sum X = Na + b\sum y$$

$$\sum XY = a\sum y + b\sum Y^2$$

By solving the normal equations simltaneously, the value of the constants ‘a’ and ‘b’ are obtained. By putting their value in Y = a +bX and in X = a +bY, the regression equations or lines are obtained through the ‘least Square method’. Note thtat ‘b’ or ‘b_{xy}’ is the regression co-efficieint of the equation ‘X on Y’ and ‘b’ or ‘b_{yx}’ is the regression co-efficieint of the equation ‘Y on X’.

Expample : Obtain the regression equation from the following :

X	:	1	2	3	4	5
Y	:	3	6	4	7	2

Solution :

	X	Y	X^2	Y^2	XY
	1	3	1	9	3
	2	6	4	36	12
	3	4	9	16	12
	4	7	16	49	28
	5	2	25	4	10
N = 5	$\sum X=15$	$\sum Y=22$	$\sum X^2=55$	$\sum Y^2=114$	$\sum XY = 65$

The Regression equation of Y on X is

$Y = a + bX$ 1

The two normal equations are

$\sum Y = Na + b \sum X$

$\sum XY = a \sum X + b \sum X^2$

Putting values in these, we get

$22 = 5a + 15b$ 2

$65 = 15a + 55b$ 3

Solving 2 and 3 simultaneously, we get the following by multiplying 2 by 3

$66 = 15a + 45b$ 4

Subtracting 4 from 3, we get

$-1 = 10b$

or $b = -0.1$

Putting the value of b = -0.1 in 2, we get

$22 = 5a + 15(-0.1)$

or $22 = 5a - 1.5$

or $5\ a = -23.5$

or $a = 4.7$

Putting the values of a and b in 1, we get

$Y = 4.7 - 0.1$ x Ans.

Simiarly, the regression equation of X on Y is

$X = a + by$ 5

The two normal equations are

$\sum X = Na + b \sum Y$

$\sum XY = a \sum Y + b \sum Y^2$.

Putting value in these, we get

$15 = 5a + 22b$ 6

$65 = 22a + 114b$ 7

Multiply 6 by 22 and 7 by 5, we get

$330 = 110\ a + 484\ b$ 8

$325 = 110\ a + 570\ b$ 9

Subtrating 8 from 9, we get

$$-5 = 86\ b \text{ or } b = -\frac{5}{86} = -0.06$$

Putting value of b in 6, we get

$15 = 5a + 22\ (-0.06)$

or $15 = 5a - 1.32$

or $5\ a = 16.32$

or $a = 3.26$

Putting value of a and b in 5, we get

$X = 3.26 - 0.06\ Y$

5.6.4. Alternative Method :

The value of constants 'a' and 'b' can also be found alternatively using the following formula.

i. Regression equation 'Y on X' or Y = a +b_{YX}X

$$a = \overline{Y} - b\overline{X}, \quad b_{YX} = \frac{N\sum XY - \sum X.\sum Y}{N\sum X^2 - (\sum X)^2}$$

ii. Regression equation 'X on Y' or X = a + b_{XY}Y.

$$a = \overline{X} - b\overline{Y}, \quad b_{XY} = \frac{N\sum XY - \sum X.\sum Y}{N\sum Y^2 - (\sum Y)^2}$$

Now, using the above data, for regression equation 'Y on X' or Y = a +b_{YX}X,

$$b_{YX} = \frac{(5\times 65)-(15\times 22)}{(5\times 55)-(25)^2} = \frac{325-330}{275-225}$$

$$= \frac{-5}{50} = -0.1$$

a = 4.4 - (-0.1 x 3) = 4.4 + 0.3 = 4.7

∴ The regression equation of Y on X is

Y = 4.7 - 0.1 X Ans.

Also, using the same data for regression equation 'X on Y' or X = a + b_{XY}Y

$$b_{XY} = \frac{(5\times 65)-(15\times 22)}{(5\times 114)-(22)^2} = \frac{325-330}{570-484} = \frac{-5}{86} = -0.06$$

and a = 3 - (-0.06 x 4.4) = 3 + 0.26 = 3.26

Hence, the regression equation of X on Y is

X = 3.26 - 0.06 Y Ans.

5.6.5 Regression Equations through Actual Mean :-

In this method, the deviations from actual mean and taken firstly, and then the regression lines are constructed.

i. The regression equation of Y on X is Y = a + b_{YX}X which is obtained by using the equation $\left(Y-\overline{Y}\right)=b_{YX}\left(X-\overline{X}\right)$

Where $b_{YX}=\dfrac{\sum XY}{\sum X^2}$,

Also, x and y are deviations taken from actual means of variable X and Y, ie. x = $X-\overline{X}$ and y = $Y-\overline{Y}$.

ii. Similarly, the regression equation X on Y is X = a +b_{XY}Y, which is obtained by using the equation

$$\left(X-\overline{X}\right)=b_{XY}\left(Y-\overline{Y}\right)$$

Where $b_{XY}=\dfrac{\sum xy}{\sum y^2}$

Also, x and y are deviations taken from actual means of variables X and Y, i.e. $x=X-\overline{X}$ and $y=Y-\overline{Y}$

Example :

Find the two regression equation from the following data :

X	:	1	2	3	4	5
Y	:	10	6	2	8	4

Solution :

	X	Y	X=X-3	Y=Y-6	X^2	Y^2	XY
	1	10	-2	+4	4	16	-8
	2	6	-1	0	1	0	0
	3	2	0	-4	0	16	0
	4	8	+1	+2	1	4	2
	5	4	+2	-2	4	4	-4
N=5	$\sum X=15$ $\overline{X}=3$	$\sum Y=30$ $\overline{Y}=6$			$\sum X^2$ =10	$\sum Y^2$ =40	$\sum XY$ =-10

i. The regression equation of Y on X is given by

$$\left(Y-\overline{Y}\right)=b_{YX}\left(X-\overline{X}\right) \text{ ....1}$$

Where $b_{YX}=\frac{\sum xy}{\sum x^2}=\frac{-10}{10}=-1$

Putting the values of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation 1, we get

(Y - 6) = (-1) (X-3)

or Y - 6 = -X + 3

or Y = 9 - X Ans.

ii. The regression equation of X on Y is given by

$$\left(X-\overline{X}\right)=b_{XY}\left(Y-\overline{Y}\right)\text{.....................2}$$

Where $b_{XY}=\frac{\sum xy}{\sum y^2}=\frac{-10}{40}=-0.25$

Putting the value of $\overline{X}$, $\overline{Y}$ and b_{XY} in equation 2, we get

(X-3) = (-0.25) (Y-6)

or X -3 = -0.25 Y + 1.5

or X = 4.5 - 0.25 Y Ans.

5.6.6 Regression Equations through actual means standard deviations and correlation co-efficient of given variables\

When the actual means and the standard deviation of the each variables is given alongwith the correlation co-efficient between the two variables, then regression equation are obtained as follows :

i. The regression equation of 'Y on X' is

$$\left(Y-\overline{Y}\right)=b_{YX}\left(X-\overline{X}\right)$$

where, $b_{YX}=r.\frac{\sigma y}{\sigma x}$

ii. The regression equation of 'X on Y' is

$$\left(X - \overline{X}\right) = b_{XY}\left(Y - \overline{Y}\right)$$

Where $b_{XY} = r.\frac{\sigma x}{\sigma y}$

Example :

Obtain two regression equations from the following data.

	X	Y
Actual Mean	25	30
Standard Deviation	5	7

Correlation Co-efficient = 0.85

Solution :

The regression equation of Y on X is given by

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right) \text{ ....1}$$

where, $b_{YX} = r.\frac{\sigma x}{\sigma y} = (0.85)\left(\frac{7}{5}\right) = 1.19$

Putting the values of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation 1,

we get (Y-30) = (1.19) (X-25)

or Y - 30 = 1.19 X - 29.75

or Y = 0.25 + 1.19X Ans.

The regression equation of X on Y is given by

$$\left(X - \overline{X}\right) = b_{XY}\left(Y - \overline{Y}\right) \text{ ....2}$$

Where $b_{XY} = r.\frac{\sigma x}{\sigma y} = (0.85)\left(\frac{5}{7}\right) = 0.61$

Putting the values of $\overline{X}$, $\overline{Y}$ and b_{XY} in equation 2, we get $(X - 25) = (0.61)(Y-30)$

or $X- 25 = 0.61Y - 18.3$

or $X = 6.7 + 0.61Y$ Ans.

5.6.7 Regression Equation through Assumed Mean :-

In this method, the deviations from assumed mean are taken firstly and then the regression equations are obtained as follows :

i. The arithmatic mean of Y-series is given by

$$\overline{Y} = A_Y + \frac{\sum dy}{N}, \overline{X} = A_X + \frac{\sum dx}{N}$$

The regression equation of 'Y on X' is given by

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right)$$

where, $b_{YX} = \dfrac{N\sum dxdy - \sum dx.\sum dy}{N\sum d^2x - (\sum dx)^2}$

ii. The arithmatic mean of X-series is given by

$$\overline{X} = A_X + \frac{\sum dx}{N}, \overline{Y} = A_Y + \frac{\sum dy}{N}$$

The regression line or equation of 'X on Y' is

given by $\left(X - \overline{X}\right) = b_{XY}\left(Y - \overline{Y}\right)$

Where $b_{XY} = \dfrac{N\sum dxdy - \sum dx.\sum dy}{N\sum d^2y - (\sum dy)^2}$

Example:

Obtain the two regression equations from the following data :

X	:	56	42	36	47	49	42	60	72	63	55
Y	:	147	125	118	128	145	140	155	160	149	150

Solution:

X	Y	Ax=55 dx=x-55	Ay=140 dy=y-140	d^2X	d^2y	dxdy
56	147	1	7	1	49	7
42	125	-13	-15	169	225	195
36	118	-19	-22	361	484	418
47	128	-8	-12	64	144	96
49	145	-6	5	36	25	-30
42	140	-13	0	169	0	0
60	155	5	15	25	225	75
72	160	17	20	289	400	340
63	149	8	9	64	81	72
55	150	0	10	0	100	0
N=10		$\sum$ dx=-28	$\sum$ dy=17	$\sum d^2x$ =1178	$\sum d^2y$ =1733	$\sum$ dxdy =1173

Now $\overline{X} = A_X + \frac{\sum dx}{N} = 55 + \left(\frac{-28}{10}\right) = 55 - 2.8 = 52.2$

$$\overline{Y} = A_Y + \frac{\sum dy}{N} = 140 + \left(\frac{17}{10}\right) = 140 + 1.7 = 141.7$$

$$b_{YX} = \frac{N\sum dxdy - \sum dx.\sum dy}{N\sum d^2x - \left(\sum dx\right)^2}$$

$$b_{YX} = \frac{(10\times1173)-(-28)(17)}{(10\times1178)-(-28)^2} = \frac{11730+476}{11780-784} = \frac{12206}{10996}$$

$$\therefore b_{YX} = 1.11$$

Also $b_{XY} = \frac{N\sum dxdy - \sum dx.\sum dy}{N\sum d^2y - (\sum dy)^2}$

$$= \frac{(10\times 1173)-(-28)(17)}{(10\times 1733)-(17)^2} = \frac{11730+476}{17330-289} = \frac{12206}{17041}$$

$\therefore b_{XY} = 0.72$

Now putting the value of $\overline{X}$, $\overline{Y}$, b_{YX} and b_{XY} to get the regression equations, we have

i. The regression equation of Y on X is

$(Y - \overline{Y}) = b_{YX} (X-\overline{X})$

or. $(Y - 141.7) = (1.11) (X-52.2)$

or $Y - 141.7 = 1.11 X - 57.94$

or $Y = 83.76 + 1.11X$ Ans.

ii. The regression equation of X on Y is

$(X-\overline{X}) = b_{XY} (Y - \overline{Y})$

or $(X-52.2) = (0.72) (Y-141.7)$

or $X - 52.2 = 0.72Y - 102.02$

or $X = -49.82 + 0.72Y$ Ans.

5.6.8 Finding Mean Value of Variables and Correlation Co-efficient from the given regression equations :

If the two regression equations are given and the average or arithmatic means of X and Y variables as well as the correlation co-efficient are to be found, then following steps should be followed :

STEP I Assume one equation as regression equation of 'Y on X' and second equation as regression equation of 'X on Y'

STEP II Multiply the regression co-efficient and take square root to get correlation co-efficieint If $-1 \leq r \leq +1$, then proceed further. If not, then reverse the assumption of regression equations in Step I and start again to calculated the correlation co-efficieint as in Step II.

STEP III If -1 $\leq$ r $\leq$ +1, then solve the regression equation of 'Y on X' and 'X on Y' simultaneously for the values of X and Y.

STEP IV The value of X is $\overline{X}$ and the value of Y is $\overline{Y}$

Example :

Two random variables have the regression equations 3X +2Y-26 = 0 and 6X + Y - 31 = 0. Find out the mean values of X and Y and their co-efficient of correlations.

Solution :

Suppose 3X +2Y-26 = 0 is the regression equation of

X on Y and 6X + Y - 31 = 0 is the regression equation of Y on X. Then

$$X = \frac{26}{3} - \frac{2}{3}Y \text{ and } Y = 31 - 6X$$

Here $b_{XY} = -\frac{2}{3}$ and $b_{YX} = -6$.

We know that $r = \pm\sqrt{b_{XY} \times b_{YX}} = -\sqrt{\left(-\frac{2}{3}\right)(-6)}$

$= -\sqrt{4}$ = -2 which is not possible, as it falls outside the limit.

Now reversing over assumption, suppose 3X+2Y-26 = 0 as regression equation of Y on X and 6X + Y -31 = 0 as the regression equation of X on Y.

So, $Y = \frac{26}{2} - \frac{3}{2}X$ or $Y = 13 - \frac{3}{2}X$ and $X = \frac{31}{6} - \frac{1}{6}Y$

$\therefore$ $b_{XY} = -\frac{1}{6}$ and $b_{YX} = -\frac{3}{2}$

Now, $r = \pm\sqrt{\left(-\frac{1}{6}\right)\left(-\frac{3}{2}\right)} = -\sqrt{\frac{1}{4}}$

$\therefore r = -0.5$ Ans.

Now solving 3X + 2Y - 26 = 0 1

and 6X + Y - 31 = 0 2

Simultaneously for X and Y, we multiply 1 by 2 and subtract it from 2

(6X + Y -31) - 2(3X+2Y-26)=0

$\therefore$ 6X+Y-31-6X-4Y+52=0

or -3Y+21 = 0

or Y = 7

or $\overline{Y}$ = 7 Ans.

Putting value of Y in 1, we get

3X + (2 x 7) - 26 = 0

or 3X + 14-26 = 0

or 3X - 12 = 0

or 3X = 12

or X = 4

or $\overline{X}$ = 4 Ans.

5.7 PROPERTIES OF THE REGRESSION CO-EFFICIENTS:-

1. The regression co-efficient are independent of change of origin but not change of scale.
2. The geometric mean of the regression co-efficient is equal to the correlation co-efficieint, i.e., $r = \pm\sqrt{b_{XY} \times b_{YX}}$
3. Both the regression co-efficieint will have the same sign, either both postive or

both negative.

4. Average of the regression co-efficieint is always greater than or equal to the correlation co-efficieint

   i.e. $= \frac{b_{XY} + b_{YX}}{2} \geq r$

5. Atleast one of the regression co-efficients must have the value less than unity.

6. Both regression co-efficients and the correlation co-efficieint will have same sign i.e. $\pm r = \pm\sqrt{(\pm b_{XY}) \times (\pm b_{YX})}$

7. If the correlation co-efficient is zero, then both the regression co-efficients will be zero.

8. The regression co-efficient are amenable to algebric treatment.

9. If the standard deviations of both the variables are same, then the correlation co-efficieint will be equal to both the regression co-efficieint.

5.8 STANDARD ERROR OF ESTIMATE :-

The statistical measure which indicates the accuracy of the production or estimates of dependent variable on basis of independent variable is known as the 'standard Error of Estimate'. It is denoted by $S_{1.2}$ where in subscript 1.2, the 1 stands for dependent variable and 2 stands for independent variable.

The Standard error of estimate for regression equation Y on X is $S_{Y.X}$ = $= \sqrt{\frac{\Sigma(Y - Y_C)^2}{N}}$ when computed from a population. Here, Y stands for original value of Y series, Y_c stands for estimated values of Y-series using the regression equation of Y on X and N stand for number of pairs of observations.

This formula changes to $S_{Y.X} = \sqrt{\frac{\Sigma(Y - Y_C)^2}{N - 2}}$ when computed from a sample rather than a population. The different denominator (N-2) is used because two constants 'a' (intercept) and 'b' (slope) are to be computed or estimated from the sample values to calculate the 'standard Error of Estimate'. The alternative formulas for computing 'Standard Error of Estimate' are $S_{Y.X} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N}}$ when computed from a population and $S_{Y.X} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N-2}}$ when computed from a sample. Simlarly, we can calculate the 'Standard Error of Estimate' for regression equation of X on Y, denoted by $S_{X.Y}$. The smaller is the value of standard error of Estimate, more accurate the estimates are and closer will be the estimated values to the actual values. In case of zero standard error of estimate, the estimated and the actual values of the variable will be same. The correlation will be perfect in that case.

5.9 LET US SUM UP :

The regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is ued for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanator while the dependent variable is also known as regressed or explained variable.

5.10 LESSON END EXERCISE :

Q1. What do you mean by regression ? Why are there two regression lines in case of a bivariate series.

Q2. Distinguish between correlation and regression analysis and indicate the utility of regression analysis in economic activities.

Q3. Fit a least square line to the following data :

(a) Using x as independent variable (b) using x as dependent variable.

x	:	1	3	4	8	9	11	14
y	:	1	2	4	5	7	8	9

Hence obtain

i. the regression coefficient of y on x and x on y

ii. $\bar{x}$, $\bar{y}$

iii. Co-efficient of correlation between x and y.

Q4. Calculate the standard error of the estimate of y from the regression of y on x for the following data

$n= 10 \ \sum y^2 = 90 \ \sum xy = 120, \ \sum x^2 = 200$

where, $x = x-\bar{x}$, $y = y-\bar{y}$

Q5. Write the properties of regression co-efficients.

SUGGESTED FOR FURTHER READINGS

- Gupta, S.P (2011), Statistical methods, Sultan Chand and Sons Educational Publisher, New Delhi.
- Gupta, S. C. (2011), Fundamental of Statistics, Himalaya Publisher House, New Delhi
- Das, N.G (2012), Statistical Methods, Tata McGraw Hill Education Private Limited, New Delhi
- Yamane, T (1970), Statistics : An Introductory Analysis, Harper and Row, New York.
- Hoel, P.G. (1954), Introduction to Mathematical Statistics, Wiley and Sons.

****

M.A. Economics **Semester - IInd** **Lesson - 6**

Course No. 203 **Unit - II**

NATURE AND PURPOSE OF INDEX NUMBERS, COMMONLY USED INDEX NUMBERS, LASPEYRES AND PAASCHE'S INDEX NUMBER

CHAPTER HIGHLIGHTS :-

This lesson inculcates the ideas about the nature and purpose of index numbers various used index numbers i.e. Laspeyres and Paasche's index numbers.

CHAPTER OUTLINES :

6.7 Lesson End excercise

6.1 INTRODUCTION :

Historically, the first index number was constructed by an Italian, Mr. Carlia, in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though orginally developed for measuring the effect of change in prices, index numbers have become today one of the most widely used statistical devices and there is hardly any field where they are not used Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies. Infact, they are described as 'barometers of economic activity'. i.e. if one wants to get an idea as to what is happening to an economy, he should look to important indices like the index number of industrial production, agricultural production, business activity etc.

6.2 NATURE AND PURPOSE OF INDEX NUMBERS :

Meaning of Index Numbers :- An index number may be defined as a measure of the average change in a group of related variables over two different situations. The group of variables may be the prices of a specified set of commodities, the volumes of production in different sectors of an industry, the marks obtained by a student in different subjects and so on. The two different situations may be either two different times or two different places.

INDEX NUMBERS DEFINED

According to Berenson and Levine, "Generally speaking, index numbers measure the size or magnitude of some object at a particular point in time as a percentage of some base or reference object in the past."

In the words of Croxton and Cowden, "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

F.Y. Edgeworth gave the classical definition of index number as follows :

"Index numbers shows by its variations the changes in a magnitude which is not susceptible.either of accurate measurement in itself or of direct valuation in pratice".

Index numbers are statistical devices designed to measure the relative change in the level of a phenomenon (Variable or a group of variables) with respect to time, geographical location or other characteristics such as income, profession etc." In other words, index numbers are specialised type of rates, ratios, percentage which give the general level of magnitude of a group of distinct but related variables in two or more situations.

e.g. :- Suppose we are intersted in studying the general change in the price level of consumer goods i.e., goods or commodities consumed by the people belonging to a particular section of society, say, low income group or middle income group or labour class and so on. Obviously these changes are not directly measurable as the price quotations of the various commodities are available in different units e.g., cereals (wheat, rice, pulses etc) are quoted in Rs per quintal or kg, water in Rs. per gallon; milk, petrol, kerosene etc, in Rs. per litre; cloth in Rs per metere and so on.

It is clear from the above definitions that an index number is a specialized average designed to measure the change in a group of related variable over a period of time. Thus, when we say that the index number of wholesale prices is 112 for Jan 2011 compared to Jan 2010, it means there is a net increase in the prices of wholesale commodities to the extent of 12 per cent during the year.

For a proper understanding of the term index number, the following points are worth considering -

i. ***Index number are specialized averages*** : Index numbers are used for purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items e.g. while constructing a consumer price index the various items are divided into broad heads namely (i) food, (ii) clothing, (iii) Fuel etc. these items are expressed in different units : thus under the head, food, wheat and rice may be quoted per quintal, ghee per kg etc. Similarly, cloth may be measured in terms of meters. An average of all these items expressed in different units is obtained by using the technique of index numbers.

ii. ***Index numbers measure the net change in a group of related variables***: Since index numbers are essentially averages they describe one single figure the increase or decrease in a group of related variables under study. The group of variables may be the prices of a specilised set of commodities, the volume of production in different sectors etc. It should be noted carefully noted that even where an index is showing a net increase, it may include some items which have actually decreased in value and others which have remained constant.

iii. ***Index numbers measure the effect of changes over a period of time*** : It should be noted that index numbers not only measure change over a period of time but also compare economic conditions of different locations, different industries, different cities.

PURPOSE OF INDEX NUMBERS

The purpose for which the index number is being constructed should be clearly and unambiguously stated, for instance, if we want to construct an index number for measuring the change in the gereral price level, we have to take the wholesale prices of finished products, intermediate products, agricultural products, mineral products etc. Similarly, the retail prices of consumer goods and the costs of services like electricity charges form the basis for the construction of a cost of living index number.

Index number are very powerful statistical tools for measuring the changes in the level of any phenomenon over two different periods of time. They are used in almost all sciences-natural, social and physical. The main uses of index numbers can be summarised as follows :-

i. ***Index numbers as Economic Barometers*** :- Like barometers which are used in Physics and chemistry to measure atmospheric pressure, index numbers are rightly termed as ‘economic barometers’ or barometers of economic activity, which measure the pressure of economic and business behaviour.

ii. ***Index Numbers Help in Studying Trends and Tendencies***: Since the index numbers study the relative changes in the level of a phenomenon at different periods of time, they are specially useful for the study of the general trend for a group phenomenon in a time series data. For instance, if a businessman is interested in establishing a new industry, the study of the trend of changes in the prices, wages and incomes in different industries is extremely helpful to him to frame a general idea of the comparative courses which the future holds for different undertakings.

iii. ***Index number help in formulating decisions and policies***: Index numbers of the data relating to prices, production, profits, imports and exports, personnel and financial matters are indispensable for any organisation in efficient planning and formulation of executive decisions.

iv. ***Price Indices Measure the pruchasing power of Money*** : The cost of living index numbers determine whether the real wages are rising or falling, money wages remaining unchanged. In other words, they help us in computing the real wages which are obtained on dividing the money wages by the corresponding price index and multiplying by 100. Real wages help us in determining the purchasing power of money. e.g. Suppose that the cost of living index for any year, say, 1979 for a particular class of people with 1970 as base year is 150. If a person belonging to that class get Rs 300 in 1970, then in order to maintain the same standard of living as in 1970, his salary in 1979 should be $\frac{150}{100}\times 300 = \text{Rs } 450$. In other words, if a person gets Rs. 450 in 1979, then his real wages are $\frac{450}{150}\times 100 = \text{Rs } 300$ i.e, the pruchasing power of money has reduced to 2/3.

v. ***Index Numbers of used for Deflation*** : Consumer price indices or cost of living index numbers are used for deflation of net national product, income value series in national accounts. The techniquie of obtaining real wages from the given nominal wages can be used to find real income from inflated money income, real sales from nominal sales and so on by taking into account appropriate index numbers.

Index numbers are primarily used to measure the relative position of business and economic conditions. There are many different types of index numbers and the use of an index number depends on its type; Index numbers of wholesale prices, retails prices cost of living, industrial production, quantum of exports and imports, business activity, to name only a few are useful in their own fields.

Price index numbers are used for various purposes. Wholesale price index number tells us about changes taking place in the value of money. Consumer price index number or cost of living index number measures changes in the real income of people. It helps in the calculation of dearness allowance, so that the real wage may not decrease. Index numbers of stock prices are used by economists, speculators and bankers in various ways. An economist uses them to measure changes in the purchasing power of money over stocks, a speculator uses them for forecasting the future course of the market, and the insurance company may require the index numbers for estimating future interest rate. Similarly, index number of industrial production reveals the comparative position in productivity and index number of business activity throws light on the progress of business conditions.

Index numbers are also used to measure the comparative position in respect of price in different regions at the same period of time. e.g. for comparing the standards of living in several cities.

6.3. COMMONLY USED INDEX NUMBER:

Index numbers may be broadly classified into various categories depending upon the type of the phenomenon or variable in which the relative changes are to be studied. Although index numbers can be constructed for measuring relative changes in any field of quantitative measurement, we shall primarily confine the discussion to the data relating to economics and business ie. data relating to prices production and consumption. In this context index numbers may be broadly classified into following three categories :-

6.3.1 Price Index Numbers :-

The price index numbers measure the general changes in the prices. They are further sub-divided into the following classes :-

a. **Wholesale Price Index Numbers** :- The wholesale price index numbers reflected the changes in the general price levelof a country.

b. **Retail price index Numbers** : These indices reflect the general changes in the retail price of various commodities such as consumption goods, stocks and shares, bank deposits, good bonds, etc. India these indices are constructed by labour ministry in the form of labour bureau Index Number of retail price Urban centres and Rural Centres.

Consumer price index, commonly known as the cost of living index is a specialised kind of retial price index.

2. Quantity Index Numbers :-

Quantity index numbers study the changes in the volume of goods produced, consumed or disturbed like the indices of agricultural production, Industrial production, imports and exports etc. They are extremely helpful in studying the level of physical output in an economy.

3. Value Index Numbers :

These are intended to study the change in the total value of production such as indices of retail sales or profits or inventories. However, these indices are not as common as price and quantity indices.

We shall now discuss the various techniques or methods used for index numbers.

Simple (Unweighted) Aggregate Method :- This is the simplest of all the methods of constructing index numbers and consist in expressing the total price i.e. aggregate of prices (of all the selected commodities) in the current year as a percentage of the aggregate of prices in the base year. Thus, the price index for the current year w.r.t. the base year is given by

$$P_{OI} = \frac{\sum P_1}{\sum P_0} \times 100$$

$\sum P_1$ = Total of current year prices for various commodities.

$\sum P_0$ = Total of base year price for variou commodities.

Based on this method, the quantity index is given by the formula

$$Q_{OI} = \frac{\sum q_1}{\sum q_0} \times 100$$

Where $\sum q_0$ and $\sum q_1$ are the quantities of all the selected commodities consumed in the base year and the current year respectively.

Example :

From the following data calculate Index number by simple aggregative method.

Commodity	A	B	C	D
Price in 1980 (Rs)	162	256	257	132
Price in 1981 (Rs)	171	164	189	145

Solution :

Computation of Price Index number

Commodity	Price (in Rupees)	
	1980 (P_o)	1981 (P_I)
A	162	171
B	256	164
C	257	189
D.	132	145
Total	$\sum p_0$=807	$\sum p_1$=669

The price index number using simple Aggregate method is given by :

$$P_{OI} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$= \frac{669}{807} \times 100 = 82.90$$

Weight Aggregate Method :- In this method, appropriate weights are assigned to various commodities to reflect their relative importance in the group. The weights can be production figures, consumption figures or distribution figures. For the construction of price index numbers, quantity weights are used i.e. the amount of the quantity consumed, purchased or marketed. If w is the weight attached to a commodity, then the price index is given by

$$P_{OI} = \frac{\sum wp_1}{\sum wp_0} \times 100$$

Where 'w' represents the weight. It should be noted that the same set of weights must be used for base year as well as for current year.

In the construction of Price index, quantitives (q) are used as weights. There are several formula for weighted aggregative index depending on the nature of weights employed :

i. If the base year quantity (q_0) is used as weight, i.e. w = q_0, we get

Laspeyres's Index (I_{on}) $= \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$

ii. If the current year quantity (q_n) is used as weight, i.e. w =q_n we get

Paasche's Index (I_{on}) $= \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$

iii. If the sum of quantities in the base year and the current year is used as weight, ie. w ($q_0 + q_n$) we get

Edgeworth - Marshall's Index (I_{on}) $= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$

iv. The geometric mean (i.e. square root of the product) of Laspeyres Index and Paasche's index is of special importance, because of certain properties **Fisher's Ideal Index**

$$= \sqrt{(\text{Laspeyres Index}) \times (\text{Paasche's Index})}$$

$$= \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} \times \frac{\Sigma P_n q_n}{\Sigma P_o q_n}} \times 100$$

The following index number of the weighted aggregative type are also sometime used :

v. The arithimetic mean of Laspeyres's index and Paasche's index is known as **Bowley's Index**

$$= \frac{1}{2} \text{(Laspeyres's Index + Paasche's index)}$$

$$= \frac{1}{2}\left[\frac{\Sigma P_n q_o}{\Sigma P_o q_o} \times \frac{\Sigma P_n q_n}{\Sigma P_o q_n}\right] \times 100$$

vi. If the geometric mean of basic year and current year quantities is used as weight i.e. $w = \sqrt{q_o q_n}$ we get

Walsh's Index = $= \dfrac{\Sigma p_n \sqrt{q_0 q_n}}{\Sigma p_0 \sqrt{q_0 q_n}} \times 100$

vii. If the weight used are kept fixed for all periods ie. weight are constant quantities (q), without any reference to base or current period, we get

Kelly's Index = $= \dfrac{\Sigma p_n q}{\Sigma p_0 q} \times 100$

This is also known as 'Aggregative Index with fixed weights."

Relative Method :- In this method, the price of each item in the current year is expressed as a percentage of the price in the base year. This is called price relative and is given by formula

Price Relative $= \dfrac{\text{Price in the given year}}{\text{Price in the base year}} \times 100$

$$\text{Price Relatives} = \frac{P_n}{P_0} \times 100$$

The average of price relatives, which shows the average percentage change for the whole group of items, gives the index number.

Price index = Average of price relatives.

Usually A.M or G.M is used for averaging the relatives. The special cases, H.M. or median, is also used again, the average employed may be either simple or weighted. If a simple average is used, the index number is called simple Average of Relative Index. If a weighted average is used, it is known as weighted average of relatives index. Thus,

Simple A.M. of Relatives Index

$= \sum$ (Price Relatives ÷ K, where K is the number of items included)

Simple G.M of Relative Index

$$= \sqrt[K]{\text{Product of price relatives}}$$

Weighted A.M of Relatives Index

$$= \frac{\sum(\text{Price Relatives}) \times w}{\sum w}$$

The weights (w) employed for averaging price relatives are the values (= price x quantity) of items. In most cases, these values are given not in absolute units, but as percentage of the total value for all the items. i.e. the weights are given as pure numbers.

6.3.2 Quantity Index Numbers

Just as price index numbers measure and permit comparison of the price of a group of related items, quantity Index numbers similarly measure and permit comparison of the physical quantity of goods produced or consumed or marketed. Quantity index number formulae may be obtained from the corresponding price index number formulae replacing p and q and q by p.

Simple Aggregative Quantity Index $= = \frac{\sum q_n}{\sum q_0} \times 100$

Laspeyres' Quantity Index $= \frac{\sum q_n p_0}{\sum q_0 p_0} \times 100$

Paasche's Quantity Index $= \frac{\sum q_n p_n}{\sum q_0 p_n} \times 100$

Edge-worth-Marshall' Index $= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$

Fisher's Ideal Index $= \sqrt{\frac{\sum P_n q_o}{\sum P_o q_o} \times \frac{\sum P_n q_n}{\sum P_o q_n}} \times 100$

Quantity Relative $= \frac{q_n}{q_0} \times 100$

Simple A.M. of Quantity Relatives Index

$= \sum$ (Quantity Relatives) ÷ K

Weighted A.M of Quantity Relatives Index =

$$= \frac{\sum(\text{Quantity relaive} \times \text{Weight})}{\sum(\text{Weight})}$$

6.4 LASPEYRES AND PAASCHE'S INDEX NUMBERS :

Laspeyres Method :- The Laspeyres Price Index is a weighted aggregate price index, where the weights are determined by quantities in the base period. The formula for constructing the index is $P_{OI}^{La} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$

$P_n \rightarrow$ Price in Current Period.

$q_0 \rightarrow$ Quantity in base period.

Steps :- Multiply the current year prices of various commodities with base year weights and obtain $\sum P_n q_0$

- Multiply the base year prices of various commodities with base year weights and obtain $\sum p_0 q_0$.
- Divide $\sum p_n q_0$ by $\sum p_0 q_0$ and multiply the quotient by 100. This gives us the price index.

Laspeyres Index attempts to answer the quesion : "What is the change in aggregate value of the base period list of goods when valued at given period prices?" This index is very widely used in practical work.

The primary disadvantage of the Laspeyres method is that it does not take into consideration the consumption pattern. The Laspeyres index has an upward bias. When price increase, there is a tendency to reduce the consumption of higher priced item. Hence, by using base year weights, too much weight will be given to those items which have increased in price the most. Similarly, when prices decline, consumers shift their purchaes to those items which decline the most. By using base period weights, too litle weight is given to those items which decrease most in price, again overstating the index.

Example :

Calculate the price index number for the year 1978 with 1976 as base using Laspeyres' formula on the basis of the following table :-

Commodity	Price in (Rs)		Money value ('oooRs)
	1976	1978	1976
A	12.50	14.00	112.50
B.	10.50	12.50	126.00
C.	15.00	14.00	105.00
D	9.40	11.20	47.00

Calculations for Laspeyre's Price Index

Commodity	P_o	P_n	P_0q_o	q_o = (4), (2)	p_nq_o
(1)	(2)	(3)	(4)	(5)	(6)
A	12.50	14.00	112.50	9	126.00
B	10.50	12.00	126.00	12	144.00
C	15.00	14.00	105.00	7	98.00
D	9.40	11.20	47.00	5	56.00
Total	-	-	390.50	-	424.00

Laspeyre's price index $= \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$

$$= \frac{424.00}{390.50} \times 100$$

= 109

Paasche's Method :- The Paasche price index is a weighted aggregate price index in which the weights are determined by quantities in the given year. The formula for constructing the index is

Paasche's Index $= \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$

$p_n \rightarrow$ Price in current year

$q_n \rightarrow$ Quantity in current year

$p_o \rightarrow$ Price in base year

<u>Steps</u>

- Multiply current year prices of various commodities with current year weights and obtain $\sum p_n q_n$
- Multiply the base year prices of various commodities with current year weights and obtain $\sum p_0 q_n$

- Divide $\sum p_nq_n$ by $\sum p_0q_n$ and multiply the quotient by 100.

In general this formula answers the question : "What would be the value of the given period list of goods when valued at base period prices".

The difficulty in computing the Paasche' index in practice is that revised weight or quantities must be computed each year or each period. For this reason, the Paasche index is not used frequently in practice where the number of commodities is large.

Q. From the following price and quantity data compute Paasche's price index number for 1980 with 1970 base year

Commodity	Price (per kg)		Quantities Sold (Kg)	
	1970	1980	1970	1980
Commodity A	4	5	95	120
Commodity B	60	70	118	130
Commodity C	35	40	50	70

Calculations for Paasche's Price Index

Commodity	p_0	p_n	q_0	q_n	p_0q_n	p_nq_n
A	4	5	95	120	480	600
B	60	70	118	130	7800	9100
C	35	40	50	70	2450	2800
Total					10730	12500

$$\text{Paasche's Price Index } = \frac{\sum p_nq_n}{\sum p_0q_n} \times 100$$

$$= \frac{12500}{10730} \times 100 = 116$$

6.5 COMPARISON OF LASPEYRES AND PAASCHE METHODS :-

Laspeyre's price index is based on the assumption that the quantities consumed in the base year and the current year are same, an assumption which is not true in general. If the consumption of some of the commodities or items decreases in the current year due to rise in their price or due to changes in the habits, tastes and customs of the people, then Laspeyre's index which is based on base year quantities as weights as weights given relatively more weightage for such commodities and consequently the numerator in is relatively larger.

Hence, Laspeyre's index is expected to have an 'upward bias' as it over estimates the true value similary, if the consumption of certain commodities increases in the current year due to decrease in their prices, then Paasche's index which uses current year quantities as weights, gives more weightage to such commodites. Accordingly, Paasche's index has a 'downward bias' and is expected to under estimate the true value. However, it should not be inferred that Laspeyre's index must be larger than Paasche's index always. The conditions under which Laspeyre's index is greater than, equal to or less than Paasche's index have been obtained.

If the price of all the goods change in the same ratio then Laspeyre's and Paasche's price index numbers will be equal, for then the weighting system is irrelevant; or if the quantities of all the goods change in the same ratio, they will be equal for then the two weighting system, are the same relatively.

In general, the true value of the price index lies somewhere between the two.

Concepts of over and under estimate

The numerical explanation of Laspeyres and Paache's Index Number

Since, $r = \frac{\text{Cov}(x,y)}{\sigma x \sigma y}$

Where, $\text{Cov}\ (x,y) \le 0$

And, $\frac{\Sigma fxy}{N} - \frac{\Sigma fx}{N} . \frac{\Sigma fy}{N} \le 0$

Where, x - price relative
Y – quantity relative
F – base value
(Now, omitting the factor 100)

And, $x=\frac{Pn}{Po}$, $y = \frac{qn}{qo}$ & $f = poqo$

And, $x=\frac{Pn}{Po}$, $y = \frac{qn}{qo}$ & $f = poqo$

$$= \text{Cov}\ \left(\frac{pn}{po},\frac{qn}{qo}\right) = \frac{\Sigma poqo.\frac{pn}{po}.\frac{qn}{qo}}{\Sigma poqo} - \frac{\Sigma poqo.\frac{pn}{po}}{\Sigma poqo}.\frac{\Sigma poqo.\frac{qn}{qo}}{\Sigma poqo}$$

$$= \text{Cov}\ \left(\frac{pn}{po},\frac{qn}{qo}\right) = \frac{\Sigma pnqn}{\Sigma poqo} - \frac{\Sigma pnqo}{\Sigma poqo}.\frac{\Sigma poqn}{\Sigma poqo}$$

$$= \text{Cov}\ \left(\frac{pn}{po},\frac{qn}{qo}\right) = \frac{\Sigma pnqn}{\Sigma poqn} . \frac{\Sigma poqn}{\Sigma poqo} - \frac{\Sigma pnqo}{\Sigma poqo}.\frac{\Sigma poqn}{\Sigma poqo}$$

$$\text{Cov}\ \left(\frac{pn}{po},\frac{qn}{qo}\right) = Pp.Lq - Lp.Lq \le 0$$

Where Pp – Paache's price index
Lp – Laspeyres price index
Lq – Laspeyres quantity index.

$$\text{Cov } \left(\frac{pn}{po}, \frac{qn}{qo}\right) = \text{Lq [Pp-Lp]} \leq 0$$

But "Lq" is positive and can't be equal to zero
$\therefore$ Pp – Lp $\leq$ 0
Or Pp $\leq$ Lp

6.6 FISHER'S TEST FOR INDEX NUMBERS

Several formulae have been suggested for constructing index numbers and the problem is that of selecting the most appropriate one in a given situation.

In order to judge the efficiency of an index number formula as a measure of the level of a phenomenon from one period to another, the noted economist Irving fisher suggested certain tests. The three most important tests of index numbers are :

1. Time revrsal test
2. Factor reversal test
3. Circular test.

These tests are based on the analogy that what for an individual item should also hold for a group of items.

6.6.1 Time reversal test :

Prof. Irving fisher has made a careful study of the various proposals for computing index numbers time reversal test is a test to determine whether a given method will work both ways in time, forward and backward.

In the words of fisher, "The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base." In other words, when the data for any two of years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals of each other so that their product is unity. Symbolically, the following relation should be satisfied :

$$\text{Po x Po} \overset{x}{=} 1 \qquad \text{Ion x Ino} = 1$$

where Ion is the index for time ‘n’ on time “O” as base and I_{no} is the index for time “O” on time “h” as base. If the product is not unity, there is said to be a time bias in the method. Thus, if from 2010 to 2011 the price of wheat increased from Rs. 1920 to 2560 per qunintal. The price in 2011 should be $133\frac{1}{3}$ percent of the price in 2010 and the price in 2010 should be 75 percent of the price in 2011. One figure is the reciprocal of the other, their product (1.33 x 0.75) is unity. This is obviously true for each individual price relative and, according to the time reversal test, it should be true for the index number.

Time reversal test is satisfied by the following index number formulae.

i. Simple aggregate index.

ii. Marshal - edgeworth formula

iii. Walsh formula

iv. Kelly’s fived weight formula

v. Simple geometric mean of price relatives formula with fixed weights.

vi. Weighted geometric mean of price relatives formula with fixed weights.

vii. Fisher’s ideal formula

Laspeyre’s and Paasche’s formula do not satisfy this test.

6.6.2 Factor reversal test :

Another test suggested by fisher is known as factor reversal test. It holds that the product of a price index and quantity index shoud be usual to the corresponding value index :

In the words of fisher, just as each formula should permit the interchange of the two times without giving incosistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent result, i.e. the two results multipled together should give the true value ratio. In other words the test is that the change in price multiplied by the change in quantity should be equal to the total change in value. The total value of a given commodity in a given year and the product of the quantity and the price per unit (total value = P x Q)

Symbolically $P_{on} \text{ x } Q_{on} = \dfrac{\sum P_n q_n}{\sum P_o q_o}$

Fisher's ideal index is the only formula which satisfied this test.

6.6.3 Circular test :

Another test of index number formula is what is known as 'circular test'. If in the use of index numbers interest attaches not merely to a comparison of two years, but to the measurement of price changes, over a period of years, it is frequently desirable to shift the base is again 200, then the 2001 as the base. A formula is said to meet this test if, for example, the 2011 index with 2006 as the base is 200, and the 2006 index with 2001 as the base of must be 400. Clearly, the desirability of this property is that it enables us to adjust the index values from period to period without referring each time to the original base. A test of this shiftability of base is called the circular test.

This is an extension of time reversal test. An index number formula is said to satisfy the circular test, if the time reversal test is satisfied through a number of intermediate years. Symbolically

$$I_{o1} \text{ x } I_{12} \text{ x } I_{23} \ldots \text{x } I_{(n-1)},\ n \text{ x } I_{no} = 1$$

This means that the relation is satisfied in a circular fashion through several years, 0 to 1, 1 to 2, 2 to 3... (m-1) to n, and finally from n back to O, simple aggregative formula and the simple geometric mean of relatives formula satisfy this test.

6.6.4 Proofs related to tests

Show that neither Laspeyres' formula nor Paasche's formula obeys time reversal or factor reversal tests of index numbers.

Solution :

I. Time reversal test may be symbolically expressed as $I_{on} \times I_{no} = 1$

a) Using Laspeyres' price index formula and omitting the factor 100,

Index number for year n with base year $o(I_{no}) = \frac{\Sigma P_n q_o}{\Sigma P_o q_o}$

Interchanging the suffixes O and n,

Index number for year O with base $n(I_{no}) = \frac{\Sigma P_o q_n}{\Sigma P_n q_n}$

$$\therefore \quad I_{on} x I_{no} = \frac{\Sigma P_n q_o}{\Sigma P_o q_o} x \frac{\Sigma P_o q_n}{\Sigma P_n q_n} \neq 1$$

Thus, Laspeyres' formula does not obey time reversal test.

b) Using Paache's price index formula and omitting the factor, 100,

$$I_{on} = \frac{\Sigma P_n q_n}{\Sigma P_o q_n}$$

Interchanging the suffixes O and n, $I_{on} = \frac{\Sigma P_o q_o}{\Sigma P_n q_o}$

$$\therefore \quad I_{on} x I_{no} = \frac{\Sigma P_n q_n}{\Sigma P_o q_n} x \frac{\Sigma P_o q_o}{\Sigma P_n q_o} \neq 1$$

Thus, Paache's formula also does not obey time reversal test

II. Factor reversal test may be expressed as

$$P_{on} \text{ x } Q_{on} = \frac{\sum P_n q_n}{\sum P_o q_o}$$

c) Using Laspeyres' formula, and omitting the factor 100,

Price Index for year n with base year $o(P_{on}) = \dfrac{\sum P_n q_o}{\sum P_o q_o}$

Interchanging p and q, quantity index for year n with base year

$$Q_{on} = \frac{\sum P_n q_o}{\sum q_o P_o} = \frac{\sum P_o q_n}{\sum P_o q_o}$$

Multiplying, we have by Laspeyres' formula

$$P_{on} \text{ x } Q_{on} = \frac{\sum P_n q_o}{\sum P_o q_o} \text{ x } \frac{\sum P_o q_n}{\sum P_o q_o} \neq \frac{\sum P_n q_n}{\sum P_o q_o}$$

i.e. $P_{on} \text{ x } Q_{on} \neq \dfrac{\sum P_n q_n}{\sum P_o q_o}$

This proves that Laspeyres' formula does not satisfy factor reversal test.

d) Applying Paasche's formula, it will be found that

$$P_{on} \text{ x } Q_{on} = \frac{\sum P_n q_n}{\sum P_o q_n} \text{ x } \frac{\sum P_n q_n}{\sum P_o q_o} \neq \frac{\sum P_n q_n}{\sum P_o q_o}$$

This proves that Paasche's formula does not satisfy factor reversal test.

Q. Examine whether fisher's ideal index formula satisfies the time reversal and factor reversal tests.

Solution :

Using fisher's "ideal" index formula, price index for year n with base year O is given by (omitting the factor 100)

$$I_{on} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n}}$$

Interchanging the suffixes O and n, price index for year o with base year n is

$$I_{no} = \sqrt{\frac{\Sigma P_o q_n}{\Sigma P_n q_n} . \frac{\Sigma P_o q_o}{\Sigma P_n q_o}}$$

Mulitplying we get

$$I_{on} \text{ x } I_{no} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n}} \sqrt{\frac{\Sigma P_o q_n}{\Sigma P_n q_n} . \frac{\Sigma P_o q_o}{\Sigma P_n q_o}}$$

$$I_{on} \text{ x } I_{no} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n} . \frac{\Sigma P_o q_n}{\Sigma P_n q_n} . \frac{\Sigma P_o q_o}{\Sigma P_n q_o}} = \sqrt{1} = 1$$

(since all term cancel one another)

i.e. $I_{on} \text{ x } I_{no} = 1$

This shows that fisher's ideal formula obeys time reversal test. In order to apply factor reversal test, we see that price index by fisher's ideal formula is

$$P_{on} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n}}$$

Interchanging p and q, quantity index is given by

$$Q_{on} = \sqrt{\frac{\Sigma q_n P_o}{\Sigma P_o q_o} . \frac{\Sigma q_n P_n}{\Sigma q_o P_n}}$$

$$Q_{on} = \sqrt{\frac{\Sigma q_o P_n}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_n q_o}} \text{, (rearranging the factor p, q)}$$

Multiplying $P_{on} = Q_{on}$, we have

$$P_{on} \text{ x } Q_{on} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n}} \sqrt{\frac{\Sigma P_o q_n}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_n q_o}}$$

$$P_{on} \text{ x } Q_{on} = \sqrt{\frac{\Sigma P_n q_o}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_n} . \frac{\Sigma P_o q_n}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_n q_o}}$$

$$P_{on} \text{ x } Q_{on} = \sqrt{\frac{\Sigma P_n q_n}{\Sigma P_o q_o} . \frac{\Sigma P_n q_n}{\Sigma P_o q_o}} = \frac{\Sigma P_n q_n}{\Sigma P_o q_o}$$

This proves that Paache's price index under estimates and Laspeyres price index over estimates the compensation variation.

Q

Commodity	Price		Quantity	
	Po	Pn	Qo	Qn
Rice	32	30	50	50
Barley	30	25	40	35
Maize	16	18	50	55

Sol [Solve it yourself]

....

....

6.7 LESSON AND EXCERCISE :

Q1. What are index number ? Explain the nature and purpose of index number.

Q2. "Index Numbers are economic barometers" - Explain.

Q3. Find the Simple Aggregative index number from the following data :

Commodity	Base Price	Current Price	Weight
Rice	140	180	10
Oil	400	550	7
Sugar	100	250	6
Wheat	125	150	6
Fish	200	300	4

Hint : Ans 149

Q4. Calculate the price index number by (a) Paasche' method (b) Laspeyre' method (c) Bowley's method

Commodities	1979		1980	
	Price (Rs)	Qty (Kgs)	Price (Rs)	Qty (Kgs)
A	20	8	40	6
B	50	10	60	5
C	40	15	50	10
D	20	20	20	15

Hint : ***Pa = 125.2***

La = 124.7

Bow = 125.0.

Q5. Explain the term 'Price Relative'. Find by Arthmetic Mean method the index number from the following :

Commodity	Base Price	Current Price	Weight
Rice	30	52	8
Wheat	25	30	6
Fish	130	150	3
Potato	35	49	5
Oil	70	105	7

Hint : Ans 145

Q6. Calculate a suitable index number from the data given below

Commodity	Price Relative	Weight
A	125	5
B	67	2
C	250	2

Hint : Ans 150.9

Q7. Explain the time reversal and factor reversal text.

Q8. Examine whether Laspeyre's and Paasche's index numbers satisfy these tests.

****

CHAIN BASE INDEX NUMBERS, OFFICIAL INDEX NUMBERS, TRUE COST OF LIVING INDEX, FISHER'S TEST FOR INDEX NUMBERS

CHAPTER HIGHLIGHTS:

This lesson contains the information about the various index numbers like chain base index numbers, official, true cost of living indexes. Along with this, discussion also revolves around the fisher's test for index numbers.

CHAPTER OUTLINES

7.1 INTRODUCTION :

In the fixed base method, discussed so far the base remains the same throughout the series of the index. This method, though convenient, has certain limitations. As time elapses conditions which were once important become less significant and it becomes more difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases it may be desirable to use the chain base method. When this method is used the comparisons are not made with a fixed base rather the base changes from year to year. For example, for 2011, 2010 will be the base; for 2010, 2009 will be the base and so on. If however, it is desired to associate these relatives to a common base the results may be chained to obtained chain index. Thus i its simplest form, the chain indices is one in which the figures for each year are first expressed as percentages of the preceding year. These percentages are then chained together by successive multiplication to form a chain index.

7.2 CHAIN BASE INDEX NUMBERS OR CHAIN INDICES

There are two methods of construction of index numbers depending on the nature of base period employed : (i) Fixed base method and (ii) Chain base method. Most of the index numbers in common use are of the fixed base type, where a fixed period is chosen as base and the index number for any given year is calculated by direct reference to this fixed base period. This fixed base index for any year is no therefore, affected by changes in price or quantity in any other year. It is however considered that the net changes in any given year are the result of gradual changes that have taken place during the past years. This idea is reflected in 'Chain Base Index" numbers.

7.2.1 Steps in constructing a chain index

In constructing a chain index following steps are desirable :

i. Express the figures for each year as percentages of the prceding year. The results so obtained are called link relatives.

ii. Chain together these percentages by successive multiplication to form a chain index. Chain index of any year is the average link relative of that year multiplied by chain index of previous year divided by 100. In the form of formula:Chain index for current year =

$$\frac{\text{Average link relative of current year x chain index of prevous year}}{100}$$

The link relatives = $\frac{\text{Current year price}}{\text{Previous year price}}$ x 100

Obtained in step (i) facilitate comparison from one year to another i.e., between closely situated periods in which the q's are a process of chaining binary comparisons facilitate long-term comparisons.

Chain relatives differ from fixed-base relatives in computation. Chain relatives are computed from link relatives whereas fixed base relatives are computed directly from the original data. The results obtained by the two different methods should be the same but they may differ from each other slightly due to rounding off of decimal places. Since the process of computing chain relatives as the fixed base relatives obtained from the original data, chain relatives is quite complicated and the results are same as should be used when the original data are not available but the link relatives are :

Q. From the following data of the wholesale prices of wheat for ten years construct index numbers taking (a) 2002 as base, and (b) by chain base method:

Year	Price of wheat (Rs. per 10 kg)	Year	Price of wheat (Rs. per 10 kg)
2002	50	2007	78
2003	60	2008	82
2004	62	2009	84
2005	65	2010	88
2006	70	2011	90

Ans. Solution

a) Construction of index number, taking 2002 as base

Year	Price of wheat	Index Number (2002 = 100)	Year	Price of wheat	Index Number (2002 = 100)
2002	50	100	2007	78	$\frac{78}{50}$ x 100 = 156

2003	60	$\frac{60}{50} x 100 = 120$	2008	82	$\frac{82}{50} x 100 = 164$
2004	62	$\frac{62}{50} x 100 = 124$	2009	84	$\frac{84}{50} x 100 = 168$
2005	65	$\frac{65}{50} x 100 = 130$	2010	88	$\frac{88}{50} x 100 = 176$
2006	70	$\frac{70}{50} x 100 = 140$	2011	90	$\frac{90}{50} x 100 = 180$

This means that from 2002 to 2003 there is a 20 percent increase, from 2003 to 2004 there is a 24% increase, from 2004 to 2005 there is 930% increase. If we are interested in finding out increase 2002 to 2003, from 2003 to 2004, from 2004 to 2005; we shall have to compute the chain indices.

b) Construction of Chain Indices

Year	Price of Wheat	Link Relatives	Chain Indices (2002 = 100)
2002	50	100.00	100
2003	60	$\frac{60}{50} x 100 = 120.00$	$\frac{120 x 100}{100} = 120$
2004	62	$\frac{62}{60} x 100 = 103.33$	$\frac{103.33 x 120}{100} = 124$
2005	65	$\frac{65}{62} x 100 = 104.84$	$\frac{104.84 x 124}{100} = 130$
2006	70	$\frac{70}{65} x 100 = 107.69$	$\frac{107.69 x 130}{100} = 140$
2007	78	$\frac{78}{70} x 100 = 111.43$	$\frac{111.43 x 140}{100} = 156$

2008	82	$\frac{82}{78} \text{x} 100 = 105.13$	$\frac{105.13 \text{x} 156}{100} = 164$
2009	84	$\frac{84}{82} \text{x} 100 = 102.44$	$\frac{102.44 \text{x} 164}{100} = 168$
2010	88	$\frac{88}{84} \text{x} 100 = 104.76$	$\frac{104.76 \text{x} 168}{100} = 176$
2011	90	$\frac{90}{88} \text{x} 100 = 102.27$	$\frac{102.27 \text{x} 176}{100} = 180$

The chaiin indices obtained in (b) above with 2002-100 are the same as the fixed based indices obtained in (a) above. In fact chain index figure will always be equal to fixed base index figure if there is only one series.

Q. Compute the chain index number with 2007 prices as base from the following table giving the average wholesale prices of the commodities A, B and C for the year 2007 to 2011 :

Average whole price (in Rs.)

Commodity	2007	2008	2009	2010	2011
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

Solution :

Computation of chain indices

Commodity	Relatives based on preceding year				
	2007	2008	2009	2010	2011
A	100	$\frac{16}{20}$ x 100 = 80	$\frac{28}{16}$ x 100 = 175	$\frac{35}{28}$ x 100 = 125	$\frac{21}{35}$ x 100 = 60
B	100	$\frac{30}{25}$ x 100 = 120	$\frac{24}{30}$ x 100 = 80	$\frac{36}{24}$ x 100 = 150	$\frac{45}{36}$ x 100 = 125
C	100	$\frac{25}{20}$ x 100 = 125	$\frac{30}{25}$ x 100 = 120	$\frac{24}{30}$ x 100 = 80	$\frac{30}{24}$ x 100 = 125
Total of Link Relatives	300	325	375	355	310
Average of Link Relatives	100	108.33	125	118.33	103.33
Chain Index (2007-100)	100 100	$\frac{108.33\text{x}100}{100}$ = 108.33	$=\frac{125\text{x}108.33}{100}$ = 135.41	$=\frac{118.33\text{x}135.41}{100}$ = 160.23	$=\frac{103.33\text{x}160.23}{100}$ = 165.57

7.2.2 Conversion of chain index to fixed index :

At times it may be desired to convert the chain base index numbers into fixed base index numbers. In such a case the following procedure is followed :

1. For the first year the fixed base index will be taken the same as the chain base index. However, if the index numbers are to be constructed by taking first year as the base in that case the index for the first year is taken as 100.

2. For calculating the indices for other years, the following formula is used:

$$\text{Current Year's F.B.I.} = \frac{\text{Current year's C.B.I x Previous Year's F.B.I.}}{100}$$

F.B.I. = Fixed base index number : C.B.I. = Chain base Index No.

Q. From the chain base index numbers given below, prepare fixed base index numbers :

2007	2008	2009	2010	2011
80	110	120	90	140

Sol. Computation of fixed base index numbers

Year	Chain Base Index No.	Fixed Base Index Numbers
2007	80	80
2008	110	$\frac{110 \text{ x } 80}{100} = 88.00$
2009	120	$\frac{120 \text{ x } 88}{100} = 105.60$
2010	90	$\frac{90 \text{ x } 105.60}{100} = 95.04$
2011	140	$\frac{140 \text{ x } 95.04}{100} = 133.06$

7.2.3 Merits of chain base method :

The merits of this method are enumerated here :

1. The chain base method has a great significance in practice because in economic and business data, we are more often concerned with making comparisons with the previous period and not with any distant past. The link relatives obtained by chain base method serve this purpose.

2. Chain base method permits the introduction of new commodities and the detection of old ones without necessitating either the recalculation of entire series or other drastic changes. Because of this flexibility, chain index is used in many types of indices such as the consumer price index and the

wholesale price index.

3. Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.

4. Index numbers calculated by the chain base method are free to a greater extent from seasonal variations than these obtained by the other method.

7.2.4 Limitations of the chain index :

The main limitation of the chain index is that while the percentages of previous year figures give accurate comparisons of year-to-year changes the long-range comparisons of chained percentages are not strictly valid. However, when the index number user wishes to make year-to-year comparison, as is so often done by the businessman, the percentages of the preceding year provide a flexible and useful tool.

7.3 TRUE COST OF LIVING INDEX:

Cost of living index numbers are special purpose index numbers which are designed to measure the relative change in the cost of level for maintaining similar standard of living in two different situations. These are generally intended to represent the average changes in prices overtime, paid by the ultimate consumer for a specified group of goods and services, and hence are also called consumer price index numbers. Generally, the consumption pattern varies with the class of people and the geographical area covered. Hence cost of living index (C.L.I.) numbers must always relate to a specified class of people and a specified geographical area.

The steps in the construction of a cost a living index are as follows :

1. The first step is to decide on the class of people for whom the index number is intended. It is extremely important to define this in clear terms.

2. The next step is to conduct a ‘family budget enquiry’ in the base period relating to the class of people concerned, by the process of random sampling. This would give us information regarding the nature and quality of goods consumed by an average family and also enable determination of weights for computing the index. Only important items among those which are used by the majority

of the class of people are included in the construction of a cost of living index.

3. The items of expenditure are classified in certain major groups, e.g. (i) Food, (ii) Clothing (iii) Fuel and light (iv) Housing (v) Miscellaneous. These major groups are further divided into smaller groups and sub-groups, so that the items are individually mentioned.

4. Arrangements should be made to collect retail prices of the items at regular intervals of time from important local markets. Price quotations are taken at least once a week.

5. For each item there will be a number of price quotations covering different qualities and markets. The simple average of price relatives of the different quotations is taken as the price relative for the particular item.

6. A separate index number is them computed for each group, using Laspeyres, formula in the form of weighted average of price relatives.

$$\text{Group Index (I)} = \frac{\Sigma w\left(\frac{P_m}{P_o} x100\right)}{100}$$

where $w = \frac{P_o q_o}{\Sigma P_o q_o} x100$

Thus, in the construction of a group index, the weight (w) of an item is the percentage expenditure of an 'average family' on that itemex in relation to the total expenditure in the group, as obtained from the family budget enquiry.

7. The weighted average of group index numbers gives the final cost of living index number.

$$\text{Cost of living index} = \frac{\Sigma IW}{100}$$

The weight (w) of the group index is the percentage of total expenditure of an average family spent on that group as shown by the family budget enquriy.

8. Cost of living index numbers are generally constructed for each week. The average of the weekly index numbers is taken as the index number for a month. The average monthly index numbers gives the cost of living index for the whole year.

7.3.1 Uses of cost of living index numbers :

1. Cost of living index numbers are used to determine the purchasing power of money and for computing the real wages (income) from the nomial or money wages (income) we have :

   $$\text{Purchasing Power of Money} = \frac{1}{\text{Cost of living index number}}$$

   $$\text{Real Wages} = \frac{\text{Money Wages}}{\text{Cost of living index}} \text{ x} 100$$

   Thus, cost of living index number enables us to find if the real wages are rising or falling the money wages remaining unchanged.

2. The government and many big industiral and business units use the cost of living index numbers to regulate the dearness allowance (D.A.) for grant of bonus to the employees in order to compensate them for increased cost of living due to price rise. They are used by the government for the formulation of price policy, wage policy and general economic policies.

3. Cost of living indices are used for deflating income and value series in national accounts.

4. Cost of living index numbers are used widely in wage negotations and wage contracts. For example, they are used for automatic adjustments of wages under ‘Escalator clauses’ in collective bargainning agreements. Escalator clause provides for certain point automatic increase in the wages corresponding to a unit increase inthe consumer price index.

Examples :

Construct the cost of living index number from the table given below :

	Group	Index for 1998	Expenditure
1.	Food	550	46%
2.	Clothing	215	10%
3.	Fuel & Lighting	220	7%
4.	House Rent	150	12%
5.	Miscellaneous	275	25%

Solution : Computation of cost of living index

Group	Index (I)	Expenditure (W)	WI
Food	550	46	25300
Clothing	215	10	2150
Fuel & Lighting	220	7	1540
House Rent	150	12	1800
Miscellaneous	275	25	6875
		$\sum w = 100$	$\sum wi = 37665$

Cost of living index number for 1998 $= \frac{\sum wi}{\sum w}$

$$= \frac{37665}{100} = 376.65 \text{ Ans}$$

Q. Calculate the cost of living index number from the following data :

Items	Prices		Weights
	Base Year	Current Year	
Food	30	47	4
Fuel	8	12	1
Clothing	14	18	3
House Rent	22	15	2
Miscellaneous	25	30	1

Solution :

Items	Weight (w)	Prices		Price Relatives $P = \frac{Pn}{Po} x\,100$	WP
		Base year (Po)	Current year Pn		
Food	4	30	47	156.67	626.67
Fuel	1	8	12	150.00	150.00
Clothing	3.	14	18	128.57	385.71
House Rent	2	22	15	68.18	136.36
Miscellaneous	1	25	30	120.00	120.00
	$\sum w = 11$				$\sum wp = 1418.74$

$$\text{Cost of living index number} = \frac{\sum WP}{\sum W}$$

$$= = \frac{1418.74}{11} = 128.98 \text{ Ans}$$

Q. In caluclating a certain cost of living index number the following weights were used food 15, clothing 3, rent 4, fuel and light 2, miscellaneous 1.

Calculate the index for a date when the average percentage increases in prices of items in the various groups over the base period were 32, 54, 47, 48, 48, 78 and 58 respectively.

Suppose a business executive was earning Rs. 2050 in the base period. What should be his salary in the current period if his standard of living is to remain the same ?

Sol. The current index number for each item is obtained on adding 100 to the percentage increase in price.

Calculation for cost of living index number

Group (1)	Average % increase in price (2)	Group Index (I) 100+(2) (3)	Weight (W)	WI
Food	32	132	15	1980
Clothing	54	154	3	462
Fuel & Lighting	47	147	4	588
House Rent	78	178	2	356
Miscellaneous	58	158	1	158
			$\sum w = 25$	$\sum wi = 3544$

Cost of living index number $= \frac{\sum wi}{\sum w}$

$= \frac{3544}{25} = 141.76$ Ans

This implies that if a person was getting Rs. 100 in the base year, then in order to fully compensate the business executive for rise in prices, his salary in the current period should be Rs. 141.76. Hence, if business executive was earning Rs. 2050 in the base period, his salary in the current period should be :

$$\frac{\text{Rs.}141.76}{100} \text{ x } 2050 = \text{Rs. } 2906.08$$

in order to enable him to maintain the same standard of living w.r.t. price rise, other factors remains constant.

7.4 LET US SUM UP

The chain base index number has a great significance in practice because in economic and business data.

The cost of living index or consumer price index numbers are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services.

We have discussed various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. As a mneasure of the formula error a number of mathematical test D, which have been suggested by Fisher, have discussed clearly.

7.5 LESSON END EXERCISE

Q1. What are the chain base index numbers ? How are they constructed ? What are their uses ?

Q2. Distinguish between fixed and chain base indices. Give a suitable illustration to show the difference.

Q3. From the fixed base index numbers given below, find out chain base index numbers :

Year	1996	1997	1998	1999	2000	2001
Index No.	200	220	240	250	280	300

Ans. 200, 220, 240, 250, 280, 300

Q4. What is a cost of living index number ? Discuss briefly its uses and limitation ?

Q5. Find the cost of living index for the following data .

Group	Group Index	Weight
Food	180	140
Clothing	150	420
Rent	100	49
Fuel & Lighting	110	56
Miscellaneous	80	63

Ans. 136

****

M.A. Economics Semester - IInd Lesson - 8
Course No. 203 Unit - II

NATURE AND DECOMPOSITION OF A TIME SERIES, ANALYSIS OF TREND : POLYNOMIAL TREND, NON-LINERA GROWTH CURVES

CHAPTER HIGHLIGHTS:

This chapter contain the information about the nature and components of time series. Apart from this, the trend analysis is discussed in a deep manner.

CHAPTER OUTLINE:

8.1 INTRODUCTION:

One of the most important tasks before economists and businessman these days is to make estimates for the future. For example, a businessman is interested in finding out his likely sales in the year 2013 or as a long-term planning in 2020 or the year 2030 so that he could adjust his product accordingly and avoid the possibility of either unsold stocks or inadequate production to meet the demand. Similarly, an economist is interested in estimating the likely population in the coming year so that proper planning can be carried out with regard to food supply, jobs for the people etc. However, the first step in making estimates for the future consits of gathering information from the past. In this connection one usually deals with statistical data which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as 'time series'. Thus when we observe numerical data at different point of time the set of observations is known as time series.

8.2 NATURE AND DECOMPOSITION OF TIME SERIES :

Definition of time series are given below :

In the words of Morris and Hamburg, "A time series is a set of statistical observations arranged in chronological order."

According to Levin and Rubin, "Time series analysis is used to detect patterns of change is statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future."

It is clear from the above definitions that time series consist of data arranged chronologically. Thus if we record the data relating to population, per capita income, prices, production etc... for the last 5, 10, 15, 20 years or some other time period, the series so emerging would be called time series.

Mathematically, a time series is defined by the functional relationship

$$y = f(t)$$

Where y is the value of the phenomenon (or variable) under consideration at time t, e.g.

i) The population (y) of a country or a place in different year (t)

ii) The number of births and deaths (y) in different methods (t) of the year,

iii) The sale (y) of the departmental store in different months (t) of the year,

iv) The temperature (y) of a place on different days (t) of the week,

and so on constitute time series. Thus,k if the values of a phenomenon or variable at times t_1, t_2,....t_n are y_1, y_2....y_n respectively, then the series.

t : t_1 t_2 t_3............t_n

y : y_1 y_2 y_3........... y_n

Constitutes a time series. Thus, a time series invariable gives a bivariate distribution, one of the two variables being time (t) and the other being the value (y) of the phenomenon at different points of time. The values of t may be given yearly, monthly, weekly, daily or even hourly, usually but not always at equal intervals of time.

Decomposition of Time Series

It is customary to classify the fluctuations of a time series into four basic types of variations which super imposed and acting all in concert amount for changes in the series over a period of time. These four types of patterns, movements, or, as they are often called, components or elements of a time series are :

i. Secular trend (t)

ii. Seasonal variatons (s)

iii. Cyclical variations (c)

iv. Irregular variations or random movement (i)

In the classical or traditional approach, it is assumed that there is a multiplicative relationshp between the four components i.e. any particular observations is considered to be the product of the effects of four components.

$y_t = t \text{ x } s \text{ x } c \text{ x } i$ (Multiplicative model)

Another approach is to assume an additive relationship b/w them

$y_t = t + s + c + i$ (Additive method)

Although the additive model facilitates easier calculations, it has been found inappropriate in many practical situations, and hence is not generally used.

8.2.1 Secular Trend :

Secular trend of time series is the smooth, regular and long-term movement exhibiting the tendency of growth or decline over a period of time. The trend is that part which the series would have exhibited, had there been no other factors affecting the values.

8.2.2 Seasonal Variations :

Seasonal variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself. Since these variations repeat during a period of 12 months they can be predicted fairly accurately. Nearly every type of business activitiy is susceptible to seasonal influence to a greater or lesser degree and as such these variations are regarded as normal phenomenon recurring every year. Although the world 'seasonal' seems to imply a connection with the season of the year, the term is meant to include any kind of variation which is of periodic nature and whose repeating cycles are of relatively short duration.

8.2.3 Cyclical Variations :

The term ‘cycle refers to the recurrent variations in time series that usually last longer than a year and it can be as many as 15 or 20 years. These variations are regular neither in amplitude nor in length. Most of the time series relating to economics and business show some kind of cylical or oscillatory variation. Cyclical fluctuations are long term movements that represent consistently recurring rises and declines in activity.

The study of cyclical variations in extremely useful in framing suitable policies for stabilizing the level of economic activitiy i.e. for avoiding periods of booms and depressions as both are bad for an economy - particularly depression which brings about a complete disaster and shatters the economy.

8.2.4 Irregular Variations :

Irregular variations also called ‘erratic accidental, random, refer to such variatons in business activity which do not repeat in a definite pattern. In fact the category labelled irregular variation is really intended to include all types of variations other than those accounting for the trend, seasonal and cyclical movements.

There are two reasons for recognizing irregular movements :

i. To suggest that on occasions it may be possible to explain certain movements in the data due to specific causes and to simplify further analysis.

ii. To emphasize the fact that predictions of economic conditions are always subject to degree of error owing to the unpredictable erratic influences which may enter.

8.3 ANALYSIS OF TREND :

Polynomial trend, non-linear growth curves.

In many time series, broad movements can be discerned which evolve more gradually than the other motions which are evident. These gradual changes are described as trends and cycles. The changes which are of a transitory nature are described as fluctuations.

In some cases, the trend should be regarded as nothing more than the accumulated efffect of the fluctuations. In other cases, we feel that the trends and the fluctuations represents different sorts of influences.

There are essentially two ways of extracting trends from a time series. The first way is to apply to the series a variety of so called filters which annihilate or nullify all of the components which are not regarded as trends.

A filter is a carefully crafted moving average which spans a number of data points and which attributes a weight to each of them. The weights should sun to unity to ensure that the filter does not systematically inflate or deflate the values of the series. Thus, for example, the following moving average might serve to eliminate the annual cycle from an economic series which is recorded at quartely intervals.

$$yt = \frac{1}{16}\left(y_{t+3} + 2y_{t+2} + 3y_{t+1} + 4_{y_t} + 3y_{t-1} + 2y_{t-2} + y_{t-3}\right)$$

The process of filtering is often a good way of deriving an index which represents the more important historical characteristics of the time series.

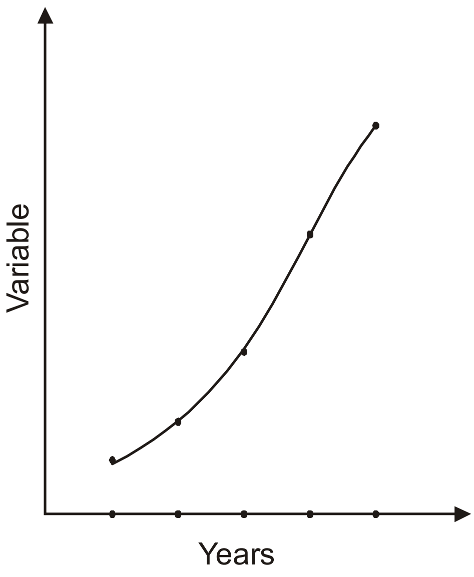
The alternative way of extracting the trend from the index is to fit some function which is capable of adapting itself to whatever from the trend happens to display. Different functions are appropriate to different forms of trend, and some functions which analysts trend to favour see almost always to be inappropriate. Once an analytic function has been fitted to the series, it may be used to provide extrapolative forecasts of the trend.

“Trends also called secular or long-term trend, is the basic tendency of a series... to grow or decline over a period of time. To concept of trend does not include short-range oscillations but rather the steady movement over a long time. This phenomenon is usually observed in most of the series relating to economics and business e.g. an upward tendency is usually observed in time series relating to population, production, sales etc.

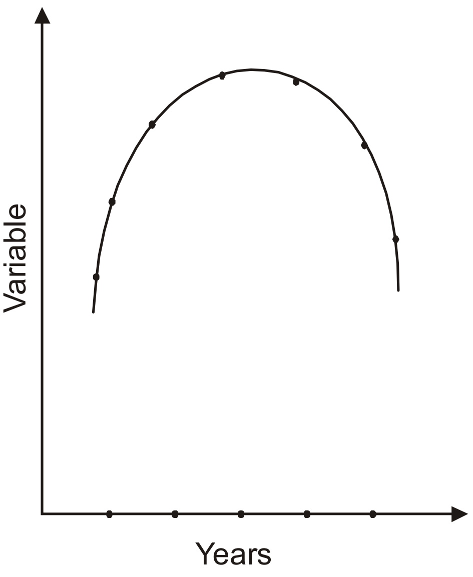
i. It should be clearly understood that trend is the general, smooth, long-term average tendency. It is not necessary that the increase of decline should be in the same direction throughout the given period. It may be possible that different tendencies of increase, decrease or stability are observed in different sections of time.

ii. The term 'long period of time' is a relative term and cannot be defined exactly. It would be much depend on the nature of the data.

iii. Linear and non-linear (Curvi-linear) trend : If the time series values plotted on graph cluster more or less round a straight line, the trend exhibited by the time series is termed as linear otherwise non-linear. In a straight line trend, the time series values increase or decrease more or less by a constant absolute amount i.e. the rate of growth is constant.



Linear Trend



Non-Linear Trend

iv. It is not necessary that all the series must exhibit a rising or declining trend. Certain phenomena may give rise to time series whose values fluctuate round a constant reading which does not change with time, e.g. the series relating to temperature of barometric readings (pressure) of a particular place.

8.4 USES OF TREND :

i. The study of the data over a long period of time enables us to have a general idea about the pattern of the behaviour of the phenomenon under consideration. This helps in business forecasting and planning future operations e.g., if the time series data for a particular phenomenon exhibits a trend in a particular direction then under the assumption that the same pattern.

ii. By isolating trend values from the given time series, we can study the short-term and irregular movements.

iii. Trend analysis enables us to compare two or more time series over different periods of time and draw important conclusions about them.

8.5 MEASUREMENT OF TREND :

There are four methods which are geneally used for the study and measurement of the trend component in a time series.

i. Graphic (or free-hand curve fitting) method.

ii. Method of semi-averages

iii. Method of curve fitting by the principle of least squares.

iv. Method of moving average.

8.5.1 Free-hand Method :

In free-hand method, the given data are plotted as points as a graph paper against time. The time series data (yt) are shown along the vertical axis and time (t) along the horizontal axis. Then a smooth free hand curve is drawn through the scatter of the plotted points, which appears to represent their pattern of movement over time. The distance of this line, known as 'trend line' gives the trend values for each time period. The advantanges of this method are that a quick estimate of the trend is obtained and that the method can be used to obtain a preliminary knowledge of the nature of trend with a view to applying more refined methods. However, the free hand method depends too much an individual judgement and different persons will obtain different trend values from the same data.

8.5.2 Semi-average Method :

Semi-average method consists in dividing the data into two parts, and then finding an average for each parts. These averages are plotted as points on a graph papers, against the mid-point of the time interval covered by each part. The straight line joining these two points gives the trend line. As before, the distances of trend line from the horizontgal axis give the trend values. If the actual trend is a straight line, the method will give quite satisfactory results. Although this method is simple to apply, it may lead to poor results when used indiscriminately. If the ratio of successive time series data are approximately equal, the method should be applied to the logrithms of values. The trend values will then be obtained by taking antilogarithims of the distances of trend line from the horizintal axis.

Example :With the help of graph under supplied, obtain the trend values.

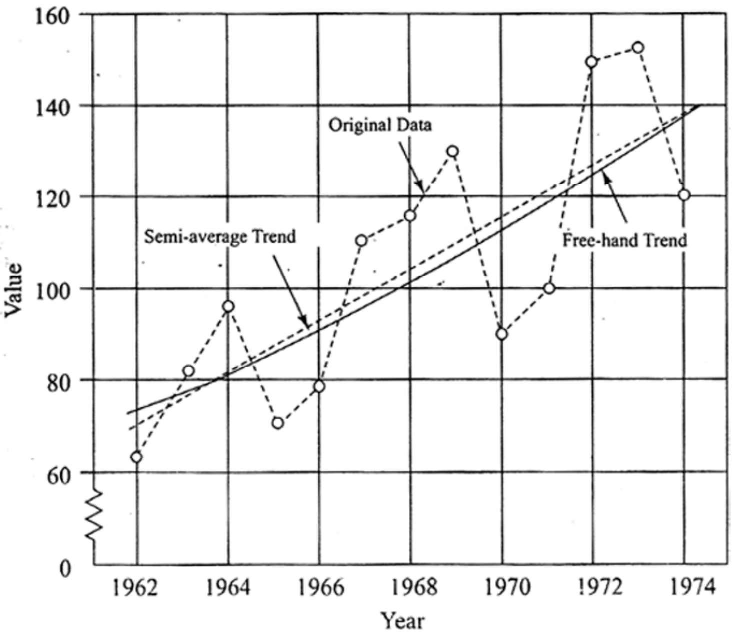
Year	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
Original data	64	82	97	71	78	112	115	131	88	100	146	150	120

Solution :

The given data are plotted as points on the graph paper (Fig) and a free-hand drawn throgh the scatter of the points. This is the 'trend line'. The distances of trend from the horizontal axis are now read from the graph for all the years, giving 'trend values'

Year	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
Value data	75	79	82	86	91	96	102	107	112	118	124	130	136
Original data	64	82	97	71	78	112	115	131	88	100	146	150	120

Note: (i) The original data are fluctating up and down (see the zig-zag line in Fig) The trend values show gradual changes over the years. The trend values from a free-hand curve, which is drawn by one's individual judgement, not agree with those obtained by others or by applying other methods.



8.5.3 Moving Average Method :

Moving average method is very commonly used for the isolation of trend and in smotthing out fluctuations in time series. In this method, a series of arithmetic means of successive observations, known as moving averages, are calculated from the given data, and these moving averages are used as trend values.

The logic behind the moving average method is that if the period of moving average is exactly equal to the period of cycle present in the series, then the method will completely eliminate cyclical fluctuations. The method is very simple and needs no complicated mathematical calculations. It cannot be used for forecasting future trend.

Example :Calculate the five-yearly moving average of the following :

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958
Original data	105	115	100	90	80	95	85	75	60
Year	1959	1960	1961	1962	1963	1964	1965	1966	
Original data	65	70	58	55	53	60	52	50	

B) Illustrate with four examples, how the four-yearly Moving Average can be calculated

Table : Calculation for 5 yearly Moving Average

Year	Value	5-Year Moving Total	5-Year Moving Average
1950	105	-	-
1951	115	-	-
1952	100	490	98
1953	90	480	96
1954	80	450	90
1955	95	425	85
1956	85	395	79
1957	75	380	76
1958	60	355	71
1959	65	328	65.6
1960	70	308	61.6
1961	58	301	60.2
1962	55	296	59.2
1963	53	278	55.6
1964	60	270	54
1965	52	-	-
1966	50	-	-

Note :

(i) Calculations for Moving Total in Col. (3) may be simplified as follows. Direct calculations give the sum of the first 5 values (105 + 115 + 100 + 90 + 80) = 490. The next moving total (115 + 100 + 90 + 80 + 95) differs from the former with the new value 95 replacing the Ist value 105, i.e., an increase of 95-105 = -10; so the new moving total is 490-10=480. Sikilarly, the next moving tota is 480 + (85 - 115) = 480 - 30 = 450; and so on. The increase or decrease over the preceding moving total may be mentally calculated and the new \moving toal obtained easily. The accuracy of calculations can be checked by verifying that the last moving total obtained by this process, ivz. 278 + (50-58) = 270, is exactly equal to the total of the last five values (55 + 53 + 60 + 52 + 50), by direct calculations.

ii) Col. (4) = Col. (3), 5]

8.5.4 Methods of Mathematical Fitting Curves or Least Squares :

Methods of fitting mathematical curves is perhaps the best and most objective method of determining trend. In this method, an appropriate type of mathematical equations is selected for trend and the constants appearing in the trend equation are determined on the basis of the given time series data. The choice of the appropriate type of equation is facilitated by graphical representation of the data.

i. If the plotted data show approximately a straight line tendency on an ordinary graph paper, the equation used is :

$y = a + bx$... (straight line)

ii. If they show a straight line on a semi-logarithmic graph paper, the equation used is

$\log y = a + bx$.... (exponential curve)

iii. Sometimes a parabola or higher order polynomial may also be fitted

$y = a + bx + cx^2$.... (parabola)

iv. Special types of curves are also used in certain cases

a) $y = a + bc^x$......(Modified exponential curve)

b) $1/y = a + bc^x$......(Logistic curve)

c) $\log y = a + bc^x$......(Gompertz curve)

The constants, appearing in the equations referred to at (a) to (c) above are obtained by applying the principle of least squares. This states that the values of contants should be such as to make the sum of the squares of vertical distances from the trend lines as small as possible. The method of fitting mathematical curves can be used for forecasting the future trend. This method also involves considerable numerical calculations.

8.5.5 Polynomial Trend :

A type of trend that represents a large set of data with many fluctuations. As more data become, available, trends often become less linear and a polynomial trend takes its place. Graphs with curved trendlines are generally used to show a polynomial trend.

For example, polynomial trending would be apparent on the graph that shows the relationship between the profit of a new product and the number of years the product has been available. The trend would likely rise near the beginning of the graph, peak in the middle and then trend downward near the end. If the compny revamps the product late in its life cycle we'd expect to see this trend repeat itself. This type of chart, which would have several waves on the graph, would be deemed to be a polynomial trend.

8.5.6 Second Degree Parabola :

The simplest example of the non-linear trend is the second degree parabola, the equation of which is written in the form :

$Y_c = a + b\,x + cx^2$

where c is still the y intercept, b is the slope of the curve at the origin and c is the rate of change in the slope. When numerical values for a, b and c have been derived, the trend values for any year may be computed by substituting in the equation the value of x for that year. The values of a,b and c can be determined by solving the following three normal equations simulteaneously :

$$\sum y = Na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Note that the first equaton is merely the summation of the given function, the second is summation of x multiplied into the given function and the third is the summation of x^2 multiplied into the given function.

8.5.7 Second Degree Curve, Fitted to Logarithms :

We may come across data which when plotted an semi-logarithmic paper continue to show curvature, being concave either upward or downward or, in other words, the ratio of change may be either increasing or decreasing. In such cases, we may fit second-degree curve to the logarithms of the y values using

$$\log y = \log a + x \log b + x^2 \log c$$

Taking the x origin at the middle of the period, the three normal equations are :

i. $\sum \log y = N\log a + \log c \sum x^2$

ii. $\sum(x.\log y) = \log b \sum x^2$

iii. $\sum(x^2.\log y) = \log a \sum x^2 + \log c \sum x^4$

It may be noted that ordinarily polynomials of third or higher degree are not fitted to time series in either arithmetic or logarithmic form.

8.5.8 Measuring Trends by Logarithms :

The trends discussed so far were plotted on arithmetic scales. Trends may also be plotted on a semi-logarithmic chart in the form of a straight line or a non-linea curve. A straight line on the semi-log chart shows the increase of y values of a time series at a consant rate. When a given time series is increasing or decreasing at a constant rate and is not approaching some upper limit, it is usually approximated least by an exponential curve. When it is a non-linear curve on the semi-log chart an upward curve shows the increase at varying rates, depending on the shapes of the slopes the steeper the slope, the higher is the rate of increase.

The types of trend usually computed by logarithims are :

i) Exponential trends

ii) Growth cruve

8.5.9 Exponential trends :

The equation of the exponential curve is of the form

$$y = ab^x$$

8.5.10 Growth Curve :

In economic data very often we come across situations in which in the beginning the growth is very slow, but as the product is accepted the demand increased by a greater amount each year and finally as the market becomes more and more fully developed, the amount of gropwth each year becomes less. The curve continues to grow more and more slowly approaching an upper limit but not reachng it. Such series are best represented by growth curves. The growth curves do not reach a maximum and turn down in the manner of the second degree parabola.

A number of different growth curves have been used to measure secular trend, but the curves used most widely to describe growth are the Gompertz curve and the pearl-reed or logistic curve - these curves are s-shaped for increasing series when plotted on graph paper with an arithmetic vertical scale and are concave downward on a semi-logarithmic chart.

Both growth curves approach a finite limit and this should necessarily be taken into account when fitting a given time series to one of the curves.

Time series analysis using growth curves analyses or depicts the manner in which or the rate at which the series approaches its limit. Thus, whenever a given time series is increasing at a constant rate, but is understood to be approaching a finite limit n a predictable manner, growth curves may be appropirate for assessing the secular trends components of the series.

8.6 LET US SUM UP:

A time series s an arrangement of statistical data in a chronological order i.e. in accordance with in time of occurence. It reflects the dynamic pace of movements of the phenomena over a period of time.

In economics, it is tradtional to decompose time series into a variety of components, some of all of which may be present in a particular instance. If yt is the sequence of values of an economic index, then its generic element is liable to be expressed as

$$Yt = Tt + Ct + st + et$$

Tt is the global trend

Ct is a secular cycle

St is the seasonal variation and

Et is an irregular component

There are two distinct purposes for which we might wish to effect such as decomposition. The first purpose is to give a summary description of the salient features of the time series. The other purpose in decomposing the series is to predict its future values.

The study of secular trends allows us to describe historical patterns. The study of secular trends permits us to project past patterns or trend into the future.

8.7 LESSON END EXERCISE :

Q1. What is a time series. Give illustrations for each of them.

Q2. Explain briefly the additive and multiplicative models of time series.

Q3. Discuss the nature and decomposition of a time series.

Q4. What are secular trend, seasonal variations and cyclical fluctuations.

Q5. With the help of graph paper, obtain the trend curve (graphic method)

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990
Value	64	82	97	71	78	112	115	131	88
Year	1991	1992	1993	1994					
Value	100	146	150	120					

Q6. Fit a straight line trend to the following data using the method of least squares and calculate the production for the year 2001 :

Year	1996	1997	1998	1999	2000
Production (000 tons)	83	92	74	90	166

Q7. Define exponential trend and growth curves.

****

M.A. Economics **Semester - IInd** **Lesson - 9**

Course No. 203 **Unit - II**

MOVING AVERAGE METHOD, SEASONAL COMPONENT CYCLICAL AND RANDOM COMPONENT FORECASTS AND THEIR ACCURACY

CHAPTER HIGHLIGHTS:

The present lesson throws light on the various component of time series and its various methods.

CHAPTER OUTLINES :

9.1 Introduction

9.2 Methods of moving averages

9.2.1 Merits and Demerits

Solved example

9.3 Seasonal component

9.4 Measurement of seasonal index

9.4.1 Method of averages

9.4.2 Moving average method

9.4.3 Trend ratio method

9.4.4 Link relative method

9.5 Disconsolation of data

Solved example

9.6 Cyclical component

9.1 INTRODUCTION :

When a trend is to be determined by the method of moving averages, the average value for a number of years (or month or weeks) 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 averages. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

While applying this method, it is necessary to select a period for moving average such as 3 yearly moving average, 5 yearly moving average, 8 yearly moving average etc. The period of moving average is to be decided in the light of the length of the cycle since the moving average method is mot commonly applied to data which are characterized by cyclical movements, it is necessary to select a period for moving average which concedes with the length of the cycle, otherwise the cycle will not be entirely removed. The danger is more severe, the shorter the time period represented by the average. When the period of moving average and the period of the cycle do not coincide, the moving average will display a cycle which has the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often we find that the cycle in the data are not of uniform length. In such a case we should take a moving average period equal to or some what greater than the average period of the cycle in the data. Ordinarily, the necessary period will range between three and ten years for general business series but even longer periods are required for certain types of data.

The 3-yearly moving average shall be computed as follows :

$$\frac{a+b+c}{3},\frac{b+c+d}{3},\frac{c+d+e}{3},.....$$

and for 5-yearly moving average

$$\frac{a+b+c+d+e}{5},\frac{b+c+d+e+f}{5},\frac{c+d+e+f+g}{5},.....$$

9.2 METHOD OF MOVING AVERAGES :

Method of moving averages is a very convenient tool for ironing out fluctuations in time series. It is generally used in all parts of time series analysis for the isolation of trend, and also in connection with seasonal, cyclical and irregular components - by eliminating the moving averages considerably reduce their intensity. This is called 'smoothing'. Moving averages are a sequence of arithmetic means, each based on a fixed number of successive observations in time series. Suppose, the time series data are given by years. Then moving averages of period 3 years, ay (also called 3-year moving averages) give us a series of arithmetic mean each of three successive observations. We start with the first 3 observations and calculate the arithmetic mean. This is placed against the 2nd year. At the next stage, we leave the first observation and calculate the A.M. of the next three, viz. 2nd to 4th and place it against the 3rd year. The process is repeated until we arrive at the last three observations. Similarly, moving averages of 4, 5, 6 or any period may be calculated.

Sometimes, weighted moving averages are also used. For the calculation of weighted moving average, the successive observations in a group are multiplied by a given set of numbers (called weight) and the weighted sum is divided by the sum of weights. The same set of weights is used for the successive observations in each group.

Each moving average is placed against the mid-point of time interval included in the calculation of A.M. Thus, when the period is odd, all moving averages will coincide with the given years. However, if the period is even, moving averages will fall mid-way between two successive years. In this case, again two-item moving averages of the moving averages already found, have to be calculated by syncronizing them with the given data. This process is known as 'centering'.

Moving averages with a period exactly equal to (or a multiple of) the period of cycle present in the series will completely eliminate the cyclical component from a time series and give an estimate of trend. So, the moving average method is used for measurement of trend from a given time series data, by taking the period of moving average exactly equal to the period of cycle present in the series. However, usually the period and intensity vary from cycle to cycle. The best results will be obtained by using moving averages with a period equal to the average period of fluctuations which can be obtained by graphical method. In all cases, however, moving averages will reduce fluctuations present in the series and smooth out the short term seasonal and irregular movements, even if the cyclical component is not removed altogether.

9.2.1 Merits and Demerits of Moving Average Method :

Merits

i. Moving average method is simple to apply and involves no difficult calculations.

ii. If the time series contain regular cyclical fluctuations, these fluctuations are not completely eliminated, moving average process reduces their intensity.

iii. Moving averages adapt themselves to the general movements of data. The shape of the trend line is thus determined by the data themselves rather than by the statistican's choice of mathematical curve.

iv. Moving average method is flexible in the sense that if some more observations are added to the original series, the entire calculations need not be changed we get only some more trend values.

Demerits

i. Trend values for all the given periods cannot be obtained. Some trend values at the beginning and at the end of the series have to be left out, their number increasing with increase in the period of moving average.

ii. The period of moving average has to be chosen very carefully. There are no hard fast rules for the purpose.

iii. Since moving averages assume no law of change, the method cannot be used for forecasting future trend.

iv. Moving averages are useful only when the trend is linear so if the actual is curvilinear, moving average may elevate considerably from the trend.

Illustration : Calculate the 3-yearly moving averages of the production figure given below and draw the trend.

Year	**Production (M. Tonnes)**	**Year**	**Production (M. Tonnes)**
1997	15	2005	63
1998	21	2006	70
1999	30	2007	74
2000	36	2008	82
2001	42	2009	90
2002	46	2010	95
2003	50	2011	102
2004	56		

Solution :

Calculation of 3-yearly moving averages

Year	**Production (M. Tonnes)**	**3 yearly total (M. tonnes)**	**3-yearly moving average**
1997	15	-	-
1998	21	66	22.00
1999	30	87	29.00
2000	36	108	36.00

2001	42	124	41.33
2002	46	138	46.00
2003	50	152	50.67
2004	56	169	56.33
2005	63	189	63.00
2006	70	207	69.00
2007	74	226	75.33
2008	82	246	82.00
2009	90	267	89.00
2010	95	287	95.67
2011	102	-	-

9.3 SEASONAL - COMPONENT

Seasonal fluctuations in time series refer to a type of periodic movement, where the period does not exceed one year. Variations in passenger traffic during the 24 hours of a day, number of books issued from a library during the seven days of a week, and values of cheque clearance during the 12 month of a year, are examples of seasonal fluctuations. In these cases, the periods are respectively 1 day, 1 week and 1 year. Most of the business and economic activities are found to have brisk and slack periods during some specific parts of a year, and these fluctuations are found to repeat with striking regularity year after year. Such up-and-down movements are the results of seasonal fluctuations. Climate changes of seasons and the customs and habits of people at different parts of the year are the main factors responsible for seasonal fluctuations.

Uses of seasonal index in time series analysis :

There are two major uses of seasonal indices :

1. Adjusting time series data for seasonal variation.
2. Forecasting on monthly or quarterly basis.

Seasonal indices give us a clear idea about the relative position of each month on quarter in time series data relation to such matters as sales, production, employment etc. The indices may be used to 'deseasonalise' (i.e. eliminate the seasonal effects of) the series - (i) by expressing the original monthly or quarterly data as percentages of the corresponding seasonal indices, when multiplicative model is used or (ii) by subtracting the seasonal component from original figures, when additive model is used. Deseasonalisation is often necessary for studying the trend or cyclical movements. Moreover, in combining or comparing data that have differing seasonal factors, it is first necessary to eliminate the effects due to seasonal variations.

Seasonal indices may be used for short-term forecasting which is so necessary for planning the future course of action. For example, in studying the production figures of a company over times, seasonal indices may be used to plan for hiring of personnel for peak periods, to accumulate an inventory of raw materials, to ready equipment and to allocate vacation time. Similarly in order to meet the possible demands of customers during the winter, department store may utilise seasonal indices of their sales data for optimum allocation of capital in the warm clothings department.

9.4 MEASUREMENT OF SEASONAL VARIATION :

There are four methods of measuring seasonal fluctuations :

i. Method of averages

ii. Moving average method

iii. Trend ratio method

iv. Link relative method

Example :

Compute the average seasonal movements by the method of quarterly total (average) for the following series of observations :

Total production of paper (thousand tons) Quarters				
Year	I	II	III	IV
1951	37	38	37	40
1952	41	34	25	31
1953	35	37	35	41

Solution :

The calculations are shown below using an additive model :

Calculations for average seasonal movement

Year/Quarter	I	II	III	IV	Total
1951	37	38	37	40	-
1952	41	34	25	31	-
1953	35	37	35	41	-
Total	113	109	97	112	431
A.M.	37.67	36.33	32.33	37.33	143.66
Average Seasonal Movement	1.75	0.42	-3.59	1.42	0
Grand Average = 143.66 ÷ 4 = 35.92					

[Note : Average seasonal movement have been obtained by subtracting the Grand Average from A.M. for each quarter. Slight adjustments are usually necessary to make the total seasonal movement zero. Thus

37.67 - 35.95 = 1.75 36.33 - 35.95 = 0.41

33.33 - 35.92 = - 3.59 37.33 - 35.92 = 1.41

Arbitarily 0.41 and 1.41 have been changed to 0.42 and 1.42 to make the total seasonal movemnet zero]

Ans. 1.75, 0.42, -3.59, 1.42 ('000 tons) for I, II, III, IV.

9.4.1 Method of Averages

This method is applied, when the given time series date do not contain trend or cyclical fluctuations to any appreciable extent. From the quarterly data the totals for each quarter and averages A_1, A_2, A_3, A_4 for the A quarters Q_1, Q_2, Q_3, Q_4 are found. The grand average

$$G = \frac{1}{A}(A_1 + A_2 + A_3 + A_4)$$

is also calculated. If the additive model is used, the deviations of quarterly averages from the grand average give seasonal variations.

$$S_1 = A_1\text{-}G,\ S_2 = A_2 - G,\ S_3 = A_3\text{-}G,\ SA = AA\text{-}G$$

if monthly figures are given, we find 12 averages A_1, A_2,....A_{12} for the month January, February,......... December respectively and then proceeding the same way as before, the seasonal index for each month is obtained. The total (or average) seasonal variation is 0, and the average season index is 100.

9.4.2 Moving Averages Method :

From the given quarterly figures the trend is estimated by taking four quarter moving averages. The effect of trend is then estimated from the original data. If the additive model is used, the moving average trend values are subtracted from the original data to give us 'derivations from trend'. Since these deviations do not contain any effect of trend, the method of quarterly averages is applied to these deviations, using additive model.

If the multiplicative model is taken, then we find 'ratios to moving averages', expressed as percentages i.e. the original values are expressed as percentage of the corresponding moving averages values. These percentages are now arranged by quarters and the average for each quarter P_1, P_2, P_3, P_4 (suppose) are found out. These are adjusted to a total of 400, multiplicative each by 100/p, where

$$P = \frac{1}{4}(P_1 + P_2 + P_3 + P_4)$$

is the grand average. The seasonal indices are

$$S_1 = \frac{P_1}{P} x\,100,\ S_2 = \frac{P_2}{P} x\,100,\ S_3 = \frac{P_3}{P} x\,100,\ S_4 = \frac{P_4}{P} x\,100$$

respectively for the quarters Q_1, Q_2, Q_3 and Q_4

Example :

Obtain seasonal fluctuation from the following time series data :

Quarterly output of coal for 4 years

Year/Quarters	I	II	III	IV
1928	65	58	56	61
1929	68	63	63	67
1930	70	59	56	52
1931	60	55	51	58

Solution :

We apply the method of moving averages. For this purpose, first find four-quarter moving averages, giving trend. Since col. (6) of Table 16.22 shows figures free from trend, the method of quarterly averages is then applied

Moving Averages and Deviations from Trend

Year/ Quarter		Output	4-quarter moving total	2-period moving total of cost (3)	4 quarter moving average	Deviation from trend
1928	I	65	-	-	-	-
	II	58	-	-	-	-
	III	56	240	-	-	-
	IV	61	243	491	61.38	- 0.38
1929	I	68	255	503	62.88	5.12
	II	63	261	516	64.50	- 1.50

	III	63	263	524	65.50	-2.50
	IV	67	259	522	65.25	1.75
1930	I	70	252	511	63.88	6.12
	II	59	237	489	61.12	- 2.12
	III	56	227	464	58.00	- 2.00
	IV	52	223	450	56.25	- 4.25
1931	I	60	218	441	55.12	4.88
	II	55	224	442	55.25	-0.25
	III	51	-	-	-	-
	IV	58	-	-	-	-

Table : Calculations for Seasonal Fluctuations

Year	Deviations from Trend				Total
Quarter	I	II	III	IV	
1928	-	-	-4.38	-0.38	-
1929	5.12	-1.50	-2.50	1.75	-
1930	6.12	-2.12	-2.00	-4.25	-
1931	4.88	-0.25	-	-	
Total	16.12	-3.87	-8.88	-2.88	0.49
Average	5.37	-1.29	-2.96	-0.96	0.16
Adjustment	-0.04	-0.04	-0.04	-0.04	-0.16
Seasonal	5.33	-1.33	-3.00	-1.00	0

Col. (5) = Col. (4) ÷ 8

Col. (6) = Col. (2) minus Col. (5)

= 56 - 60 = 38 = -4.38

9.4.3 Trend Ratio Method :

In this method, the multiplicative model is always taken. Trend values are obtained by fitting a mathematical curve and the original data are expressed as percentages of the corresponding trend. As in the moving average method, these percentages are arranged by quarters and the average trend ratio for each quarters and the average trend - ratio for each quarter viz. P_1, P_2, P_3, P_4 are found out. Each of these is now multiplied by 100/p, to give the seasonal indices.

$$S_1 = \frac{P_1}{P} \text{ x } 100, \; S_2 = \frac{P_2}{P} \text{ x } 100, \; S_3 = \frac{P_3}{P} \text{ x } 100, \; S_4 = \frac{P_4}{P} \text{ x } 100$$

corresponding to the quarters Q_1, Q_2, Q_3, Q_4 respectively where

$$P = \frac{1}{4}(P_1 + P_2 + P_3 + P_4)$$

Example :

The number of traffic accidents in Calcutta in four quarters of a year during the period 1977-79 are given below :

	Quarters			
Year	I	II	III	IV
1977	165	135	140	180
1978	152	121	127	163
1979	140	100	105	158

Find seasonal indices by Trend-ratio method, assuming a linear trend for the data.

Solution :

First, we have to find the trend values. Let y = a + bx be the equation of trend (origin : mid-point of quarters II and III, 1978; unit of $x = \frac{1}{2}$ quarter). By the method of least squares, the values of a and b are obtained from the equations $\sum y = an + b \sum x$ and $\sum xy = a \sum x + b \sum x^2$

Fitting Linear Trend to Quarterly Data

Year		Time series				Trend Values	Trend-Ratio
Quarter		(y)	x	x^2	xy	(T)	(y/T) x 100
1977	I	165	-11	121	-1815	155.7	106
	II	135	-9	81	-1215	152.9	88
	III	140	-7	49	-980	150.2	93
	IV	180	-5	25	-900	147.4	122
1978	I	152	-3	9	-456	144.6	105
	II	121	-1	1	-121	141.9	85
	III	127	1	1	127	139.1	91
	IV	163	3	9	489	136.4	120
1979	I	140	5	25	700	133.6	105
	II	100	7	49	700	130.8	76
	III	105	9	81	945	128.1	82
	IV	158	11	121	1739	125.3	126
Total		1686	0	572	-788	-	-

Substituting the value from the table in the normal equations,

$a = 1686/12 = 14.5,$

$b = - 788/572 = - 1.38$

The trend equation is therefore y = 140.5 - 1.38x (origin: mid-point of quarters II and III, 1978; unit of $x = \frac{1}{2}$ quarter). Putting appropriate values of x, we get the trend values.

The next step is to express the original data as percentages of trend, giving "trend-ratios". The trend ratios are arranged by quarters as in Table and the seasonal index is calculated by the methods shown in example

Calculations for seasonal index

Year	I	II	III	IV	Total
1977	106	88	93	122	
1978	105	85	91	120	
1979	105	76	82	126	
Total	316	249	266	368	-
Average	105	83	89	12	400
Seasonal Index	105	83	89	123	400

The seasonal indices are 105, 83, 89, 123 for quartes I and IV respectively.

9.4.4 Link Relative Method :

If quarterly data are given, each value is expressed as percentage of the value for the immediately preceding period. There are known as Link Relatives of course, the link relative for the first quarter (Q1) of the first year cannot be obtained the L.R.s are arranged by quarters and the average L.R. for each quarter is found either by using the arithmetic mean or median. The average link relatives show the average relation of each quarterly value to the value of previous quarter. From these average L.R.s we find chain relatives by relating them to a common base e.g. the first quarter, for which C.R. is taken as 100. The C.R. for and quarter is not obtained on multiplying the L.R. for that quarter by the C.R. for the immediately preceding quarter and dividing by 100. Preceeding this way we find a second C.R. for the first quarter (Q_1) by the relation second C.R. for Q1 by the relation. Second C.R. for Q1

$$= \frac{(\text{C.R. for } Q_4) \text{ x } (\text{L.R. for } Q_1)}{100}$$

usually the second C.R. for Q_1 will differ from the originally assumed C.R. 100, owing to the presence of trend. Some adjustment to the C.R.S are therefore necessary.

Let C be the average quarterly deviation of the 2nd C.R. from 100, i.e.

$$C = \frac{1}{4} \text{ (Second C.R. for } Q_1 \text{ - 100)}$$

Subtracting C, 2C, 3C and 4C from the C.R.s for Q_2, Q_3, Q_4 and the 2nd C.R. for Q_1, we find that both the C.R.s for Q_1 are now equal to 100. The adjusted C.R.s for Q_1, Q_2, Q_3, Q_4 are now expressed as Percentages of their arithmetic mean to give the seasonal indices. The total of these seasonal indices will be 400.

Example :

Compute seasonal indices from the data of example, using the method of Link Relatives.

Solution :

Each quarterly figure is expressed as a percentage of the figure in the preceding quarter, giving the Link Relatives, shown below :

Seasonal by Link Relative Method

	Link Relatives				
Year/Quarter	I	II	III	IV	I
1928	-	89.23	96.55	108.93	-
1929	111.49	92.65	100.00	106.35	-
1930	104.49	84.29	94.92	92.86	-
1931	115.38	91.67	92.73	113.73	
Total	331.31	357.84	384.20	421.87	-
A.M.	110.45	89.46	96.05	105.47	-
Chain Relative	100	89.46	85.93	90.63	100.10
Adjusted C.R.	100	89.44	85.93	90.56	100.10
Seasonal Index	109	98	94	99	-

Working Notes :

i) Link Relatives (L.R.) :

(58 ÷ 65) x 100 = 89.23, (56 ÷ 58) x 100 = 96.55

(61 ÷ 56) x 100 = 108.93 (68 ÷ 61) x 100 = 111.48 etc.

The A.M. of chain relatives for each quarter shows on an average the percentage value of that quarter in relation to the preceding quarter.

9.5 DESEASONALISATION OF DATA :

The objective of studying seasonal variations is (i) to measure them and (ii) to eliminate them from the given series. Elimination of the seasonal effects from the given values is termed as deseasonalisation of the data. It helps us to adjust the given time series for seasonal variations, thus leaving us with trend component, cyclical and irregular movements. Assuming multiplicative model of the time series, the diseasonalised values are obtained on dividing the given value by the corresponding indices of seasonal variations.

$$\text{Deseasonalised Date } = \frac{y}{s} = \frac{TCSI}{S} = TCI$$

Deseasonalistation is specially needed for the study of cyclical component. It also helps businessman and management executives for planning future production programmes, for forecasting and for managerial control. It also helps in proper interpretation of the data.

e.g. if the values are not adjusted for seasonality, then seasonal upswings (or downswings) may be misinterpreted as periods of of boom and prosperity in business.

In case of absolute seasonal variations (additive model) of the time series, the deseasonalised values are obtained on subtracting the seasonal variations from the given values. Thus

$$\begin{aligned}\text{Deseasonalised Date } &= y - s = (T + S + C + I) - S \\ &= T + C + I\end{aligned}$$

Example :

Deseasonalise the following data with the help of the seasonal data given below :

Month	Cash Balance (000 Rs.)	Seasonal Index
Jan	360	120
Feb	400	80
Mar	550	110
Apr	360	90
May	350	70
June	550	100

Solution :

Deasonalised values are obtained on dividing the given time series values (y) by the seasonal effect - assuming that the given series data follows multiplicate model of decomposition. we have

$$\text{Seasonal effect} = \frac{\text{Seasonal Index}}{100} = \frac{\text{S.I.}}{100}$$

Hence, using multiplicative model : y = T x S x C x I

$$\text{Deseasonalised value} = \frac{y}{\text{Seasonal Effect}}$$

$$= \frac{y}{\text{S.I.}} x100$$

Computation of deseasonalised values

Month	Cash Balance (000 Rs.)	Seasonal Index (S.I.)	Deseasonalised Value $=\frac{Y}{S.I.} x\,100$
Jan	360	120	$\frac{360}{120} x\,100 = 300$
Feb	400	80	$\frac{400}{80} x\,100 = 500$
Mar	550	110	$\frac{550}{110} x\,100 = 500$
Apr	360	90	$\frac{360}{90} x\,100 = 400$
May	350	70	$\frac{350}{70} x\,100 = 500$
June	550	110	$\frac{550}{100} x\,100 = 550$

If we assume the aditive model of decomposition, then the deseasonalised values are given by (y - S.I).

9.6 CYCLICAL COMPONENT :

Business cycles are perhaps the most important type of fluctuations in economic data. Certainly they have received a lot of attention in economic literature. Despite the importance of busienss cycles, they are the most difficult type of economic fluctuations to measures.

An approximate or crude method of measuring cyclical variations is this 'Residual Method' which consists in first estimating trend (t) and seasonal components (s) and then eliminating their effect from the giventime series.

Residual method consists in removing the three components viz. trend, seasonal and irregular, from the original data, in order to isolate the cyclical components. At first trend and seasonal effects are measured by suitable methods, say, by the movig average method. The two components are then removed from the data.

If the multiplicative model (y_t = T x S x C x I) is assumed, in order to get C x I, we remove T and S from Y_t by division we may either divide Y_t by T first, and thens, or divide yt by s first and then by t, or divide yt by the product t x s called normal value.

$$\frac{y_i}{\text{NormalValue}} = \frac{\text{T x S x C x I}}{\text{T x S}} = \text{C x I}$$

If, however, we assume the additive model (y_i = T + S + S + C + I), trend and seasonal components are removed by subtraction :

$$(y_t - T - X = (T \text{ x } S \text{ x } C \text{ x } I) - T - S = C + I$$

The residual now consists of a combination of cyclical and irregular components the irregular component (either C x I or C + I). At the next stage, the irregular component (I) is removed from residuals by smoothing, using moving averages. The appropriate period of the moving average depends on the average duration of irregular movements.

Residual method is laborious, but gives accurae results

Example : Obtain the estimates of the cyclical variations for the data following:

Computation of indices of cyclical variations

Year	Quarter	Original	Seasonal	$\frac{y}{s}$ x 100	Trend	$\frac{\text{Col 5}}{\text{Col 6}}$ x 100	3 quarterly
		values (y)	index (s)	= TCI	M.A.	= 100CI	M.A. of col (7((8)
1992	1	75	122.36	61.29	-	-	-
	2	60	92.42	64.92	-	-	-
	3	54	84.69	63.76	63.375	100.61	-
	4	59	100.51	58.70	65.375	89.79	98.37

1993	1	86	122.36	70.285	67.125	104.70	97.91
	2	65	92.42	70.3.	70.875	99.23	101.49
	3	63	84.69	74.39	74.00	100.53	101.78
	4	80	100.51	79.59	75.375	105.59	100.70
1994	1	90	122.36	73.55	76.625	95.99	100.65
	2	72	92.42	77.91	77.625	100.37	98.13
	3	66	84.69	77.93	79.500	98.03	100.725
	4	85	100.51	84.57	81.500	103.77	100.09
1995	1	100	122.36	81.73	83.000	98.47	100.61
	2	78	92.42	84.40	84.750	99.59	-
	3	72	84.69	85.02	-	-	-
	4	93	100.51	92.53	-	-	-

Last column (8) of table gives indices of cyclical variation.

9.7 RANDOM COMPONENT :

The irregular or random component, in a time series, represents the residue of fluctuations after trend cyclical and seasonal movements have been accounted for. Thus, if the original data is divided by T, S and C we get I, i.e. $\left(\frac{\text{TSCI}}{\text{TSC}}=\text{I}\right)$

In practice, the cycle itself is so erratic and is so interwoven with irregular movements that it is impossible to separate them. In the analysis of a time series into its component fluctuations, therefore, trend and seasonal movements are usually measured directly, while cyclical and irregular fluctuations are left together after the other elements have been removed. The cycle behaves in an erratic manner because successive cycles vary widely in period, amplitude and pattern and accordingly, it is very difficult to measure the cyclical variations accurately. Moreover, they are so much inter-mixed with irregular variations that, quite often, it becomes practically impossible to separate them. Accordingly, in analysis of time series, trend and seasonal components are measured separately and after eliminating their effect the cyclical and irregular fluctuations (C x I) are left together.

Although the random component cannot be estimated accurately, we can obtain an estimate of the variance of the random component by the 'variate difference' method.

9.8 BUSINESS FORECASTING:

All businessmen have to make a certain amount of forecasting regarding business conditions. Forecasts in business are necessary for various purposes e.g. judging future markets, making decisions on production, inventories, selling, priceing, etc.

Business forecasting is neither pure guesswork, nor finding the exact figures of businss conditions. The scientific methods of forecasting refer to the analysis of past and present conditions with a view to arriving at rough estimates about the future conditions.

Business forecasting is essentially statistical nature. The study of past conditions is done by analysing time series data into various components - trend, seasonal, cyclical and irregular. Here trend is determined by free-hand method or by fitting a mathematical curve, and is projected into the future. Another important part of the work of business forecasting lies in the construction of index numbers of business activity.

Business forecasting is now-a-day on scientific principles. Some of the important theories of business forecasting are :

i. Economic rhythm theory

ii. Specific history analogy theory

iii. Action and reaction theory

iv. Time lag or sequence theory and

v. Cross-cut analysis theory

Business forecasting has attained such an important position that in the economically advanced countries of the world, many forecasting agencies e.g. Harvard committee an economic research (U.S.A), London and Cambridge Economic Service (U.K.), Swedish Board of Trade etc. undertake the work of business forecasting on regular basis.

9.9 LET US SUM UP:

We conclude the lesson with this graph and its explanation.

The original data in this graph is represented by curve (a). The general movement persisting over a long period of time represented by the diagonal line (b) drawn through the irregular curve is called secular trend.

Next, the type of fluctuation which completes the whole sequence of changes within the span of a year and has about the same pattern year after year, is called seasonal variation.

Furthermore, looking at the broken curve superimposed on the original irregular curve, we find pronounced fluctuations moving up and down every few years throughout the length of the chart. These are known as business cyclical fluctuations.

Finally, the little saw-tooth irregularities on the original curve represent what are referred to as irregular movements.

In traditional or classical time series analysis, it is ordinarily assumed that there is a multiplicate relationship between these four components.

9.10 LESSON END EXERCISE:

Q1. Explain the method of moving average. How is it used in measuring trend in the analysis of a time series ?

Q2. Write down the merits and demerits of moving average method.

Q3. What do you understand by seasonal variations in the time series data ? Explain with few examples, the utility of such as study.

Q4. What are seasonal indices ? What methods are used to determine them ?

Q5. How do cyclical variations differ from seasonal variations ?

Q6. Calculate the trend values by the method of moving average, assuming a four-yearly cycle, from the following data relating to sugar production in India.

Year	Sugar Production (lakh tonnes)	Year	Sugar Production (lakh tonnes
1971	37.4	1977	48.4
1972	31.1	1978	64.6
1973	38.7	1979	58.4
1974	39.5	1980	38.6
1975	47.9	1981	51.4
1976	42.6	1982	84.4

Q7. Calculate seasonal indices by the ratio to moving average method from the following data.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
1991	68	62	61	63
1992	65	58	66	61
1993	68	63	63	67

Q8. Compute the seasonal indices by the link relatives method for the following data :

Wheat prices (Rs. 10 per kg)				
Quarter	1992	1993	1994	1995
Ist (Jan - Mar)	75	86	90	100
2nd (Apr - Jun)	60	65	72	78
3rd (Jul - Sept_	54	63	66	72
4th (Oct - Dec)	59	80	85	93

SUGGESTED FOR FURTHER READING:

- Nagar, et.al (1976), Basic statistics, Oxford University Press, New Delhi
- Croxton, et.al (1970), Applied general statistics, Prevtice Hall of India.
- Goon, A.M, M.K. Gupta and B. Dasgupta (1993). Fundamentals of statistics Vol. I, the world press, Calcutta.
- Das, N.G. (2012), statistical methods, Tata MCgraw Hill Education Private Limited, New Delhi.
- Williams, S.W.(2002), Stastics for Business and Economics, Thomson South-Western Pvt. Ltd. Singapore.

****

M.A. Economics **Semester - IInd** **Lesson - 10**

Course No. 203 **Unit - II**

MEASURES OF INEQUALITY - DESIRABLE PROPERTIES OF MEASURE OF INEQUALITY, GINI COEFFICIENT, LORENZ CURVE, KUZNET RATIO, CO-EFFICIENT OF VARIATION, RELATIVE RANGE

CHAPTER HIGHLIGHTS:

This lesson is a review of the recent advances in the measurement of inequality. Also discuss the desirable criteria for measures of inequality.

CHAPTER OUTLINES

10.1 INTRODUCTION

There is ongoing and increasing interest in measuring and understanding the level, causes and development of inequality during the 1990s. This period signified a shift in research previously focused on economic growth, the identification of the determinants of economic growth and contergence in GDP per capita across countries to analysis of distribution of income, its development overtime and identification of factors determining the distributino of income. This shift in focus is specifically from the issues of convergence or divergence of per capita incomes to the long term equalisation or polarisation of incomes across regions and countries of the world.

This shift is not only a reflection of technological change and raised human capacity to create growth, wealth and in the effective use of resource, but also due to awareness of the growing disparity and importance of redistribution and poverty reduction. The growing dispartiy calls for analysis of possible trends in global income inequality.

Inequality can have many dimensions. Economists are concerned specifically with the economics or monetarily measurable dimension related to individual or household income and consumption. However, this is just one perspective and inequality can be linked to inequality in skills, education, opportunities, happiness, health, life expectancy, welfare, assets and social mobility.

10.2 MEANING OF INEQUALITY

Inequality (also described as the gap between rich and poor, income, inequality, health dispartiy, wealth and income difference or wealth gap) is the state of affairs in which assets, wealth or income are distributed unequally among individuals in a group, among groups in a population or among countries.

Economic inequalities are most obviously shown by people's different positions within the economic distribution income-pay, wealth. However, people's economic positions are also related to other characteristics such as whether or not they have a disability, their ethnic background, or whether, they are a man or a women.

Income inequality uses the dispersion of capital to identify how economic inequality is defined among individuals in a given economy.

There are several reasons why development agencies should be concerned with inequality including :

1. **Inequality matters for poverty :** For a given level of average income, education, land ownership etc. increased inequality of these characteristics will almost always imply higher levels of both absolute and relative deprivation in these dimensions.

2. **Inequality matters for growth :** There is increasing evidence that countries with high levels of inequality - especially of assets - achieve lower economic growth rates on average.

3. **Inequality matters in its own right :** There is a strong and quite widely accepted, ethical basis for being concerned that there is a reasonable degree of equality between individuals though disagreement about the question equality of what ?

Inequality is clearly an important issue requiring much more attention in policy discussion than has been the case to date. Rather than conflicting with a poverty reduction focus or with the attainment of the MDGs, this is likely to be important for the successful attainment of these.

Conventional techniques for measuring and decomposing inequality remain useful and valuable, but it is important to broaden concepts of inequality beyond those typically considered in discussion on this issue. This includes developing a more multidimensional perspective on equality, but also other aspects such as considering inequality at different levels of aggregation and different time horizons. Drawing on both qualitative and quantative techniques is likely to be particularly valuable.

It is important to enrich our understanding the processes behind inequality and changes in inequality and to bring this to the forefront of policy debate.

10.3 MEASURES OF INEQUALITY

Inequality can have several dimensions. Economists are mostly concerned with the income and consumption dimensions of inequality. Several inequality indices including the most widely used index of inequality namely the Gini coefficient non-income inequality includes inequality in skills, education, opportunities, happiness, wealth and others.

The simplest measurement of inequality sorts the population from poorest to richest and shows the perentages of expenditure (or income) attributable to each fifth (qunitile) or tenth (decile) of the population. the poorest quntile typically accounts for 6-10 percent of all expenditure, the top quantile for 35-50 percent.

There are various ways of measuring economic inquality. The choice of measure does not change what inequality looks like dramatically. However, changes in inequality is often a significant factor behind crime, social unrest or violent

Conflict : These are often important contributors to povertyin their own right. Inequalities even perceived ones - between clearly defined groups, e.g. according to ethinicity may be important issue here.

v. Inequality is like to be critically important for the attainment of the millenium development goals.

Inequality is different from poverty but related to it. Inequality concerns variations in living standards across a whole population inequality over time within individual countries can look different if different measures are used.

10.4 COMMONLY USED MEASURES OF INEQUALITY:

10.4.1 Decile Dispersion Ratio :

A simple and popular measure of inequality is the decile dispersion ratio, which presents the ratio of the average consumption (or income) of the richest 10 percent of the population to the average consumption (or income) of the poorest 10 percent. This ratio can also be calculated for other percentiles (for instance, dividing the average consumption of the richest 5 percent, the 95th percentile, bythat of the poorest 5 percent, the 5th percentile.

The decile dispersion ratio is readily interpretable, by expressing the income of the top 10 percent as a multiple of that of those in the poorest decile. However it ignores information about incomes in the middle of the income distribution and does not even use information about the distribution of income within the top and bottom deciles.

10.4.2 Generalised Entropy Measures (Theil Index)

There are a number of measues of inequality. Among the most widely used are their indexes and the mean log deviation measure. Both belong tothe family of generalised entrotypes (GE) inequality measures. The general formula is given by

$$\text{GE}\,(\alpha = \frac{1}{\alpha(\alpha-1)}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\bar{y}}\right)^{\alpha}-1\right]$$

where $\bar{y}$ is the mean incom per person (or expenditure per capita). The values of GE measures vary between zero and infinity, with zero representing an equal distribution and higher values representing higher levels of inequality. The parameter α in the GE class represents the weight given to distances between incomes at different parts of the income distribution, and can take any real value. For lower values α, GE is more senstive to changes in the lower that of the distribution and for higher value

GE is more sensitive to change that affect the upper tail. The most common values of α used are 0, 1 and 2. GE(I) is Theil's index, which may be written as :

$$GE(I) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}} I_n\left(\frac{y_i}{\bar{y}}\right)$$

GE(O), also known as Theil's L, and sometimes referred to as the mean log deviation measures is given by

$$GE(o) \ = \frac{1}{N}\sum_{i=1}^{N} I_n\left(\frac{y_i}{\bar{y}}\right)$$

Once again, users of state do not need to program the computation of such measures from scratch the 'Pheqdeco' command allows one to obtain these measures even when weights need to be used with the data.

10.4.3 Atkinson's Inequality Measures :

Atkinson (1970) has proposed another class of inequality measures that are used from time to time. This class also has a weighting parameter (which measures aversion to inequality). The Atkinson class, which may be computed in stata using 'ineqdeco' command is defined as

$$A_t = 1 - \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_t}{\bar{y}}\right)^{1-E}\right]^{1/(1-E)} \qquad \in \neq 1$$

10.4.4 Rato Measures :

Ratio measures compare how much people at one level of the income distribution have compared to people at another. For instance, the 20:20 ratio compares how much richer the top 20% of people are, compared to the bottom 20%

Common examples :

- 50/10 ratio - describes inequality between the middle and the bottom of the income distribution.
- 90/10 - describes inequality between the top and the bottom

- 90/50 - describes inequality between the top and the middle
- 99/90 - describes inequality between the very top and top

10.4.5 Palma Ratio :

The palma ratio is the ratio of the income share of the tpp 10% to that of the bottom 40%. In more equal socieities this ratio will be one or below, meaning that the top 10% does not recieve a larger share of national income than the bottom 40%. In very unequal societies, the ratio may be as large as 7.

The Palma ratio addressed the Gini index's over-senstivitiy to changes in the middle of the distribution and insenstivity to changes at the top and bottom.

The UK Palma ratio is 1.07.

The Palma ratio is commonly used in international development discourse. The ratio for Brazil e.g. is 2.23.

10.4.6 Hoover Index :

Often touted as the simplest measurement to calculate, the Hoover index derives the overall amount of income in a system and divides it by the population to create the perfect proportion of distribution in the system. In a perfectly equal economy this would equate to income levels, and the deviance from this (on a percentile scale) is representative of the inequality in the system.

10.5 DESIRABLE PROPERTIES OF MEASURES OF INEQUALITY :

Inequality is the fundamental dispartiy that permits one individuals certain material choices, while denying another individuals those very same choices.

The inequality measures are each a scalar X. But many scalar inequality measures exist and can and do give conflicting results. Consider a set of four criteria of desirable properties of inequality measures. Each criteria relates to ethical judgement. Need to be cognizant of the relationship.

10.5.1 Anonymity Principle :

It does not matter who earns the income. Anonymity because we care about the ordering but not the identity of each earner. All that matters is the ranking from lowest to highest :

$y_1, y_2......y_n \rightarrow$ (y(i), y(2),..... y(n) order statistics

$y(1) \leq y(2) \leq y(3) \leq .... \leq y(n)$

Since identity of person doesn't matter we can dispense with the order statistics notation y(k) and use the simpler notation where index i represents the levels of the 9th income level in the society

$y(1) \leq y(2) \leq y(3) \leq .... \leq y(n)$

e.g. again n = 3

A = (10,20,30), B = (20, 10, 30)

Anonymity : W(A) = W(B)

10.5.2 Population Principle :

If we compare an income distribution over n people and another populations with 2n people with the same income pattern repeated twice, there should be no difference in inequality among the two income distributions.

Anonymity states that no information is lost by retaining only the sequence of individual incomes (and not the identities of each)

The population principle states it doesn't matter how large the population is, we can convert everything to percentiles (bottom 1%, lowest 20%, top 25%)

E.g. n = 3 A = (10, 20, 30)

n = 6 B = (10, 20, 30, 10, 20, 30)

Population principle : W(A) = W(B)

10.5.3 Relative Income Principle :

Only the relative incomes should matter and the absolute levels of these incomes should not.

Thus if we transform one distributions by multiply by a positive constant (e.g. $y^1 = \lambda y^o$) then inequality should be the same for the two distributions.

Roughly think of poverty as a measure of location (level) and inequality as a measure of dispersion.

10.5.4 Dalton Principle :

The two principles not controversial, third more difficult (well-being is proportional to income). This is fundamental to the construction of inequality measures.

Let (y_1, y_2.... y_n) be an income distribution and consider two incomes y_i and y_j with $y_i \le y_j$

A transfer of income from individual i to individual j is called a regressive transfer.

If inequalities is strict $y_i \le y_j$ the regressive transfer is from the poorer individual to the richer individual.

With weak inequality ($\leq$) use the language 'not richer' to 'not poorer'.

Our inequality index as a function of the form.

$I = I\ (y_1, y_2....y_n)$

with I defined over all conceivable distributions of income (y_1, y_2....y_n)

Dalton Principle :

If one income distribution can be achieved from another by constructing a sequence of regressive transfers, then the former distribution must be deemed more unequal than the latter.

If for every income distribution (y_1, y_0....y_n) and every transfer s >0,

$1(y_1, ... y_i ....y_j...........y_n) < 1\ (y_1......y_2........ \varsigma ......y_j + \varsigma ,........y_n)$

Three simple measures of inequality

Range $R = \mu^{-1}(y_a - y_1)$, crude but sometimes useful.

Kuznets Ratio Richest x 1% to the poorest y% where x and y stand for numbers such as 10, 20 or 40. Also crude, focuses on ends of distribution.

Mean Absolute Deviation $M = \frac{1}{N\mu}\sum_{g=1}^{m} n_g \left| y_g - \mu \right|$. Do not use

Three measures that satisfy all four principles

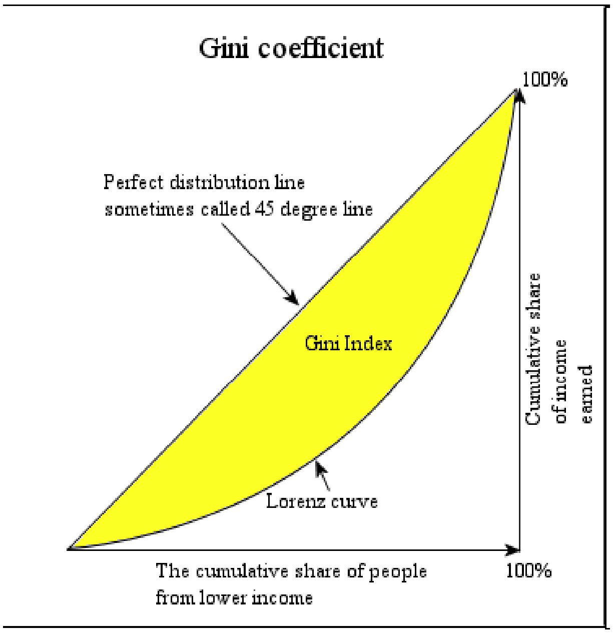
Coefficient of variation $C = \sqrt{\sum_{g-1}^{m} \frac{n_g}{N} \left(\frac{y_g - \mu}{\mu} \right)^2}$

Gini coefficient $G = \frac{1}{2n^2 \mu} \sum_{j=1}^{m} \sum_{k=1}^{m} n_j n_k \left| y_j - y_k \right|$

Theil Index $T = \frac{1}{N\mu} \sum_{g=1}^{m} y_g^{\,in} (yg/\mu)$

10.6 GINI COEFFICIENT

The Gini coefficient is a measure of inequality of a distribution. It is defined as a ratio with values between 0 and 1, the numerator is the area between the Lorenz curve of the distribution and the uniform distribution line, the denominator is the area under the uniform distribution line. It was developed by the Italian statistician Corrado Gini and published in his 1912 paper 'Variabilita e mutabilita (variability and mutability). The Gini index is the Gini coefficient multiplied by 100. (The Gini coefficient is equal to half of the relative mean different



The Gini coefficient is often used to measure income inequality. Here, O corresponds to perfect income equality (i.e. everyone has the same income) and 1 corresponds to perfect income inequality (i.e. one person has all the income, while everyone else has zero income.

The Gini coefficient can also be used to measure wealth inequality. This use requires that no one has a negative net wealth. It is also commonly used by the measurement of discriminatory power of rating systems in the credit risk management.

10.6.1 Calculation

The Gini coefficient is defined as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is Z, and the Lorenz curve is B, then the Gini coefficient is A(A+B). Since A+B=0.56, the Gini cofficient, G=2A A=1-2B. If the Lorenz curve is represented by the function y = L(x), the value of B can be found with integration and :

$$G = 1 - 2\int_0^1 L(x)dx$$

In some cases, this equation can be appied to calculate the Gini coefficient without direct reference to the Lorenz curve. For example :

For a population with value y_i, i=1 to n, that are indexed in non-decreasing order $(y_i \leq y_{i+1})$

$$G = \frac{1}{N}\left(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i} \right)$$

For a discrete probability function f(y), where y_i, i=1 to n, are the points with non zero probabilities and which are indexed in increasing order $(y_i \leq y_{i+1})$

$$G = 1 - \frac{\sum_{j=1}^{i} \int (y_i)(s_{i-1} + s_i)}{\varsigma_n}$$

where

$$s_i = \sum_{j=1}^{i} \int(y_i)y_j \text{ and } \varsigma_o = 0$$

For a cumulative distribution function fey) that is piecewise differentiable,has a mean u , and is zero for all negative values of y :

$$G = 1 - \frac{1}{\mu}\int_{o}^{d} \int(1 - F(y)^2\, dy$$

Since the Gini coefficient is half the relative mean difference, it can also be calculated using formula for the relative mean difference.

For a random sample s consisting of values Yi' i=I to n, that are indexed in non-decreasing order (Yi s Yi+l)' the statistic.

$$G(s) = \frac{1}{n-1}\left(n+1-2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} yi} \right)$$

is a consistent estimator of the population Gini coefficient, but is not, in general unbiased, Like the relative mean difference there does not exist a samole statistic that is in general an sed estimator of the population Gini coefficient can be calculated using boot strap techniques.

Sometimes the entire Lorenz curve is not known and only values at certain intervals are given the that case the Gini coefficient can be approximated by using various techniques for interpolating the missing values of the Lorenz curve. If (X_k, Y_k) are the known points are the Lorenz curve, with the X_k indexed in increasing order $(x_{k-1}\, x_k)$ so that, Y_k is the cumulated proporition of the population variable, for k=0, n, with $x_0 = o$, $x_n = 1$

Y_k is the cumulated proportion of the income variable, for k = 0, n, with $y_o = o$, $y_n = 1$

If the Lorenz curve is approximated on each interval as a line between cosecutive points, then the area B can be approximated with trapezoids and :

is the resulting approximation for G. More accurate results cum be obtained using other methods to approximate the area B, such as approximating the Lorenz curve with a quadratic function across pairs of intervals, or building an appropriately smoth approximation to the underlying distribution funtion that matches the known data, If the population mean and boundary values for each interval are also known, these can also ten be used to improve the accuracy of the approximation.

While most developed European national tend to have civil coefficient between 0.24 and 0.36, the United State Gini coefficient is above 0.4, indicating that the United State has greater inequality. Using the Gini can help quantify difference in welfare and compensation policies and philosophies. However, it should be borne in mind that Gini coefficient can be misleading when used to make political comparisons between large and small countries.

10.6.2 Advantages as a measure of a inequality :

- ♦ The Gini coefficient's advantage is that it is a measure of inequality by means of a ratio analysis, rather than a variable unrepresentative of most of the population, such as per capital income or gross domestic product.
- ♦ It can be used to compare income distributions across different population sectors as well as countries, for example the Gini coefficient for urban areas differs from that of rural areas in many countries.
- ♦ It is sufficiently simple that it can be compared across countries and be easily interpreted. GDP statistics are often criticised as they do not represent changes for the whole populations, the Gini coefficient demonstrates how income has changed for poor and rich. If the Gini coefficient is rising as well as GDP, poverty may not be improving for the majority of the population.
- ♦ The Gini coefficient can be used to indicate low the distribution of income has changed within a country over a period of time, thus it is possible to see if inequality is increasing or decreasing.
- ♦ The Gini coefficient satisfies four important prnciples :
  - **Anonymity :** It does not matter who the high and low earners are.

- **Scale independence** : The Gini coef;ficient does not consider the size of the economy, the way it is measured, or whether it is a rich or poor country on average.
- **Population independence** : It does not matter how large the population of the country is.
- **Transfer principle** : If income (less than the difference), is transferred from a rich person to a poor person the resulting distribution is more equal

10.6.3 Disadvantages as a measure of inequality :

- ♦ The Gini coefficient measured for a large economically diverse country will generally result in a much higher coefficient than each of the its regions has individually. For this reason the scores calculated for individual countries within the EU are difficult to compare with the score of the entire US.
- ♦ Comparing income distributions among countries may be difficult because benefits systems may differ. For example, some countries give benefits in the form of money while others give food stamps, which may not be counted as income in the Lorenz curve and therefore not taken into account in the Gini coefficient.
- ♦ The measure will give different results when applied to individuals instead of households. When different populations are not measured with consistent definitions, comparison is not meaningful.
- ♦ The Lorenz curve may undertake the actual amount of inequality if richer households are able to use income more efficiently than lower income households from another point of view, measured inequality may be the result of more or less efficient use of household incomes.
- ♦ As for all statistics, there will be systematic and random errors in the data. The meaning of the Gini coefficient decreases as the data becomes less accurate. Also, countries may collect data differently, making it difficult to compare statistics between countries.
- ♦ Economies with similar incomes and Gini coefficient can still have very different

income distributions. This is because the Lorenz curve can have different shapes and yet still yield the same Gini coefficient . As an extreme example, an economy where half the households have no income, and the other half share income equally has a Gini coefficient of ½, but an economy with complete income equality, except for one wealthy household that has half the total income, also has a Gini coefficfient of ½.

10.7 LORENZ CURVE

Lorenz curve is a simple diagrammtic way to depict the distribution of income. On the horizon axis we list the cumulative percentage of the population arranged in increasing order of income.

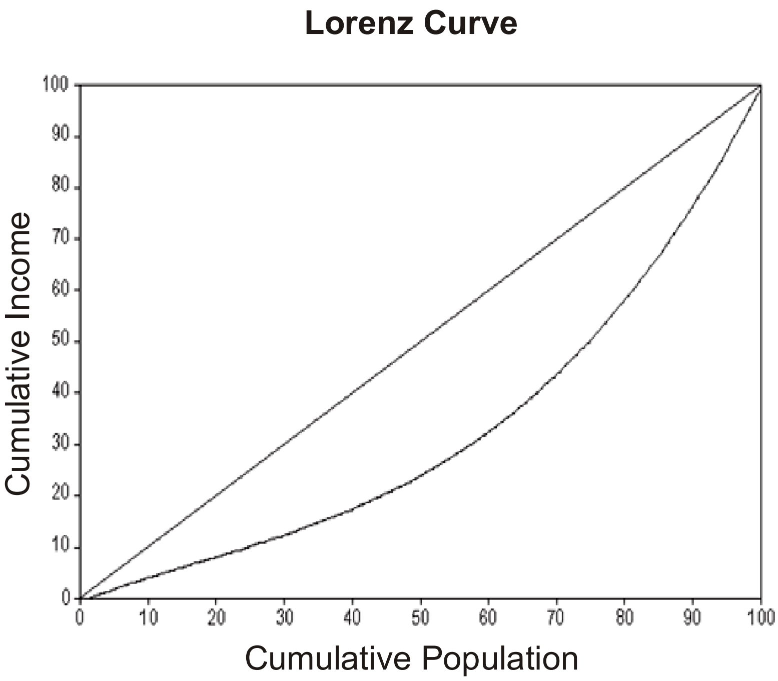
Thus point on the axis refer to the poorest 20% of the population, the poorest half, etc.

On the vertical axis we measure the percentage of the national income accuring to any particular fraction of the population thus arranged. The diagonal line (45°) represents equal distribution income.

The slope of the Lorenz curve is the contribution of the person at that point to the cumulative share of national income.

Ordered from poorest to richest the ‘marginal contribution’ can never fall.

Equivalently, the Lorenz curve can never get flatter as we move from left to right. The overall distance between the 45° and the Lorenz curve represents the amount of inequality present in the society.



10.8 THE CO-EFFICIENT OF VARIATION (CV)

This measure of income inequality is calculated by dividing the standard deviation of the income distribution by its mean.

Symbolically -

$$\text{CV}=\frac{\sqrt{Var}}{\bar{y}} \text{ or } \frac{S.D.}{\bar{y}} \text{ or } \frac{\sigma}{\bar{y}}$$

Where, σ - Standard deviation

$\bar{y}$ – Mean Income

The income distribution is more equal, when there is smaller standard deviation; as such the CV will be smaller in more equal societies. Despite being one of the simplest measures of inequality, use of the C.V. has been fairly limited in the public health literature and it has not featured in research on the income inequality hypothesis.

Thus, the important limitations attributed to the CV measure are -

- ♦ It doesn't have an upper bound, unlike the Gini Co-efficient, making interpretation and comparison some what more difficult; and
- ♦ The two components of the CV (the S.D. and the $\overline{Y}$) may be exceedingly influenced by anamalously low or high income values.

In nutshell, it might not be an appropriate choice for income inequality measure.

10.9 THE RELATIVE RANGE MEASURE

A range is an interval that defines the minimum and maximum values for any set of numbers or for the variation of a particular variable, for example - a stock price on the market. The percent relative range refers to the percentage ratio of the range to the average value in the set.

Sum up the maximum and minimum values in the range. For example - if the stock price changes in the range from

$ 34.68 to $ 41.12.
Then, $ 34.68 + $ 41.12 = $ 75.80

Now, Divide the sum by "2" to calculate the average value. In the example, the average price is

$$\frac{\$75.80}{2}=\$37.90$$

Now, subtract the minimum value from maximum one to calculate the range. In this example, the range is

$ 41.12 - $ 34.68 = $ 6.44

Now, finally, divide the range by the average value, and then multiply the result by 100 to calculate the relative percent range. In this example, the relative percent range is

$$\frac{\$6.44}{\$37.90}\text{x}\,100=16.99\%$$

10.10 THE KUZNET RATIO

The Kuznet Ratio is defined in two ways -

- One is a measure of inequality, and
- Other as a measure of equality.

As a measure of Inequality -

The Kuznets Ratio is defined as the share of total income received by the 20% richest people divided by the share of total income received by the 40% poorest people in the society. This is a measure of inequality, i.e., if the value is high (or the number is high), then the society is unequal.

As a measure of Equality -

The Kuznets Ratio is defined as the share of total income received by the 40% poorest people is divided by the share of total income received by the 20% richest people of the society. This is a measure of equality, i.e., if the value or the number is high, then the society is quite equal.

10.11 LET US SUM UP

Inequality is typically viewed as different people having different degrees of something. Often considered in terms of income or consumption but equally appliable to other dimensions of living standards that show a continuous pattern of variation.

Many inequality indices have been developed, and some of these have additonal desirable properties. Economists are mostly concerned with the income and consumption dimesnions of inequality. Among other non-income inequality dimensions, we can include inequality in skills, education, opportunities, happiness, health, life-years, welfare and assets.

Several inequality indices can be derived from the Lorenz diagram. The divergence of a Lorenz curve for perfect equality and the Lorenz curve for a given distribution is measured by some index of inequality. Several inequality indices follow along with some basic properties that one would expect the indices to satisfy. These properties are to be used in their ranking, relevance and performance evaluation. The most widely used index of inequality is the Gini coefficient. Gini is generalised to accomodate differing aversion of inequality.

10.12 LESSON END EXERCISE

Q1. What do you mean by inequality in economics ?

Q2. Discuss the various measures of inequality.

Q3. Explain the meaning of Gini coefficient with diagramtically ?

Q4. Write down the merits and demerits of Gini coefficient as a measure of inequality ?

Q5. How Lorenz curve helps to show the income distribution in the economy.

****

M.A. Economics **Semester - IInd** **Lesson - 11**

Course No. 203 **Unit - III**

PROBABILITY THEORY - DIFFEENT CONCEPTS AND APPROACHES, LAWS AND AXIOMS OF PROBABILITY, CONDITIONAL THEORY AND CONCEPT OF INTERDEPENDENCE, BAYES' THEOREM AND ITS APPLICATIONS

STRUCTURE

11.1 INTRODUCTION :

The theory of probability has its origin in the games of chance related to gambling, for instance, throwing of dice or coin, drawing cards from a pack of cards and so on. However, a systematic and scientific foundation of the mathematical theory of probability was laid in mid-seventeenth century by the French Mathematicians Blaise Pascal (1623-62) and Piesse de Fermat (1601-65) while solving a problem for sharing the stake in an incomplete gambling match posed by a notable French gambles and nobleman chevalies-de-Mere. Today the subject has been developed to a great extent and there is not even a single discipline in social, physical or natural sciences where probability theory is not used. It is extensively used in the quantitative analysis of business and economic problems. It is an essential tool in statistical nference and forms the basis of the 'Decision Theory', viz., decision making in the face of uncertainty with calculated risks.

11.2 CONCEPT OF PROBABILITY THEORY

If an experiement is performed repeatedly under essentially homogenous and similar conditions, the result or what is commonly termed as outcome may be classified as follows :

a) It is unique as certain

b) It is not definite but may be one of the various possibilities depending on the experiment.

The phenomenon under category (a) whose the result can be predicted with certainty is known as deterministic or predictable phenomenon. In a deterministic phenonemon, the conditions under which an experiment is performed, uniquely determine the outcome of the experiment. For instance :

i. In case of a perfect gas we have Boyle's law which states,

Pressure x volume = Constant i.e. PV = Constant => provided the temperature remains constant.

ii. The distance (s) covered by a particle after time (t) is given by $s = ut + \frac{1}{2}at^2$

where u is the initial velocity and a is the acceleration.

iii. If dilute sulphuric acidd is added to zinc, we get hydrogen.

Thus, most of the phenomena in physical and chemical sciences are of a deterministic nature. However, there exist a number of phenomena as generated by category (b) where the results cannot be predicted with certainity and are known as unpredictable or probabilistic phenomena. Such phenomena are frequently observed in economics, business and social sciences or even in our day-to-day life. For example,

a. The sex of a baby to be born cannot be predicted with certainty.

b. A sales (or production) manager cannot say with certainty if he will achieve the sales (or production) target in the season.

c. If an electric bulb or tube has lasted for 3 months, nothing can be said about its future life.

d. In toss of a uniform coin, we are not sure if we shall get head or tail.

e. A produces can not ascertain the future demand of his product with certainty.

Even in out day-to-day life we say or lear phrases like "It may rain today"; "Probably I will get a first class in the examination"; "India might draw or win the cricket series against Australia", and so on. In all the above cases there is involved an element of uncertainty or chance. A numerical measure of uncertainty is provided by a very important branch of statistics called the 'Theoryof Probability'. In the words of Prof. Ya-Lin-Chou, "Statistics is the science of decision making with calculated risks in the face of uncertainity."

Following concepts of probability will be discussed with reference to simple experiments relating to tossiong of coins, throwing of a die (cube with six faces bearing number 1 to 6) or drawing cards from a pack of cards.

a) **Random Experiment :** An experiment is called a random experiment if when conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the various possible outcomes.

b) **Trial and Event :** Performing of a random experiment is called a trial and outcome as combination of outcomes are termed as events. For example:

   i) If a coin is tossed repeatedly, the result is not unique. We may get any of the two faces, head or tail. Thus tossing of a coin is a random experiment or trial and getting of a head or tail is an event.

   ii) Similarly, throwing of a die is a trial and getting any one of the faces 1, 2, ..., 6 is an event, or getting of an odd number of an even number is an event; or getting a number greater than 4 or less than 3 are events.

   iii) Drawing of two balls from an urn containing 'a' red balls and 'b' white balls is a trial and getting of both red balls, as both white balls, as one red and one white ball are events.

   Event is called simple if it corresponds to a single possible outcome of the experiment as trial otherwise it is known as a compound or composite event. Thus, in tossing of a single die, the event of getting '5' is a simple event but the event getting an even number', is a composite event.

c) **Exhaustive Cases :** The total number of possible outcomes of a random experiments called the exhaustive cases for the experiment. Thus, in toss of a single coin, we can get head (H) as tail (T). Hence exhaustive number of cases is 2, viz., (H,T). If two coins are tossed, the various possibilities are HH, HT, TH, TT whose HT means head on the first coin and tail on second coin, and TH means tail on the first coin and head on the second coin and so on. Thus, in case of toss of two coins, exhaustive number of cases is 4, i.e. 2^2. Similarly, in a toss of three coins the possible number of outcomes is :

(H, T) x (H, T) x (H, T)

= (HH, HT, TH, TT) x (H, T)

= HHH, HTH, THH, TTH, HH,T, HTT, THT, TTT

Therefore, in case of toss of 3 coins the exhaustive number or cases is $8=2_3$. In general, in a throw of n coins, the exhaustive number of cases is 2_n.

d) **Favourable cases or events :** The number of outcomes of a random experiment which entail (or result in) the happening of an event are termed as the cases favourable to the event. For example :

In a toss of two coins, the number of cases favourable to the event 'exactly on ehead' is 2, viz., HT, TH and for getting 'two heads' is one viz., HH.

e) **Mutually exclusive events or cases :** Two are more events are said to be mutually exclusive the happening of all others in the same experiment. For example, in toss of a coin, the events 'head' and 'tail' are mutually exclusive because of head comes, we can't get tail and if tail comes we can't get head. Similarly, in the throw of a die, the six faces numbered 1, 2, 3, 4, 5 and 6 are mutually exclusive. Thus, events are said to be mutually exclusive if no two or more of them can happen simultaneously.

f) **Equally likely cases :** The outcomes are said to be equally likely or equally probable if none of them is expected to occur in preference to other. Thus, in tossing of a coin (dice), all the outcomes, viz., H, T (the faces 1, 2, 3, 4, 5, 6) are equally likely if the coin (die) is unbiased.

g) **Independent events** : Events are said to be independent of each other if happening of any one of them is not affected by and does not affect the happening of any one of others. For example: In tossing of a die repeatedly, the event of getting '5' in Ist throw is independent of getting '5' in second, third as subsequent throws.

11.3 APPROACHES TO PROBABILITY:

There are 3 approaches to probability:

i) Classical approach

ii) Empirical approach

iii) Axiomatic approach

11.3.1 Mathematical or classical or 'A priori' Probability :

Definition : If a random experiment results in N exhaustive mutually exclusive and equally likely outcomes (cases) out of which m are favourable to the happening of an event A, then the probability of occurence of A, usually denoted by P(A) is given by :

$$P(A) = \frac{\text{Favourable number of cases to A}}{\text{Exhaustive number of cases}} = \frac{M}{N} \qquad ....(1)$$

This definition was given by James Bernoulli who was the first man to obtain a quantitative measure of uncertanity

Remarks :

1) Obviously, the number of cases favourably to the complementry event $\overline{A}$ i.e., non-happening of event A are (N-M) and hence by definition, the probability of non-occurence of A is given by :

$$P(\overline{A}) = \frac{\text{Favourable No. of Cases to } \overline{A}}{\text{Exhaustive Number of Cases}} = \frac{N-M}{N} = 1 - \frac{M}{N}$$

$$\Rightarrow P(\overline{A}) = 1 - P(A) \text{ ....(2)}$$

$$\Rightarrow P(A) + P(\overline{A}) = 1 \text{ ....(3)}$$

2) Since M and N are non-negative integers, $P(A) \geq 0$. Further, since the favourable numbr of cases to A are always less than or equal to the total number of cases N, i.e., $M \leq N$, we have $P(A) \leq 1$. Hence probability of any event is a number lying between 0 and 1, i.e.,

$$O \leq P(A) \leq 1, \text{.............(4)}$$

for any event A. If P(A) = O, then A is called an impossible or null event. If P(A) = 1, then A is called a certain event.

3) The probability of happening of the event A, i.e., P(A) is also known as the probability of success and is usually written as p and the probability of the non-happening fA, i.e., P($\overline{A}$) is known as the probability of failure, which is usually denoted by q. Thus, from (2) and (3), we get

$$q = 1 - P \Rightarrow P + q = 1..........(5)$$

4) According to the above definition, the probability of getting a head in a toss of an unbiased coin is $\frac{1}{2}$, since the two exhaustive cases H and T (assuming the coin does not stand on its edge), are mutually exclusive and equally likely and one is favourable to getting a head. Similarly, in drawing a card from a well shuffled pack of cards, the probability of getting an ace is $\frac{4}{52} = \frac{1}{13}$. Thus, the classical definition of probability does not require the actual experimentation, i.e. no experimental data are needed for its computation, nor it is based on previous experience. It enables us to obtain probability by logical reasoning prior to making any actual trial and hence it is also known as 'a prior' or theoretical or mathematical probability.

Limitations:

The classical probability has its short-comings and fails in the following situations :

1. If N, the exhaustive number of outcomes of the random experiment is infinite.
2. If the various outcomes of the random experiment are not equally likely. For example, if a person jumps from the top of Qutab Minar, then the probability of his survival will not be 50%, since in this case the two mutually exclusive and exhaustive outcome, viz., survival and death are not equally likely.

3. If the actual value of N is not known. Suppose an turn contains some balls of twocolours, say red and white, their number being unknown. If we actually draw the balls from the urn, then we may from some idea about the ratio of red to the white balls in the urn. In the absence of any such experimentation (which is the case in classical probability), we cannot draw any conclusion in such a situation regarding the probability of drawing a white or a red ball from the urn. This drawback is overcome, in the statistical or empirical probability which we discuss below.

11.3.2 Statistical or Empirical Probability:

Definition (Von Mises): If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the numbers of trials, as the number of trials becomes indefinitely large, is called the probability of happening of the event, it being assumed that the limit is finite and unique.

Suppose that an event A occurs m times in N repetitions of a random experiment. Then the ratio M/N gives the relative frequency of the event A and it will not vary appreciably from one trial to another. In the limiting case when N becomes sufficiently large, it more or less settles to a number which is called the probability of A. Symbolically,

$$P(A) = \lim_{n \to \infty} \frac{M}{N} \text{.................}(6)$$

Remarks

1. Since in the relative frequency approach, the probability is obtained objectively by repetitive empirical observations, it is also known as empirical probability.
2. The empirical probability provides validity to the classical theory of probability. If an unbiased coin is tossed at random, then the classical probability gives the probability of a head as $\frac{1}{2}$. Thus, if we toss an unbiased

coin 20 times, then classical probability suggests we should have 10 heads. However, in practice, this will not generally be true. In fact in 20 throws of a coin, we may get no head at all or 1 or 2 heads. However, the empirical probability suggests that if a coin is tossed a large number of times, say 500 times, we should on the average expect 50% head and 50% tails. Thus, the empirical probability approaches the classical probability as the number of trials becomes indefinitely large.

Limitations :

It may be remarked that the empirical probability P(A) defined in (6) can never be obtained in practice and we can only attempt at a close estimate of P(A) by making N sufficiently large. The following are the limitations of the experiments.

1. The experimental conditions may not remain essentially homogeneous and identical in a large number of repititions of the experiment.
2. The relative frequency M/N, may not attain a unique value, no matter however large N may be.

e.g.1 : A uniform die is thrown at random. Find the probability that the number on it is :

i) 5 ii) Greater than 4 iii) Even

Solution :

Since the dice can fall with any one of the faces 1, 2, 3, 4, 5 and 6 the exhaustive number of cases is 6.

i) The number of cases favourable to the event of getting ‘5’ is only 1.

$$\therefore \quad \text{Required probability} = \frac{1}{6}$$

ii) The number of cases favourable to the event of getting a number greater than 4 is 2, viz., 5 and 6

$$\therefore \quad \text{Required probability} = \frac{2}{6} = \frac{1}{3}$$

iii) Favourable cases for getting an even number are 2, 4 and 6 i.e. 3 in all.

$\therefore$ Required probability $= \frac{3}{6} = \frac{1}{2}$

11.3.3 Axiomatic probability (approach)

The modern theory of probability is based on the axiomatic approach introduced by the Russian mathematician A.N. Kolmogorov in 1930's. He axiomised the theory of probability and his small book 'Foundations of Probability', published in 1933, introduces probability as a set function and is considered as a classic. In axiomatic approach, some concepts are laid down and certain properties as postulates, commonly known as axioms, are defined and from these axioms alone the entire theory is developed by logical of deduction. The axiomatic definition of probability includes both the classical and empricaldefnitions of probability and at the same time is free from their drawbacks. Before giving examples axiomatic definition of probability, we shall explain certain concepts, used therein.

Sample Space: The set of all possible outcomes of a random experiment is known as the sample space is denoted by S. In other words, sample space is the set of all exhaustive cases of the random experiment. The outcomes of the experiment are also known as sample points. Mathematically, if e_1, e_2...., e_n are the mutually exclusive possible outcomes of a random experiment, their the set S = {e_1, e_2.....,e_n} is said to be sample space of the experiment. The elements of S possess the following properties :

a) Each of the e_i's (i = 1, 2,.....n) is outcome of the experiment.

b) Any repetition of the experiment results in an outcome corresponding to one and only one of the e_i's.

Remarks:

We shall write n(s) to denote the number of elements i.e., sample points in S.

Illustration 1:

If a coin is tossed at random, the sample space is S = (H, T) and n(s) = 2.

If two coins are tossed then the sample space is given by :

$S = \{(H,T) \text{ x } (H,T)\} = \{HH, HT, TH, TT\}$ and $n(s) = 4 = 2^2$

In a toss of three coins

$S = \{(H, T) \text{ x } (H, T) \text{ x } (H, T)\} = \{(HH, HT, TH, TT) \text{ x } H,T)\}$

$S = \{HHH, HTH, THH, TTH, HHT, HTT, THT, TTT\}$

$n(s) = 8 = 2^3$

In general, in a random toss N coins, $n(s) = 2^n$

Event : Of all the possible outcomes in the sample space of a random experiments, some outcomes satisfy a specified description, which we call an event. For example, as already discussed, in a toss of 3 coins the sample space is given by:

$S = \{HHH, HTH, THH, TTH, HHT, HTT, THT, TTT\}$

$= \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8\}$, say.... (7)

where $w_1 = HHH$, $w_2 = HTH$, $w_3 = THH$, ..., w_8=TTT.

for this example space we can define a number of events, some of which are given below:

E_1 : Event of getting all heads = $\{HHH\} = \{w_1\}$

E_2 : Event of getting exactly two heads = $\{HTH, THH, HHT\}$

$= \{w_2, w_3, w_5\}$

E_3 : Event of getting at least two heads

$= \{w_1, w_2, w_3, w_5\} = \{w_1\} \cup \{w_2, w_3, w_5\} = E_1 \cup E_2$,

Where E_1 and E_2 are disjoint.

E_4: Event of getting exactly one head = $\{w_4, w_6, w_7\}$

E_5 : Event of getting at least one head

$= \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\} = \{w_1, w_2, w_3, w_4, w_5\} \cup \{w_4, w_6, w_7\}$

$= E_3 \cup E_4 = E_1 \cup E_2 \cup E_4$, where E_1, E_2 and E_4 are disjoint

E_6 : Event of getting all tails = {TTT} = $\{w_8\}$.

Thus, rigorously speaking an event may be defined as a non-empty sub-set of the sample space. Every event may be expressed as a disjoint union of some subjects of S or a disjoint union of some subjsets of S. Since events are nothing but sets, the algebra of sets may be used to deal with them.

The two events A and B are said to be disjoint or mutually exclusive if they cannot happen simultaneously i.e., if their intersection in a null set. Thus if A and B are disjoint events, then

$$A \cap B = \phi \Rightarrow P(A \cap B) = P(\phi) = 0 \text{ .... } (8)$$

Thus $P(A \cap B) = 0$, provides us with a criterion for finding if A and B are mutually exclusive.

Axiomatic Probability

Definition : Given a sample space of a random experiment, the probability of the occurence of any event A is defined as a set function P(A) satisfying the following axioms.

Axiom 1 : P(A) is defined, is real and non-negative

i.e. , $P(A) \geq 0$ (Axiom of non-negativity).........(9)

Axiom 2 : PCS = 1 (Axiom of certainity)...........(10)

Axiom 3 : If $A_1, A_2, \ldots, A_n$ is any finite as infinite sequence of disjoint events of s, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ or } P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \text{ (Axiom of additivity)....(11)}$$

events as set - glossary of probability terms

If A and B are two events then :

$A \cup B$: An event which represents the happening of at least one of the events A and B, (i.e. either A occurs or B occurs or both A and B occur)

$A \cap B$: An event which represents the simultaneous happening of both the events A and B.

$\overline{A}$: A does not happen.

$\overline{A} \cap \overline{B}$: Neither A nor B happens i.e., none of A and B happens.

$\overline{A} \cap B$: A does not happen but B happens.

$(A \cap \overline{B}) \cup (\overline{A} \cap B)$: Exactly one of the two events A and B happens.

The above notations can be generalised for n events, say $A_1, A_2, \ldots, A_n$. Thus :

$A_1 \cap A_2 \cap \ldots \cap A_n$: A compound event which represents the simultaneous happening of all the events $A_1, A_2, \ldots., A_n$.

$A_1 \cup A_2 \cup \ldots \cup A_n$: An event which represents the happening of at least one of the events $A_1, A_2, \ldots, A_n$.. This involves the events of the type $A_1, A_2, \ldots, A_n$.(one at a time).

$A_i \cap A_j (i \neq j \neq k = 1, 2, \ldots., n)$, i.e., simultaneous happening of three at a time,...., and $A_1 \cap A_2 \cap \ldots \cap A_n$ i.e., all the n at a time.

However, if $A_1, A_2, \ldots, A_n$ are mutually disjoint, they cannot happen simultaneously, i.e., $A_i \cap A_j, A_i \cap A_j \cap A_k, \ldots\ldots, A_1 \cap A_2 \cap, \ldots., \cap A_n$ are all null events and in that case $A_1 \cup A_2 \cup \ldots \cup A_n$ will represents the happening of any one of the events $A_1, A_2, \ldots, A_n$.

11.3.4 Comparison of Classical & Axiomatic approach to probability :-

Classical Approach	Axiomatic Approach
All the theorems and results are obtained by logical arguments	All results theorems are derived from theaxioms by using the mathematical properties of sets
Based on the concept of equally likely cases when the no. of possible outcomes cases when the no. of possible outcomes is finite	Quite general and embraces all cases, irrespective of whether no. of possible outcomes in finite or not

It defines probability as a ratio of two positive whole numbers showing the cases favourable to the events and total no. of all events and total no. of all events which are generally equally likely.	It defines probability as a non-negative number associated with the event i.e. probability is a set function obeying the three axioms
The concept of conditional probability and independent events are introduced in it.	These are defined by only mathematical statements in it.

11.4 PROBABILITY - MATHEMATICAL NOTION :

Let us suppose that s is the sample space of a random experiment with a large number of trials with sample points (number of all possible outcomes) N, i.e., n(s) = N. Let the number of occurences (sample points) favourable to the event A be denoted by n(A). Then the frequency interpretation of the probability gives:

$$P(A) = \frac{n(A)}{n(S)} = \frac{n(A)}{N}$$

But in practical problems, writing down the elements of s and counting the number of cases favourable to a given event often proves quite tedious. For example, if a die is thrown three times, then total number of sample points would be $6^3 = 216$ and if 3 cards are drawn from a pack of cards without replacement three could be 52 x 51 x 50 = 132,600 sample points. To write them is a very difficult task and is quite often unnecessary. However, in such situations the computation of probabilities can be facilitied to a great extent by the two fundamental theorems of probability - the addition theorem and the multiplication theorem.

11.5 THEOREMS OF PROBABILITY

Addition Theorem of Probability

Theorem 1: The probability of occurence of at least one of the two events A and B is given by :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)......(1)$$

Proof : Let us suppose that a random experiment results in a sample space S with N sample points (exhaustive numbers of cases). Then by definition:

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N} \quad .....(2)$$

$\cap$

Where n $(A \cup B)$ is the number of occurences (sample points) favourable to the event $(A \cup B)$.

From the fig 3.1 we get :

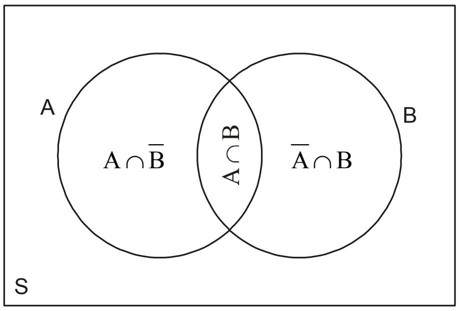


Fig 3.1

$$P(A \cup B) = \frac{[n(A) - n(A \cap B)] + n(A \cap B) + [n(B) - n(A \cap B)]}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$$

$$P(A) + P(B) - P(A \cap B)$$

11.5.1 Addition Theoremof Probability for Mutually Exclusive Events:

If the events A and B are mutually disjoint, i.e., if $P(A \cap B) = \phi$ then

$$P(A \cap B) = \frac{n(A \cap B)}{N} = \frac{n(\phi)}{N} = 0 \text{ ,.....(3)}$$

because $n(\phi)=0$, as a null set does not contain any sample point. In case of disjoint events, $A \cup B$ represents the happening of anyone of the events A and B. Hence, substituting from (3) in (1) we get the addition theorem as follows:

Theorem 2: The probability of happening of any one of the two mutually disjoint events is equal to the sum of their individual probabilities. Symbolically, for disjoint events A and B

$P(A \cup B) = P(A) + P(B)$..............(4)

Generalisation of (1) equation

For three events A, B and C, the probability of the occurence of at least one of them is given by

$$P(A \cup B \cup C) = \frac{n\,(A \cup B \cup C)}{N}$$

$$= \frac{1}{N}\left[n(A) + n(B) + n(C) - n(A \cap B) - n(B \cap C) - n(A \cap C) + n(A \cap B \cap C)\right]$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} + \frac{n(C)}{N} - \frac{n(A \cap B)}{N} - \frac{n(B \cap C)}{N} - \frac{n(A \cap C)}{N} + \frac{n(A \cap B \cap C)}{N}$$

$$= P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C) \text{ ...(5)}$$

In particular, if A, B and C are mutually exclusive (disjoint), then

$$A \cap B = A \cap C = B \cap C = \phi \text{ and } A \cap B \cap C = \phi$$

$$\Rightarrow \quad n(A \cap B) = n(A \cap C) = n(B \cap C) = n(A \cap B \cap C) = 0$$

Hence, substituting in (5), the probability of occurence of any one of the mutually exclusive events A, B and C is equal to the sum of their individual probabilities given by :

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$......(6)

In general, if $A_1, A_2, ..., A_n$ are mutually exclusive then

$P(A_1 \cup A_2 \cup .... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$...... (7)

i.e., the probability of occurence of any of the n mutually disjoint events A_1, $A_2, ..., A_n$ is equal to the sum of their individual probabilities.

11.5.2 Theorem of compound probability or multiplication theorem of probability:

Theorem: The probability of simultaneous happening of two events A and B is given by :

$$\left.\begin{array}{ll} & P(A\cap B)=P(A)P(B/A); P(A)\neq 0 \\ \text{or} & P(B\cap A)=P(B)P(A/B); P(B)\neq 0 \end{array}\right\} ....(1)$$

where P(B/A) is the conditional probability of happening of B under the condition that A has happened and P(A/B) is the conditional probability of happening of A under the condition that B has happened.

In other words, the probability of the simultaneous happening of the two events A and B is the product of two probabilities, namely: the probability of the first event times the conditional probability of the second event, given that the first event has already occured. We may take any one of the events A or B as the first event.

Proof: Let A and B be the events associated with the sample space S of a random experiment with exhaustive number of outcomes (sample points) N, i.e. n(s)=N. Then by definition:

$$P(A\cap B)=\frac{n(A\cap B)}{n(s)} \qquad ....(2)$$

For the conditional event A/B (i.e., the happening of A under the condition that B has happened), the favourable outcomes (sample points) must be out of the sample points of B. In other words, for the event A/B, the sample space is B and hence

$$P(A/B)=\frac{n(A\cap B)}{n(B)} \; ....(3)$$

Similarly, we have

$$P(B/A)=\frac{n(B\cap A)}{n(A)} \; ....(4)$$

Rewriting equ. (2), we get

$$P(A \cap B) = \frac{n(A)}{n(s)} x \frac{n(A \cap B)}{n(A)} = P(A).P(B/A) \quad \{\text{from } (4)\}$$

Also $P(A \cap B) = \frac{n(B)}{n(s)} x \frac{n(A \cap B)}{n(B)} = P(B).P(A/B) \quad \{\text{from } (3)\}$

Generalisation of Multliplication Theorem of Probability:

The multiplication theorem of probability can be extended to more than two events. Thus, for three events A_1, A_2, and A_3, we have

$$P(A_1 \cap A_2 \cap A_3) = P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2) \ldots\ldots\ldots\ldots(5)$$

In general for n events $A_1, A_2, \ldots, A_n$, we have

$$P(A_1 \cap A_2 \cap \ldots\ldots \cap A_n) = P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2)x\ldots\ldots x_1P(A_n/A_1 \cap A_2 \cap \ldots\ldots \cap A_n - 1)$$

Multiplication Theorem for Independent Events :

Two events A and B are independent if and any if

$$P(A \cap B) = P(A).P(B) \ldots\ldots\ldots\ldots(6)$$

i.e., if the probability of the simultaneous happening of two events is equal to the product of their individual probabilities.

Proof : For two events A and B, we have

$$P(A \cap B) = P(A).P(B/A) = P(B/A) = P(B).P(A/B) \ldots\ldots\ldots\ldots(7)$$

If part : If A and B are independent then

$$P(A/B) = P(A) \text{ and } P(B/A) = P(B) \ldots\ldots\ldots\ldots(8)$$

Subtituting (7), we get

$$P(A \cap B) = P(A).P(B)$$

Only if past : If equation (6) holds, then using equation (7), we get

$$P(B/A) = P(B) \text{ and } P(A/B) = P(A)$$

$\Rightarrow$ A and B are independent

Hence, $P(A \cap B) = P(A).P(B)$,

provides a necessary and sufficient condition for the independence of two events A and B.

By this we mean that if A and B are independent events, then equation (6) holds and conversely, if equ. (6) holds, then A and B are independent events.

11.5.3 Some Other Theorems :

Theorem 1: $P(\overline{A}) = 1 - P(A)$ ....(1)

Proof: We have, $A \cup \overline{A} = s$

$\Rightarrow \quad P\left(A \cup \overline{A}\right) = P(s)$

$\Rightarrow \quad P(A) + P(\overline{A}) = P(s)$

by addition theorem of probability (or by axiom of additivity), since A and $\overline{A}$ are mutually disjoint. Further,

$$P(s) = \frac{n(s)}{n(s)} = 1$$

substituting in equation (7), we get

$P(A) + P(\overline{A}) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$

Theorem 2: i) $P(\overline{A} \cap B) = P(B) - P(A \cap B)$ ....(2)

ii) $P(A \cap \overline{B}) = P(A) - P(A \cap B)$ ....(3)

Proof: From the Venn diagram, (3.2) it is obvious that the events A and B can be expressed as disjoint unions as given below :

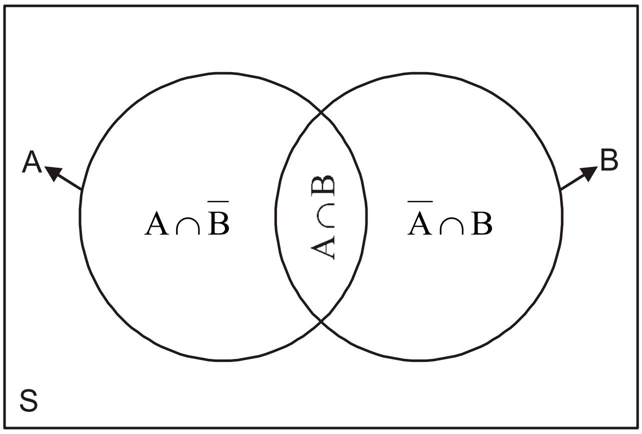


Fig 3.2

$$A = (A \cap \overline{B}) \cup (A \cap B) \text{ ....(4)}$$

B $B = (A \cap B) \cup (\overline{A} \cap B) \text{ ....(5)}$

Hence, by the axiom of addivity or by the addition theorem of probability for mutually disjoint events, we get from (4)

$$P(A) = P\left[(A \cap \overline{B}) \cup (A \cap B)\right]$$

$$= P(A \cap \overline{B}) \cup P(A \cap B)$$

$$\Rightarrow \qquad P(A \cap \overline{B}) = P(A) - P(A \cap B)$$

which is the result equation (3), similarly from equation (5), we get

$$P(B) = P(A \cap B) + P(\overline{A} \cap B)$$

$$\Rightarrow \qquad P(\overline{A} \cap B) = P(B) - P(A \cap B)$$

which is the result equation (2)

Theorem 3: If $A \subset B$, then $P(A) \leq P(B)$

Proof : Let $A \subset B$. Then from the fig. 3.3, the event B can be expressed as disjoint union of the two events A and $\overline{A} \cap B$, i.e.,

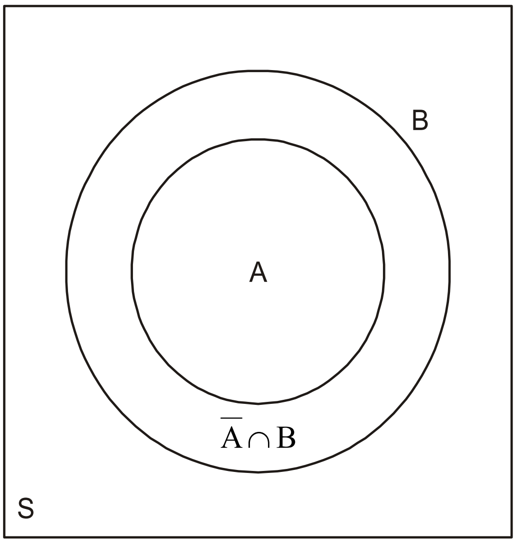


Fig 3.2

$$B = A \cup \left(\overline{A} \cap B\right)$$

since A and $\overline{A} \cap B$ are disjoint, by addition theorem of probability, we get

$$P(B) = P(A) + P(\overline{A} \cap B)$$

$\Rightarrow \quad P(B) - P(A) = P(\overline{A} \cap B) \leq 0 \, [\because p(E) \leq 0 \text{ for every } E \subset S]$

$\Rightarrow \quad P(B) \geq P(A) \Rightarrow P(A) \leq P(B)$, as desired.

Theorem 4: If events A and B are independent then the events (i) A and $\overline{B}$ are independent; (ii) $\overline{A}$ and B are independent; (iii) $\overline{A}$ and $\overline{B}$ are independent.

Proof : Since the events A and B are independent, we have

$P(A \cap B) = P(A)P(B)$..........(6)

i) $P(A \cap \overline{B}) = P(A) - P(A \cap B)$

$= P(A) - P(A)P(B)$ (from equation ....(6))

$= P(A)\,[1 - P(B)]$

$= P(A)\,P(\overline{B})$

$\Rightarrow$ A and $\overline{B}$ are independent

ii) $P(\overline{A} \cap B) = P(B) - P(A \cap B)$

$= P(B) - P(A)P(B)$ (from equation ....(6))

$= P(B)\,[1 - P(A)]$

$= P(B)\,P(\overline{A})$

$\Rightarrow$ $\overline{A}$ and B are independent

iii) $P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B) - P(A \cap B)]$

$= 1 - P(A) - P(B) + P(A)P(B)$ (using equation (6))

$= [1 - P(A)] - P(B)\,[1 - P(A)]$

$= [1-P(A)\,]\,[1-P(B)]$

$= P(\overline{A}).P(\overline{B})$

$\Rightarrow$ $\overline{A}$ and $\overline{B}$ are independent events

Theorem 5: If A_1,......,A_2,........,A_n are independent events with respective probabilities of occurence p_1, p_2....,p_n then the probability of occurence of at least one of them is given by:

$$P(A_1 \cup A_2 \cup .... \cup A_n) = 1 - (1 - p_1)(1 - p_2)......(1 - p_n) \text{ ....(7)}$$

Proof: We are given:

$$P(A_i) = P_i \Rightarrow P(\overline{A_i}) = 1 - P_i \text{ .... (i)}$$

we know that for any event E, $P(E) + P(\overline{E}) = 1$..........(ii)

Taking $E = A_1 \cup A_2 \cup .... \cup A_n$ in (ii), we get

$$P(A_1 \cup A_2 \cup .... \cup A_n) + P(A_1 \cup A_2 \cup .... \cup A_n)^c = 1$$

$$\Rightarrow \quad P(A_1 \cup A_2 \cup .... \cup A_n) + P(\overline{A_1} \cap \overline{A_2} \cap .... \cap \overline{A_n})^c = 1 \quad ....(8)$$

[By De-Morgan's law of complementation i.e., the complement of the union of sets is equal to the intersection of their complements]

$$\Rightarrow \quad P(A_1 \cup A_2 \cup .... \cup A_n) = 1 - P(\overline{A_1} \cap \overline{A_2} \cap .... \cap \overline{A_n}) \quad ....(9)$$

$$1 - P(\overline{A_1})(\overline{A_2})...P(\overline{A_n})$$

by compound probability theorem, since A_1, A_2,....,A_n and consequently $\overline{A_1}, \overline{A_2}, ..., \overline{A_n}$ are independent (c.f. theorem 4).

Hence substituting from (i), we get

$$P(A_1 \cup A_2 \cup .... \cup A_n) = 1(1 - P_1)(1 - P_2)....(1 - P_n)$$

Now, let take some numerical problems.

Example 1: A committee of 4 persons is to be a appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner.

i) There must be one from each category

ii) It should have at least one from the purchase department

iii) The chartered accountant must be in the committee.

Solution : There are in all 3 + 4 + 2 + 1 = 10 people. A committee of 4 can be formed out of these 10 people in $^{10}C_4$ ways. Hence the exhaustive number of cases is :

$$\overset{10}{\underset{4}{C}} = \frac{10 \text{ x } 9 \text{ x } 8 \text{ x } 7}{4!} = 210$$

i) The number of favourable cases for the committee to consist of one number from each category (Production, Purchase, Sales & C.A.) is :

$$ {}^{3}C_{1} \text{ x } {}^{4}C_{1} \text{ x } {}^{2}C_{1} \text{ x } {}^{1}C_{1} = 3 \text{ x } 4 \text{ x } 2 \text{ x } 1 = 24 $$

$$ \therefore \qquad \text{Required probability } = \frac{24}{210} = \frac{4}{35} = 0.1143 $$

ii) The probability ‘p’ that the committee of 4 has at least one member from the purchase department is given by :

p = P[1 from purchase department and 3 others] + P[2 from purchase department and 2 others] + P[3 from purchase department and 1 other] + P[4 from purchase department]

$$ = \frac{{}^{4}C_{1} \text{ x } {}^{6}C_{3}}{{}^{10}C_{4}} + \frac{{}^{4}C_{2} \text{ x } {}^{6}C_{2}}{{}^{10}C_{4}} + \frac{{}^{4}C_{3} \text{ x } {}^{6}C_{1}}{{}^{10}C_{4}} + \frac{{}^{4}C_{4}}{{}^{10}C_{4}} $$

$$ = \frac{1}{210}\left[4 \text{ x } \frac{6 \text{ x } 5 \text{ x } 4}{3!} + \frac{4 \text{ x } 3}{2!} \text{ x } \frac{6 \text{ x } 5}{2!} + 4 \text{ x } 6 + 1\right] $$

$$ = \frac{1}{210}\left(80 + 90 + 24 + 1\right) = \frac{195}{210} = 0.9286 $$

iii) The probability p_1 that the charted accountant must be in the committee of 4 is given by :

$$ p_1 = \text{P[Chartered Accountant and 3 others]} = \frac{{}^{1}C_{1} \text{ x } {}^{9}C_{3}}{{}^{10}C_{4}} $$

$$ = \frac{9 \text{ x } 8 \text{ x } 7}{3!} \text{ x } \frac{4!}{10 \text{ x } 9 \text{ x } 8 \text{ x } 7} = \frac{4}{10} = 0.4 $$

Example 2:

A card is drawn from a well shuffled pack of playing cards. Find the probability that it is either a diamond or a king.

Solution:

Let A denote the event of drawing a diamond and B denote the event of drawing a king from a pack of cards. Then we have

$P(A) = \frac{13}{52} = \frac{1}{4}$ and $P(B) = \frac{4}{52} = \frac{1}{13}$ and we want $P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{4} + \frac{1}{13} - P(A \cap B) \qquad ....(1)$$

There is only one case of favourable to the event $A \cap B$ viz., king of diamond. Hence, $P(A \cap B) = \frac{1}{52}$ substituting in 1, we get

$$P(A \cup B) = \frac{1}{4} + \frac{1}{13} - \frac{1}{52} = \frac{13+4-1}{52} = \frac{16}{52} = \frac{4}{13}$$

Laws and Axioms of Probability

Let S be a sample of a random experiment. If to each event A of the set of all possible events of S, we associate a real number P(A), then P(A) is called "probability" of event A, if the following axioms hold:

Axioms 1 : For every event A

$$P(A) \geq 0 \qquad ....(1)$$

Axiom 2: For the sure event s.

$$P(S) = 1 \qquad ....(2)$$

Axiom 3 : For any finite number or countability infinite number of mutually exclusive events A_1, A_2,..... of S

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...(3)$$

In particular, for two mutually exclusive events A and B

$$P(A \cup B) = P(A) + P(B) \qquad ....(4)$$

It should be noted that we can speak of the 'probability' only if the event is a subset of a specified sample space s, and to each subset of S a real number, satisfying the axioms can be assigned. A sample space on which 'probability' has been defined is called a probability space.

Deducation from the axioms

Theorem I : The probability of the impossible event is zero

$P(\phi) = 0$ ....(5)

Proof : Any event A and the impossible event ϕ are mutually exclusive

Also $A \cup \phi = A$ Hence, by Axiom 3,

$P(A) = P(A \cup \phi) = P(A) + P(\phi)$

$\therefore$ $P(\phi) = 0$

Theorem II: The probability of the complementary event is

$P(A^1)=1 - P(A)$ .... (6)

Proof : A and A^1 are mutually exclusive events, and

$A \cup A^1 = S$

Hence, $P(S) = P(A \cup A^1) = P(A) + P\left(A^1\right)$, by Axiom 3

i.e., $1 = P(A) + P(A^1)$,

$\therefore$ $P(A^1) = 1- P(A)$

Theorem III: The probability of an event lies between 0 and 1

$0 \leq P(A) \leq 1$ ....(6)

Proof : By Axiom 1, $0 \leq P(A)$. Also, from equation (6), we have $P(A)=1-P(A^1)$. Since $P(A^1)$ is a probability, it cannot be negative (Axiom 1); therefore $P(A) \leq 1$. Combining both the inequalities, the result follows.

Theorem IV: If $A = A_1 \cup A_2 \cup ...... \cup A_n$, where $A_1, A_2, ... A_n$ are mutually exculsive events, then

$P(A) = P(A_1) + P(A_2) + ... + P(A_n)$ ....(8)

In particular, if A = S, the sample space, then

$$P(A_1) + P(A_2) + ... + P(A_n) = 1 \qquad ....(9)$$

Proof: Relation equation (8) follows from Axioms 3 for the case when the number of mutually exclusive events is finite, say n. In addition, using Axiom 2, the result (equation 9) follows :

Theorem V: If $A \subseteq B$ (i.e. event A implies event B), then

$$P(A) \leq P(B) \qquad ....(10)$$

Proof: If $A \subseteq B$, then events A and $A^1 \cap B$ are mutually exclusive, and their union $A \cup (A^1 \cap B) = B$

Hence, by Axiom 3,

$$P(B) = P(A) + P(A^1 \cap B)$$

Since by Axiom 1, $P(A^1 \cap B)$ cannot be negative,

Hence $\quad P(B) \geq P(A)$

Theorem IV: For any two events A and B

$$P(A) = P(A \cap B) + P(A \cap B^1) \; ....(11)$$

$$P(B) = P(A \cap B) + P(A^1 \cap B) \; .... (12)$$

Proof: Event $A \cap B$ and $A \cap B^1$ are mutually exclusive and their union is the event = A. Hence by Axiom 3, relation (equation 11) can be proved. Similarly, event B is the union of mutually exclusive events $A \cap B$ and $A^1 \cap B$, and result (equation 12) can be proved

Theorem VII: For any two events A and B (which may or way not be mutually exclusive)

$$P(A \cap B) = P(A) + P(B) - P(A \cap B) \; ....(13)$$

Proof: Events $A \cap B^1$, $A \cap B$ and $A^1 \cap B$ are mutually exclusive, and their union is the event. $A \cup B$. Hence, by Axiom 3

$$P(A \cup B) = P(A \cap B^1) + P(A \cap B) + P(A^1 \cap B)$$

But from (equation 11) and (equation 12), we have

$$P(A\cap B^1) = P(A) - P(A\cap B) \text{ and } P(A^1\cap B) = P(B) - P(A\cap B)$$

Substituting these values,

$$P(A\cap B) = [P(A) - P(A\cap B)] + P(A\cap B) + [P(B) - P(A\cap B)]$$

$$= P(A) + P(B) - P(A\cap B)$$

Theorem VIII: For any three events A, B, C

$$P(A\cup B\cup C) = P(A) + P(B) + P(C) - P(A\cap B) - P(A\cap C) - P(B\cap C) + P(A\cap B\cap C)$$

...........(14)

Proof: In equation (13) let us replace the event B by $B\cup C$.

Then, $P[A\cup(B\cup C)] = P(A) + P(B\cup C) - P[A\cap(B\cup C)]$

$$= P(A) + [P(B) + P(C) - P(B\cap C)] - P[(A\cap B)\cup(A\cap C)],$$

by distributive law.

Again using (equation 13) for the union of events $A\cap B$ and $A\cap C$,

$$P[(A\cap B)\cup(A\cap C)] = P(A\cap B) + P(A\cap C) - P[(A\cap B)\cap(A\cap C)]$$

$$= \; = P(A\cap B) + P(A\cap C) - P[A\cap B\cap C]$$

Hence the result follows.

Techniques of Counting

Some mathematical methods are shown below, which are often helpful for determining without direct enumeration the number of outcomes of a random experiment or the number of cases favourable to an event. These are referred to as "Combinational Methods".

1. **Fundamental principle of counting:** If several processes can be performed in the following manner: the first process in 'p' ways, the second 'q' ways, the third in r ways, and so on then the total numnber of ways in which the whole process can be performed in the order indicated is given by the product

p.q.r .... (1)

2. **Permutation** : The total number of ways of arranging (called permutation) 'n' distinct objects taken 'r' at a time is given by

$$^{n}p_{r} = n(n-1)\ (n-2)\ ....(n - r + 1) \qquad ....(2)$$

3. **Arrangement in a line or circle:** The total number of way in which 'n' distinct objects can be arranged among themselves is

i) in a line $n\,! = 1.\ 2.\ 3....n$ ....(3)

ii) in a circle $(n - 1)\,!$ ....(4)

4. **Permutation with Repition** : The number of ways of arranging n objects, among which p are alike, q are alike, r and alike, etc. is

$$\frac{n!}{p!q!r!...} \qquad ....(5)$$

5. **Combination** : The total number of possible groups (called combination) that can be formed by taking r object out of n distinct objects is given by

$$\overset{n}{\underset{r}{C}} = \frac{n(n-1)(n-2)...(n-r-+1)}{r!}$$

$$= \frac{n!}{r!(n-r)!} \qquad ....(6)$$

6. **Combination (any number at a time)** : The total number of ways of forming groups by taking any numbers from n distinct objects i

$$\overset{n}{\underset{1}{C}} + \overset{n}{\underset{2}{C}} + \overset{n}{\underset{3}{C}} + ... + \overset{n}{\underset{n}{C}} = 2^{n} - 1 \qquad ....(7)$$

7. **Choosing balls from an Urn** : The total number of ways of choosing 'a' white balls and 'b' black balls from an urn containing A white and B black balls is

$$\overset{A}{\underset{a}{C}}.\overset{B}{\underset{b}{C}} \qquad ....(8)$$

This may be extended to more than two categories of balls.

8. **Ordered Partitions (Distinct Objects)**: The total number of ways of

distributing 'n' distinct objects into 'r' compartments marked 1, 2,....,r is

r^n ....(9)

The number of ways in which the 'n' objects can be distributed so that the compartments contain respectively $n_1, n_2, \dots n_2$ object is

$$= \frac{n!}{n_1! n_2! \dots n_r!} \qquad \text{...............(10)}$$

9. **Ordered partitions (Identical objects)**: The total number of ways of distributing 'n' identical objects into 'r' compartments marked 1, 2,...,r is

$$\overset{n+r-1}{\underset{r-1}{C}} \qquad \text{..........(11)}$$

if none of the compartment should remain empty, the total number of ways of distributing the balls is

$$\overset{n-1}{\underset{r-1}{C}} \qquad \text{..........(12\_}$$

10. **Sum of points on the dice** : When 'n' dice are thrown, the number of ways of getting a total of 'r' points is given by the

Coefficient of x^r in $(x + x^2 + x^3 + x^4 + x^5 + x^6)^n$ ....(13)

11. **Dearrangements and Matches** : If 'n' objects numbered 1, 2, 3, ...,n are distributed at random in 'n' places also numbered 1, 2, 3, ..., n, a "match" is said to occur if an object occupies the place corresponding to its number. The number of permulations in which no match occurs is

$$tn = n!\left\{1 - \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots\dots + (-1)^n \frac{1}{n!}\right\} \qquad \text{......(14)}$$

This is also known as "derrangement"

$$\overset{n}{\underset{r}{C}} t_{n-r} = \frac{n!}{r!}\left\{1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots\dots + \frac{(-1)^{n-r}}{(n-r)!}\right\} \qquad \text{..........(15)}$$

11.6 CONDITIONAL PROBABILITY[1] :

The multiplication theorem explained above is not applicable in case of dependent events. Two events A and B are said to be dependent when B can occur only when A is known to have occured (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P(A/B) or, in other words, probability of A given that B has occured.

If two events A and B are dependent, then the conditional probability of B given A is :

$$P(B/A) = \frac{P(AB)}{P(A)}$$

Proof: Suppose a_1 is the number of cases for the simultaneous happening of A and B out of $a_1 + a_2$ cases in which A can happen with or without happening of B.

$$\therefore \qquad P(B/A) = \frac{a_1}{a_1 + a_2} = \frac{a_{1/n}}{(a_1 + a_2)_{/n}} = \frac{P(AB)}{P(A)}$$

Similarly it can be shown that :

$$P(A/B) = \frac{P(AB)}{P(B)}$$

The general rule of multiplication in its modified form in terms of conditional probability becomes :

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

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

For three events A, B and C, we have

P(ABC) = P(A) x P(B/A) x P(C/AB)

[1] Footnote : When we are computing the probability of a particular event A, given information about the occurence of another event B, this probability is referred to as conditional probability.

i.e. the probability of occurence of A, B and C is equal to the probability of A, times the probability of B given that A has occured, times the probabilityof C given that both A and B have occured.

Example 1: A bag cotains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.

Solution : Probability of drawing a black ball in the first attempt is

$$P(A) = \frac{3}{5+3} = \frac{3}{8}$$

Probability of drawing the second black ball given that the first ball drawn is black

$$P(B/A) = \frac{2}{5+2} = \frac{2}{7}$$

The probability that both balls drawn are black is given by

$$\therefore \quad P(AB) = P(A) \text{ x } P(B/A) = \frac{3}{8} \text{x} \frac{2}{7} = \frac{3}{28}$$

11.7 BAYES' THEOREM AND ITS APPLICATIONS :

The Bayes' theorem named after the British mathematician Rev. Thomas Bayes (1702-61) and published in 1763 in a short paper has become one of the most famous memories in the history of science and one of the most controversial.

Theorem : One of the most interesting applications of the results of probability theory involves estimating unknown probability and making decisions on the basis of new (sample) information. Since World War II, a considerable body of knowledge has developed known as Bayesian decisions theory whose purpose is the solution of problems involving decision-making under uncertainity.

The concept of conditional probability discussed above takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to "revise" probabilities based on new information and to determine the probability that a particular effect was due to a specific cause. The procedure for revising these probabilities is known as Bayes' theorem.

Bayes contribution consists primarily of a unique method for calculating conditional probabilities. The so-called "Bayesian" approach to this problem addresses itself to the question of determining the probability of some event. A given that another event, B, has been (or will be) observed, i.e, determining the value of P(A/B). The event A is usually thought of as sample information so that Bayes' rule is concerned with determining the probability of an event given certain sample information. For example, a sample output of 2 defectives in 50 trials (event A) might be used to estimate the probability that a machine is not working correctly (event B) or you might use the results of your first examination in statistics (event A) as sample evidence in estimating the probability of getting a first class (event B).

Baye's Theorem :- Baye's Theorem was given by British mathematician, Thomas Bayes in 1763 which provides a means for making the probability calculations. It is basically an application of conditional probability.

The conditional probability of P(Bi/A) of a specified event Bi when Ai is stated to have acheally occurred is given by

$$P(Bi/A) = \frac{P(Bi).P(A/Bi)}{\sum\limits_{i=1}^{n} P(Bi).P(A/Bi)}$$

Proof :-

The event A can occur in 'n' mutually exclusive ways $B_1A, B_2A, B_3A, \ldots, B_nA$

i.e. either when B_1 has occurred or B_2 or ..... B_n

so, by the theorem of total probability .

$$P(A) = P(B_1A + B_2A + \ldots + B_nA)$$
$$= P(B_1A) + P(B_2A) + \ldots + P(B_nA)$$

We know that $P(A/Bi) = \frac{P(B_1A)}{P(B_1)}$

$= P(B_1A) = P(A/B_1)P(B_1)$

$= P(A) = P(B1).\ P(A/B1) + P(B2).P(A/B2) + \ldots\ldots + P(Bn).P(A/Bn)$

$= P(A) = \sum_{i=1}^{n} P(Bi).P(A/Bi)$

Again, P(ABi) = P(A).P(Bi/A)

& P(BiA) = P(A/Bi).P(Bi)

Since the events ABi and BiA are equivalent So, their probabilities are also equal.

i.e. P(ABi) = P(BiA)

= P(A).P(Bi/A) = P(Bi).P(A/Bi)

$$P(Bi/A) = \frac{P(Bi).P(A/Bi)}{P(A)}$$

Substituting the value of 1 [i.e. P (A)]

$= P(Bi/A) = \frac{P(Bi).P(A/Bi)}{\sum_{i=1}^{n} P(Bi).P(A/Bi)}$ which is Baye's Theorem

Some interesting points worth noting about Bayes' theorem are:

1. Though it deals with a conditional probability, its interpretation is different from that of the general conditional probability theorem. The general conditional probability theorem asks, "what is the probability of the sample as experimental results given the state value" ? Whereas Baye's theorem ask : "What is the probability of the event given the sample or experimental result ?

2. When we talk of Bayes' theorem, different decision-makes may assign different probability to the same set of states of nature. Also we may conduct a new experimnent by using posterior probabilities of the preceding experiment as prior probabilities. As we proceed with repeated experiments, evidence accumulates and modifies the intitial prior probabiities, thereby modifying the intensity of a decision-maker's belief in various states of nature. In other words, the more evidence we accumulate the less important are the prior probabilities.

3. The notions of "prior" and "posterior" in Bayes' theorem are relative to a given sample outcome. That is, if a posterior distribution has been determined from a particular sample, this posterior distribution would be considered the prior distribution relative to a new sample.

The following exsample shall illustrate the application of Bayes' theorem:

Example 1: Assume that a factory has two machines. Past records show that machine 1 produces 30% of the items of output and machine 2 produces 70% of the items. Further, 5% of the items produced by machine 1 were defective and only 1% produced by machine 2 were defective. If a defective item is drawn at random, what is the probability that the defective item was produced by machine 1 or machine 2 ?

Solution : Let A_1 = the event of drawing an item produced by machine 1,

A_2 = The event of drawing an item produced by machine 2, and

B = The event of drawing a defective item produced either by machine 1 or machine 2.

Then from the first information,

$P(A_1) = 30\% = 0.30$, $P(A_2) = 70\% = 0.70$

from the additional information

$P(B/A_1) = 5\% = 0.05$, $P(B/A_2) = 1\% = 0.1$

The required value are tabuled below :

Computation of Posterior Probabilities

Events	Prior probability	Conditional probabbility B given event A $P(BA_i)$	Joint probability $A(A_1$ and B) (2) x (3)	Posterior (revised) probability $P(A_1/B)$ (4) ÷ P(B)
A_1	0.30	0.05	0.015	0.015/0.022=0.682
A2	0.70	0.01	0.007	0.007/0.022=0.318
Total	1.00		P(B) = 0.022	1.000

Without the additional information, we may be inclined to say that the defective item is drawn from machine 2 output since $P(A_2) = 70\%$ is larger than $P(A_1) = 30\%$ with the additional information, we may give a better answer. The probability that the defective item was produced by machine 1 is 0.682 or 68.2% and that by machine 2 is only 0.318 or 31.8%. We may now say that the defective item is more likely drawn from the output produced by machine 1.

The above answer may be checked by actual number of items as follows:

If 10,000 items were produced by the two machines in a given period, the number of items produced by machine 1 is

$$10,000 \text{ x } 30\% = 3,000$$

and the number of items produced by machine 2 is

$$10,000 \text{ x } 70\% = 7,000$$

The number of defective items produced by machine 1 is

$$3,000 \text{ x } 5\% = 150$$

and the number of defective items produced by machine 2 is :

$$7,000 \text{ x } 1\% = 70$$

The probability that a defective item was produced by machine 1 is

$$= \frac{150}{150+70} = 0.682$$

and by machine 2 is

$$= \frac{70}{150+70} = 0.318$$

Example 2 : A manufacturing firm produces units of a product in four plants. Define event A_i : a unit is produced in plant i, i = 1, 2, 3, 4 and event B: a unit is defective. From the past records of the proportions of defectives produced at each plant the following conditional probabilities are set :

$$P(B/A_1) = 0.05,$$

$$P(B/A_2) = 0.10,$$

$P(B/A_3) = 0.15,$

$P(B/A_4) = 0.02,$

The first plant produces 30 percent of the units of the product, the second plant 25 percent, third plant 40 percent and the fourth plant 5 percent. A unit of the product made at one of these plants is tested and is found to be defective. What is the probability that the unit was produced in plant 3 ?

Solution : We have to determine $P(A_3/B)$. From the general form of the Baye's theorem, the probability is given by

$$P(A_3/B) = \frac{P(B/A_3)P(A_3)}{\sum_{i=1}^{4} P(B/A_i)P(A_i)}$$

The computation of $P(A_3/B)$ is shown below :

Plant Event	P(A)	$P(B/A_i)$	$P(A_i)\ P(B/A_i)$	$\frac{P(A_i)\ P(B/A_i)}{P(A_i)\ P(B/A_i)}\sum_{i=1}^{4}$
1	0.30	0.05	0.015	0.015/0.101 = 0.1485
2	0.25	0.10	0.025	0.025/0.01 = 0.2475
3	0.40	0.15	0.060	0.06/0.101= 0.5941
4	0.05	0.02	.001	.001/0.101 = 0.0099
	1.00		0.101 = P(B)	1.00

It is clear from the table that $P(A_1B) = 0.5941$. The 4th column when summed gives that denominator of Bayes' theorem from the table $P(B) = 0.101$, i.e., the probability that a defective part is produced by this firm is 0.101.

It is clear from the above illustrations that Bayes' theorem provides a powerful method in improving the quality of probability for aiding the management in decision-making under uncertainity. As we proceed with repeated experiments, evidence accumulates and modifies the initial prior probabilies and thereby, modifies the intensity of a decision-maker's belief in various hypotheses. Repeated estimates will soon produce such low posterior probabilities for some hypotheses that they can be eliminated from further consideration. In other words, the more evidence we accumulate, the less important are the prior probabilities. The only restriction on the application of Bayesian rule is that all hypotheses must enable in a given situation and that none is assigned a prior probability of 0 or 1.

11.8 LET US SUM UP:

Today the probability theory has been developed to a great extent and there is not even a single discipline in social, physical or natural sciences where probability theory is not used. It is extensively used in the quantitative analysis of business and economic problems. It is an essential tool in statistical inference and forms the basis of the 'Decision Theory', viz., decision making in the face of uncertainty with calculated risks.

11.9 LESSON END EXERCISE :

Q1. What do you understand by conditional probability ? If

Prob. (A+B) = Prob. A + Prob. B,

are the two events A and B statistically independent ?

Q2.a) Define independent events.

b) Obtain the necessary and sufficient condition for the independence of two events A and B generalise the result to 'n' events $A_1, A_2, ... A_n$.

Q3. The result of an examination given to a class on three papers A, B and C are

40% failed in paper A; 30% failed in B;

25% failed in paper C; 15% failed in paper A and B both;

12% failed in B and C both; 10% failed in paper A and C both;

3% failed in all the the three A, B and C

What is the probability of a randomly selected candidates passing in all the three paper ?

Q4. An box contains 6 red 4 white and 5 blue balls. From this box 3 balls are drawn in succession. Find the probability that they are drawn in the order red, white and blue if each balls is

i) replaced ii) not replaced

Ans. i) 8/225 ii) 4/91

Q5. A group of 200 persons was classified according to age and sex as given below:

Age in years	Male	Female	Total
Below 30	60	50	110
30 and above	80	10	90
Total	140	60	200

i) What is the probability that a randomly chosen person from this group is a male below 30 years of age ?

ii) What is the probability that a person is below 30 years of age, given that he is a male ?

****

M.A. Economics	Semester - IInd	Lesson - 12
Course No. 203		Unit - III

RANDOM VARIABLE, PROBABILITY, MASS AND DENSITY FUNCTIONS, EXPECTATIONS

CHAPTER HIGHLIGHTS:

This lesson contains the information regarding the meaning of random variable, mathematical expectation and random variable. Alongwith this moments about zero and mean are also discussed below.

CHAPTER HIGHLIGHTS:

12.1 CONCEPT OF RANDOM VARIABLE:

Meaning : Intuitively, by a random variable (r.v.) we means a real number x associated with the outcomes of a random experiment. It can take any one of the various possible values each with a definite probability. For example, in a throw of a die, if x denotes the number obtained, then x is a random variable which can take any one of the values 1, 2, 3, 4, 5 or 6, each with equal probability 1/6. Similarly, in toss of a coin if x denotes the number of heads, then x is a random variable which can take any one of the two values; O (No head, i.e., tail) or 1 (i.e, head), each with equal probability $\frac{1}{2}$.

Let us now consider a random experiment of three tosses of a coin (or three coins tossed simultaneously). Then the sample space S consists of $2^3 = 8$ points as given below :

$$S = \{(H,T) \text{ x } (H,T) \text{ x } (H,T)\}$$

$$= \{(HH, HT, TH, TT) \text{ x } (H,T)\}$$

$$= \{HHH, HTH, THH, TTH, HHT, HTT, THT, TTT\}$$

Let us consider the variable x, which is the number of heads obtained. Then x is a random variable which can take any one of the values 0, 1, or 2.

Outcome :	HHH	HTH	THH	TTH	HHT	HTT	THT	TTT
Values of x :	3	2	2	1	2	1	1	0

If the the sample points in the above order be denoted by $w_1, w_2, w_3, \ldots, w_8$ then to each outcome w of the random experiments, we can assign a real number x = x(w). For example,

$$x(w_1) = 3, \; x(w_2) = 2, \; x(w_3) = 2, \ldots\ldots, \; x(w_8) = 0.$$

Thus, rigorously speaking, random variable may be defined as a real values function on the sample space, taking values on the real line $R(-\infty, \infty)$. In other words, random variable is a function which takes real values which are determined by the outcomes of the random experiment.

Remarks :

1. A random variable is denoted by the capital letter x, y, z,..., etc of the English alphabet and particular values which the random variable takes

are denoted by the corresponding small letters of the English alphabet.

2. It should be clarly understood that the actual values which the event assumes is not a random variable. For example, in three tosses of a coin, the number of heads obtained is a random variable which can take any one of the three values 0, 1, 2, or 3 as long as the coin is not tossed. But, after it is tossed and we get two heads, then 2 is not a random variable.

12.2 RANDOM VARIABLE AND PROBABILITY DISTRIBUTION :

A variable whose value is determined by the outcome of a random experiment is called a random variable. A random variable is also known as a chance variable or stochastic variable. A random variable may be discrete or continuous. If the random variable takes on the integer values such as 0, 1, 2... then it is called a discrete random variable. The number of printing mistakes in each page of a book, the number of telephone calls received by the telphone operator of a firm are examples of discrete random variable. If the random variable takes on all values, within a certain interval, then the random variable is called a continuous random variable. The amount of rainfall on a rainy day or in a rainy season, the height and weight of individuals are examples of continuous random variable.

In terms of symbols if a variable X can assume discrete set of values x_1, x_2...x_k with respective probabilities p_1, p_2....p_k where $p_1 + p_2 +....+ p_k = 1$, we say that a discrete probability distribution for x has been defined. The function P(x) which has the respective values p_1, p_2,... for $x = x_1, x_2,...,x_k$ is called the probability function or frequency function of x.

The probability distribution of a pair of fair dice tossed is given below :

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Where x denotes the sum of the points obtained. For example, the probability of getting sum 4 is 3/36. Thus in 1,200 tosses of the dice we would expect 100 tosses to give the sum 4.

It should be noted that a probability distribution is analogous to relative frequency distribution with probabilities replacing relative frequencies. Thus we can think of probability distributions as theoretical or ideal limiting forms of relative frequency distribution when the number of observations is made very large. For this reason, we can think of probability distributions as being distributions for populations, whereas relative frequency distributions for populations, whereas relative frequency distributions are distributions drawn from this population.

Example 1: A dealer in refrigerator extimates from his past experience the probabilities of his selling refregerators in a day. These are as follows :

No. of refrigerators sold in a day :	0	1	2	3	4	5	6
Probability :	0.03	0.20	0.23	0.25	0.12	0.10	0.07

Find the mean number of refrigerators sold in a day.

Solution : Mean number of refrigerators sold

= 0 x 0.03 + 1 x 0.2 + 2 x 0.23 + 3 x 0.25 + 4 x 0.12 + 5 x 0.10 + 6 x 0.07

= 0 + .2 + .46 + .75 + .48 + .5 + .42 = 2.81

Hence mean number of refrigerators sold in a day is 3.

Probability distribution of a discrete random variable

Let us consider a discrete r.v,. X which can take the possible values x_1, x_2, x_3....x_n. With each value of the variable X, we associate a number,

$$p_i = P(x = x_i); i = 1, 2,...,n$$

which is known as the probability x_1 and satisfied the following conditions:

i) pi = $P(x= x_i) \geq 0$, (i = 1, 2,...,n)

i.e. p_i's are all non-negative and

ii) $\sum p_i = p_1 + p_2 + ... + p_n = 1$

i.e. the total probability is one

More specifically, let X be a discrete random variable and define :

$$P(x) = P(X = x)$$

such that $p(x) \geq 0$ and $\sum p(x) = 1$, summation being taken over various values of the variable

The function pi = $P(X = X_i)$ or p(x) is called the probability function or more precisely probability mass function (p.m.f.) of the random variable X and the set of all possible ordered pairs {x, p(x)}, is called probability distribution of the random variable X.

Remark : The concept of probability distribution is anlogous to that of frequency distribution. Just as frequency distribution tells us how the total frequency is distributed among different values (or classes) of the variable, similarly a probability distribution tells us how total probability of 1 is distributed among the various values which the random variable can take. It is usually represented in a tabular form gven below :

Table : Probability distribution of r.v.X.

x	x_1	x_2	x_3	.....	x_n
p(x)	p_1	p_2	p_3	.....	p_n

Probability distribution of a continuous random variable

Unlike a discrete probability distribution, a continuous probability distribution can not be presented in a tabular form. It has either a formula form or a graphical form.

A frequency polygon gets smoother and smoother as the sample size gets larger, and the class intervals become more numerous and narrower. Ultimately the destiny polygon becomes a smooth curve called density curve. The function that defines the curve is called the probability density function.

Probability Density Function (p.d.f.) of Continuous Random Variable:

Let X be a continuous random variable taking values on the interval (a, b).

A function p(x) is said to be probability density function of the continuous random variable X if it satisfies the following properties:

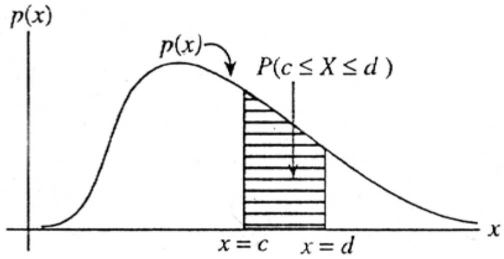
i) $p(x) \geq 0$ for all x in the interval (a, b).

ii) For two dintinct numbers c and d in the interval (a, b)

$P(c \leq x \leq d)$ = [Area under the probability curve between the ordinates (vertical lines) at x = c and x = d] (Fig.)

iii) Total area under the probability cruve is 1, i.e.

$P(a \leq x \leq b) = 1$



Remarks 1: The areas under any probability curve with probability density function p(x) between x = c and x = d can be obtained very conveniently by using the technique of integration from integral calculus viz.,

$$P(c \leq x \leq d) = \int_c^d p(x)dx$$

or by the use of some numerical methods for any complicated form of the function p(x). However, since this is beyond the scope of this book, we will not discuss the problem on continuous random variables.

2. We have

$P(c \leq x \leq d)$ = Area under the probability curve between the vertical lines at x=c and x = d.

In particular, taking c = d : we get :

p(x=c) = Area under the probability curve and the vertical line x = c

$= 0$

because the area of the rectangular strip with width 0 is 0.

This leads to the following very important concept.

For a continuous random variable, the probability at a point is always zero

i.e. $p(x = c) = 0$, for all single point values of c.

Hence, in case of continuous random variable, we always talk of probabilities in an interval and not at a point (which is always zero)

3. Since in the case of continuous random variable, the probability at a point is always zero, we have

$$p(x = c) = 0 \text{ and } p(x = d) = 0 \qquad \text{............}(*)$$

Writing

$$c \le x \le d = c(c < x \le d) \cup (x = c)$$

$$c \le x \le d = c(c \le x \le d) \cup (x = d)$$

$$c \le x \le d = c(c < x < d) \cup (x = c) \cup (x = d)$$

all the events on the right hand being mutually exclusive (disjoint), and using (*), and the addition theorem of probability [Axiom of additivity], we get

$$P\left(c \le x \le d\right) = P(c \le x \le d) = P\left(c \le x \le d\right) = P\left(c \le x \le d\right)$$

Hence, in case of continuous random variable, it does not matter if one or both the end points of the interval (c, d) are included or not.

However, this result is not true, in general, for discrete random variables.

Illustration: If x is a continuous random variable, then

$$P(x \le 5) = P(x < 5)$$

However, if X is a discrete random variable taking positive integer values, then

$$P(x \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$$

and $\quad P(x < 5) = P(x \leq 4)$

$$= P(0) + P(1) + P(2) + P(3) + P(4)$$

Therefore $\quad P(x \leq 5) \neq P(x < 5)$, in general

4. We come back to the point in Remark 5

   Let us consider the distribution of the heights of 100 soldiers, which has, mean = 69" and s.d. = 2.8". If the continuous random variable x denotes the heights (in inches), of the soldiers, then by Remark 2,

   $$P(x=67.4) = 0$$

   a result which does not make much sense

   Theoretically, it means that there is no soldier in a group of 100 soldiers, whose height is 67.4", a result which sounds absurd.

We should, therefore, be careful in interpretng such results and should try to look into their practical implications. Since no measuring instrument can take measurements with exact magnitudes, from practical point of view x = 67.4, simply means that the height of a soldier is 67.4" measured to the nearest first decimal. Practically, 67.4" can be thought of as the value lying in a small interval surrounding it, say, (67.395", 67.405"), which has considerable non-zero area under the given probability curve. Hence, we may compute the required probability as :

$$P(x = 67.4) = P(67.395 \leq x \leq 67.405)$$

5. **Probability Density Function (Continuous r.v.)** :

In case of a continuous random variable, we do not talk of probability at a particular point (which is always zero) but we always talk of probability in an interval. If p(x) dx is the probability that the random variable x takes the value in a small interval of magnitude dx, e.g., $(x, x + dx)$ or $\left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)$, then p(x) is called the probability density function (p.d.f.) of the r.v.x

Distribution Function or Cumulative Probability Function:

If X is a discrete r.v. with probability function p(x) then, the distribution function, usually denoted by F(x) is defined as :

$$F(x) = P(X \le x)$$

It X takes integral values, viz., 1, 2, 3... then

$$F(x) = P(X = 1) + P(X = 2) +.....+ P(X = x)$$

$$\Rightarrow \qquad F(x) = p(1) + p(2) +.....+ P(x)$$

Remarks 1: In the above cases,

$$F(x\text{-}1) = p(1) + p(2) +...+ P(x\text{-}1)$$

$$\therefore \qquad (Fx) - F(x\text{-}1) = p(x) \Rightarrow p(x) = F(x) - F(x\text{-}1)$$

Hence, if X is a random variable which can take positive integral values, then probability function can be obtained from distribution.

2. If X is a continuous r.v. with probability density function p(x), then the distribution function is given by the integral

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(x)dx$$

Moments:

If X is a discrete r.v. with probability function p(x) then :

$$\mu'_r = \text{rth moment about any arbitrary point 'A'} = \Sigma(x - A)^r.p(x) ..(1)$$

$$\mu_r = \text{rth moment about mean } \left(\bar{x}\right) = \Sigma(x - \bar{x})^r.p(x)$$

In particular,

$$\text{Mean } \left(\bar{x}\right) = \text{First moment about origin} = \Sigma x\, p(x)$$

[Taking A and 0 and r = 1]

$$\text{Variance } (x) = \mu_2 = \Sigma(x - \bar{x})^2.p(x) \qquad ....(2)$$

In the expression from (1 to 2), the summation is taken over the various values of the r.v.X.

In the case of continuous r.v. with p.d.f.p(x), the above formulae hold with the only difference that summation is replaced by integration $\left(\int\right)$ over the values of the variable.

Example 1: A die is tossed twice. Getting 'an odd number' is termed as a success. Find the probability distribution of the number of successes.

Solution: Since the cases favourable to getting an odd number in a throw of a die are (1, 3, 5), i.e., 3 in all,

Probability of success(S) $=\frac{3}{6}=\frac{1}{2}$; Probability of failure (F) $=1=\frac{1}{2}=\frac{1}{2}$

If X denotes the number of successes in two throws of a die, then X is a random variable which takes the values 0, 1, 2.

P(X=0) = P[F in Ist throw and F in 2nd throw] =P(FF)=P(F)xP(F)= $=\frac{1}{2}\text{x}\frac{1}{2}=\frac{1}{4}$

P(X=1) = P(S and F) + P(F and S) = P(S) P(F) + P(F) P(S) $=\frac{1}{2}\text{x}\frac{1}{2}+\frac{1}{2}\text{x}\frac{1}{2}=\frac{1}{2}$

P(X=2) = P(S and S) = P(S) $=\frac{1}{2}\text{x}\frac{1}{2}=\frac{1}{4}$

Hence the probability distribution of X is given by :

x	0	1	2
p(x)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 2: Two cards are drawn

a) successively with replacement

b) simultaneously (successively without replacement),

from a well shuffed deck of 52 cards. Find the probability distribution of the number of aces.

Solution : Let X denote the number of aces obtained in a draw of two cards.

Obviously, X is a random variable which can take the values 0, 1 or 2.

a) Probability of drawing an ace $= \frac{4}{52} = \frac{1}{13}$

$\Rightarrow$ Probability of drawing a non-ace $= 1 - \frac{1}{13} = \frac{12}{13}$

Since the cards are drawn with replacement, all the draws are independent.

$$P(X = 2) = P(\text{Ace and Ace}) = P(\text{Ace}) \text{ x } P(\text{Ace}) = \frac{1}{13} \text{ x } \frac{1}{13} = \frac{1}{169}$$

$$P(X=1) = P(\text{Ace and Non-Ace}) = P(\text{Non-ace and Ace})$$

$$= P(\text{Ace}) \text{ x } P(\text{Non-ace}) + P(\text{Non-ace}) \text{ x } P(\text{Ace})$$

$$= \frac{1}{13} \text{ x } \frac{12}{13} + \frac{12}{13} \text{ x } \frac{1}{13} = \frac{24}{169}$$

$$P(X=0) = P(\text{Non-Ace and Non-ace}) = P(\text{Non-ace}) \text{ x } P(\text{Non-ace})$$

$$= \frac{12}{13} x \frac{12}{13} = \frac{144}{169}$$

Hence, the probability of X is :

x :	0	1	2
p(x) :	$\frac{144}{169}$	$\frac{24}{169}$	$\frac{1}{169}$

b) If cards are drawn without replacement, the exhaustive number of cases of drawing 2 cards out of 52 cards is $^{52}C_2$.

$$\therefore\ P(X=0) = P(\text{No ace}) = P(\text{Both cards are non-aces}) = \frac{^{48}C_2}{^{52}C_2} = \frac{48 \text{ x } 47}{52 \text{ x } 51} = \frac{188}{221}$$

$$P(X=1) = P(\text{One Ace}) = P(\text{One ace and one non-ace})$$

$$= \frac{^{4}C_1 \text{x}\, ^{48}C_1}{^{52}C_2} = \frac{4\text{x}48 \text{ x } 2}{52 \text{ x } 51} = \frac{32}{221}$$

$$P(X=2) = P(\text{both aces}) = \frac{{}^{4}C_{2}}{{}^{52}C_{2}} = \frac{4 \text{ x } 3}{52 \text{ x } 51} = \frac{1}{221}$$

Hence, the probability distribution of X becomes :

x :	0	1	2
p(x) :	$\frac{188}{221}$	$\frac{32}{221}$	$\frac{1}{221}$

***Example 3*:** Obtain the probability distribution of X, the number of heads in tosses of a coin (or a simultaneous toss of three coins).

Solution: Obviously, X is a random variable which can take the values 0,1, 2 or 3. The sample space S consists of $2^3 = 8$ sample points, as given below:

$$S = \{(H, T) \text{ x } (H, T) \text{ x } (H, T)\}$$
$$= \{(HH, HT, TH, TT) \text{ x } (H, T)\}$$
$$= \{HHH, HTH, THH, TTH, HHT, HTT, THT, TTT\}$$

The probability distribution of X is given in Table

Table : Probability distribution of number of heads in 3 tosses of a coin

No. of heads (x)	Favourable events	No. of favourable	Probability
	cases	p(x)	
0	(TT)	1	$\frac{1}{8}$
1	(TTH, HTT, THT)	3	$\frac{3}{8}$
2	(HTH, THH, HHT)	3	$\frac{3}{8}$
3	(HHH)	1	$\frac{1}{8}$

***Example 4*:** Two dice are rolled at random. Obtain the probability distribution of the sum of the numbers on them.

Solution : When two dices are rolled, the sample space S consists of 6^2=36, sample points as shown.

Let X denote the sum of the numbers on the two dice. Then X is a random variable which can take the values 2, 3, 4,...,12 with the probability distribution given in Table.

$$S = \begin{Bmatrix} (1,1), & (1,2), & \ldots.., & (1,6) \\ (2,1), & (2,2), & \ldots.., & (2,6) \\ : & & & : \\ (6,1), & (6,2), & \ldots.., & (6,6) \end{Bmatrix}$$

Table : Probability Distribution of Sum of points in toss of two dice

Sum of numbers (x)	Favourable sample Points	No. of favourable cases	Probability P(x)
2	(1, 1)	1	$\frac{1}{36}$
3	(1, 2), (2, 1)	2	$\frac{2}{36}$
4	(1, 3), (3, 1), (2, 2)	3	$\frac{3}{36}$
5	(1, 4) (4, 1), (2, 3), (3, 2)	4	$\frac{4}{36}$
6	(1, 5) (5, 1), (2, 4), (4, 2), (3, 3)	5	$\frac{5}{36}$
7	(1, 6) (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)	6	$\frac{6}{36}$

8	(2, 6) (6, 2), (3, 5), (5, 3), (4, 4)	5	$\frac{5}{36}$
9	(3, 6) (6, 3), (4, 5), (5, 4)	4	$\frac{4}{36}$
10	(4, 6) (6, 4), (5, 5)	3	$\frac{3}{36}$
11	(5, 6), (6, 5)	2	$\frac{2}{36}$
12	(6, 6)	1	$\frac{1}{36}$

12.3 MATHEMATICAL EXPECTATION AND RANDOM VARIABLE:

The concept of mathematical expectation is of great importance in statistical work. The mathematical expectation (also called the expected value) of a random variable is the weighted arithmetic mean of the variable, the weights used to find the mathematical expectation are all the respective probabilities of the values that the variable can possibly assume.

If X denotes a discrete random variable which can assume the values x_1, x_2, x_3....x_k, with respective probabilities p_1, p_2, p_3....,p_k where $p_1 + p_2 + p_3 +...+p_k=1$ the mathematical expectation of X denoted E(x) is defined as :

$$E(x) = p_1x_1 + p_2x_2 + p_3x_3 +...+p_kx_k$$

Thus the expected value equals the sum of each particular value within the set (x) multiplied by the probability that x equals that particular value.

Example 1: A petrol pump properietor sells on an average Rs. 80,000 worth of petrol on rainy days and an average of Rs. 95,000 on clear days.

Statistics from the meteorological department show that the probability is 0.76 for clear weather and 0.24 for rainy weather on coming Monday. Find the expected value of petrol sale on coming Monday.

Solution : x_1= 80,000 p_1 = 0.24

x_2= 95,000 p_2 = 0.76

$$\Sigma(x) = p_1x_1 + p_2x_2 .....$$

$$= 0.24\ (80,000) + 0.76\ (95,000)$$

$$= 19,200 + 72,200 = \text{Rs. } 91,400$$

Thus the expected value of petrol sale on Monday is Rs. 91,400.

12.4 MOMENTS :

Given 'n' observation x_1, x_2,....,$x_{n,}$ and an arbitrary constant A,

$\frac{1}{n}\Sigma(x - A)$ is called the Ist moment about A,

$\frac{1}{n}\Sigma(x - A)^2$ is called the 2nd moment about A,.....(1)

$\frac{1}{n}\Sigma(x - A)^3$ is called the 3rd moment about A,

and so on. Let us denate these moments successively by m_1', m_2', m_3', etc.(sometimes, these are also represented by μ_1', μ_2', μ_3', etc)

Then $m_1' = \Sigma(x - A)/n = (\Sigma x - \Sigma A)/n = (\Sigma x - nA)/n$

$= \bar{x} - A$ ....(2)

i.e., the Ist moment about A equals $(\bar{x} - A)$. Moments abot zero (i.,e when A = 0), and moments about mean (i.e, when A = x) are particularly important.

12.4.1 Moments about zero (or raw moments):

Ist moment about zero $= \frac{1}{n}\Sigma x = \bar{x}$

2nd moment about zero $= \frac{1}{n}\Sigma x^2$ ....(3)

3rd moment about zero $= \frac{1}{n}\Sigma x^3$

and so on. Note that the Ist moment about zero is the mean $\overline{x}$

$$m_1^{'} = \overline{x} \qquad .... (4)$$

12.4.2 Moments About Mean (or Cental Moments) :

Ist moment about mean $= \frac{1}{n}\Sigma\left(x - \overline{x}\right) = 0$

2nd moment about mean $= \frac{1}{n}\Sigma\left(x - \overline{x}\right) = \sigma^2$

3rd moment about mean $= \frac{1}{n}\Sigma\left(x - \overline{x}\right)^3$

4th moment about mean $= \frac{1}{n}\Sigma\left(x - \overline{x}\right)^4$ ....(5)

and so on. These are usually denoted by m_1, m_2, m_3, m_4, etc. (Sometimes, these are represented by $\mu_1, \mu_2, \mu_3, \mu_4$, etc.) Note that the Ist central moment is always zero, and the 2nd central moment is the variance σ^2. The 3rd central moment is used to measure skewness and the 4th central moment m_4 to measure Kurtosis. Higher order moments m_5, m_6, etc. are seldom used.

In general, given 'n' observations x_1, x_2,....x_n, the rth order, moments (r = 0, 1, 2, ...) are defined as follows :

r - th moment about A: $m_r^{'} = \frac{1}{n}\Sigma\left(x - A\right)^r$

r - th raw moment : $m_r^{'} = \frac{1}{n}\Sigma x^r$

r - th central moment : $m_r = \frac{1}{n}\Sigma\left(x - \overline{x}\right)^r$ ......(6)

For a frequency distribution,

$$\text{r - th moment about A} = m_r^{'} = \frac{1}{n}\Sigma f(x - A)^r$$

$$\text{r - th raw moment : } m_r^{'} = \frac{1}{N}\Sigma fx^r$$

$$\text{r - th central moment : } m_r = \frac{1}{N}\Sigma f(x - \bar{x})^r \qquad \text{......... (7)}$$

where $N = \Sigma f$ (Note that moments about mean are written without dashes ('), but moments about any other origin, i.e, non-central moments, with dashes)

There are important relations between central and non-central moments. For example, if the non-central moments ($m_1^{'}, m_2^{'}, m_3^{'}$ etc.) about any arbitrary origin A are known, the central moments can be obtained by using the relations (equation 2) viz.,

$$m_2 = m_2^{'} - m_1^{'2}$$

$$m_3 = m_3^{'} - 3m_2^{'}m_1^{'} + 2m_1^{'3} \qquad \text{...........(8)}$$

$$m_4 = m_4^{'} - 4m_3^{'}m_1^{'} + 6m_2^{'}m_1^{'2} - 3m_1^{'4}$$

In particular, using the first two moments $m_1^{'}$ and $m_2^{'}$ about an arbitrary origin A, the mean and variance may be obtained :

$$\bar{x} = m_1^{'} + A, \sigma^2 = m_2^{'} - m_1^{'2} \qquad \text{...........(9)}$$

Example 1: The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and the variance.

Solution : Using formulae (equation 9), we have here

$m_1^{'} = 2, m_2^{'} = 20$, and A = 5

$$\bar{x} = 2 + 5 = 7$$

$$\sigma^2 = 20 - 2^2 = 16$$

Example 2 **:** The first four moments of a distribution about the value 3 are 2, 10, 40 and 218. Find the moments about the origin and mean.

Solution: Given

$$\sum(x-3)/n = 2, \qquad \sum(x-3)^2/_n = 10,$$

$$\sum(x-3)^3/_n = 40, \qquad \sum(x-3)^4/_n = 218,$$

We have to find (a) $\sum x/n$, $\sum x^2/n$, $\sum x^3/n$, $\sum x^4/n$; and (b) m_1, m_2, m_3, m_4.

a) $\sum(x-3)/n = 2$;

or $(\sum x - 3n)/n = 2$

or, $\sum x/n - 3 = 2$

$\therefore$ $\sum x/n = 2+3=5$ ....(i)

$\sum(x-3)^2/_n = 10$

or $\sum(x^2-6x+9)/n = 10$

or, $\sum x^2/n - 6\sum x/n + 9 = 10$; putting $\sum x/n = 5$ from (i)

$\sum x^2/_n = 10+30-9 = 31$ ....(ii)

Again $\sum(x-3)^3/n = 40$

Using the binomial $(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$ we have

$\sum(x^3 - 3.x^2.3 + 3.x3^2 - 3^3)/n = 40$

or $\sum x^3/_n - 9\sum x^2/_n + 27\sum x/n - 27 = 40$

or $\sum x^3/_n - 9 \text{ x } 31 + 27 \text{ x } 5 - 27 = 40$; from (i) and (ii)

$\therefore$ $\sum x^3/_n = 40 + 279 - 135 + 27 = 211$ ....(iii)

Also, $\sum(x-3)^4/_n = 218$

Using the binomial expansion $(a-b)^4 = a^4 - 4a^3b + 6a^2b^2 - 4ab^3 + b^1$

$$\Sigma(x^4 - 4.x^3.3 + 6.x^2.3^2 - 4.x.3^3 + 3^4)/n = 218$$

or, $\Sigma x\,{}^1\!/_n - 12\Sigma x\,{}^3\!/_n + 54\Sigma x\,{}^2\!/_n - 108\Sigma x/n + 81 = 218$

or, $\Sigma x\,{}^1\!/_n - 12 \text{ x } 211 + 54 \text{ x } 31 - 108 \text{ x } 5 + 81 = 218$; from (i), (ii) and (iii)

$\therefore$ $\Sigma x\,{}^4\!/_n = 218 + 2532 - 1674 + 540 - 81 = 1535$ ....(iv)

b) The moments about 3 are given as

$$m_1^{'} = 2, m_2^{'} = 10, m_3^{'} = 40, m_4^{'} = 218$$

Hence, using relations (equation 8) the moments about mean are

$m_1 = 0$ (in all cases)

$$m_2 = m_2^{'} - m_1^{'2} = 10 - 2^2 = 6$$

$$m_3 = m_3^{'} - 3m_2^{'}m_1^{'} + 2m_1^{'3}$$

$$= 40 - 3 \text{ x } 10 \text{ x } 2 + 2 \text{ x } 2^3 = -4$$

$$m_4 = m_4^{'} - 4m_3^{'}m_1^{'} + 6m_2^{'}m_1^{'2} - 3m_1^{'4}$$

$$= 218 - 4 \text{ x } 40 \text{ x } 2 + 6 \text{ x } 10 \text{ x } 2^2 - 3 \text{ x } 2^4$$

$$= 218 - 320 + 240 - 48 = 90$$

Note : The moments about any arbitrary origin may be used to obtain the central moments. Hence, we may also use the moments about origin, obtained earlier, viz., 5, 31, 211, 1535, to find the central moments.

$$m_2 = 31 - 5^2 = 6$$

$$m_3 = 211 - 3 \text{ x } 31 \text{ x } 5 + 2 \text{ x } 5^3 = 211 - 465 + 250 = -4$$

$$m_4 = 1535 - 4 \text{ x } 211 \text{ x } 5 + 6 \text{ x } 31 \text{ x } 5^2 - 3 \text{ x } 5^1$$

$$= 1535 - 4220 + 4650 - 1875 = 90$$

12.5 LESSON END EXERCISE :

Q1. The first two moments of a distribution about the value of 5 of the variable are

2 and 20. Find the mean and the variance ?

Ans. 7, 16

Q2. If are m_2' and m_2 respectively the second moment about an arbitrary origin a and that about $\overline{x}$ then show that $m_2' = m_2 + d^2$, where $d = \overline{x} - a$

Q3. Explain mathematical expectation with the help of example ?

Q4. Analyse the frequency distribution by the method of moments :

X :	2	3	4	5	6
F :	1	3	7	3	1

Q5. Explain probability distribution with random variable ?

****

M.A. Economics **Semester - IInd** **Lesson - 13**
Course No. 203 **Unit - III**

PROBABILITY DISTRIBUTIONS : BINOMIAL AND POISSON

STRUCTURE :

13.1 INTRODUCTION:

The statistical measures like the averages, dispersion, skewness, kurtosis, correlation, etc., for the sample frequency distributions, not only give us the nature and form of the sample data but also help us in formulating certain ideas about the characteristics of the population. However, a more scientific way of drawing inferences about the population characteristics is through the study of theoretical distributions.

Meaning of Probability Distributions : In the population, the values of the variable may be distributed according to some definite probability law which can be expressed mathematically and the corresponding probability distribution is known as theoretical probability distribution. Such probability laws may be based on a 'prior' consideration or 'a posterior' inferences. These distributions are based on expectations on the basis of previous experience. Theoretical distributions also enable us to fit a mathematical model or a function of the form y = p(x) to the given data.

In the previuos lessons, we have already discussed the random variable mathematical expectation, probability function and distribution function, moments, mean and variance in terms of probability function. These provide us the necessary tools for the study of theoretical distributions.

Now, we will study univariate probability distributions :

i) Binomial distribution

ii) Poisson distribution

iii) Normal distribution

The first two distributions are discrete probability distributions and the third is a continuous probability distribution.

BINOMIAL DISTRIBUTION :

Binomial distribution is also known as the 'Bernoulli distribution' after the Swiss mathematician James Bernoulli (1654-1705) who discovered it in 1700 and was first published in 1713, eight years after his death.

13.2 BINOMIAL DISTRIBUTION'S CONDITIONS

This distribution can be used the following conditions:

i) The random experiment is performed repeatedly a finite and fixed number of times. In other words n, the number of trials, is finite and fixed.

ii) The outcomes of the random experiment (trial) results in the dichatomous classification of events. In other words, the outcome of each trial may be classified into two mutually disjoint categories called success (the occurrence of the event) and failure (the non-occurence of the event).

iii) All the trials are independent, i.e, the result of any trial, is not affected in any way, by the preceding trials and doesn't affect the result of succeeding trials.

iv) The probability of success (happening of an event) in any trial is 'p' and is constant for each trial. q = 1 - p, is then termed as the probability of failure (non-occurence of the event) and is constant for each trial.

For example, if we toss a fair coin n times (which is fixed and finite), then the outcomes of any trial is one of the mutually exclusive, viz., head (success) and tail (failure). Further, all the trials are independent, since the result of any throw of a coin does not affect and is not affected by the result of other throws. Moreover, the probability of success(head) in any trial is $\frac{1}{2}$, which is constant for each trial. Hence the coin tossing problems will give rise to Bionomial distribution.

Similarly dice throwing problems will also conform to Binomial distribution.

More precisely, we expect a Binomial distribution under the following conditions :

i) n, the number of trial is finite.

ii) Each trial results in two mutually exclusive and exhaustive outcomes, termed as success and failure.

iii) Trials are independent.

iv) p, the probability of success is constant for each trial. Then q = 1 - p, is the probability of failure in any trial.

The trials satisfying the above four conditions are known as Bernoulli trials.

1. **Probability Function of Binomial Distribution :** If X denotes the number of successes in 'n' trials satisfying the above conditions, then X is a random variable which can take the values 0, 1, 2,...., n; since in n trials we may get no success (all failures), one success, two successes,...., or all the 'n' successes.

We are interested in finding the corresponding probabilities of 0, 1, 2,....,n successes. The general expression for the probability of 'r' successes is given by :

$$p(r) = P(x = r) = {}^nC_r\, p^r.q^{n-r};\ r = 0, 1, 2, \ldots.., n \ldots\ldots\ldots\ldots(1)$$

Proof : Let S_i denote the success and F_i denote the failure at the ith trial; i = 1, 2, ...,n. Then, we

$$P(S_i) = p \text{ and } P(F_i) = q\ ;\ i = 1, 2,\ldots.., n \qquad \ldots\ldots\ldots\ldots(2)$$

The probability of r success and consequently (n - r) failures in a sequence of n-trials in any fixed specified order, say, $S_1\ F_2\ S_3\ S_4\ F_5\ F_6 \ldots\ldots.S_{n-1}F_n$ where S occurs r times and F occurs (n-r) times is given by :

$$P[S_1 \cap F_2 \cap S_3 \cap S_4 \cap F_5 \cap F_6 \cap \ldots.. \cap S_{n-1} \cap F_n]$$

$$= P(S_1).P(F_2).P(S_3).P(S_4).P(F_5).P(F_6)\ldots\ldots..P(S_{n-1}).P(F_n)$$

[By compound probability theorem, since the trials are indepdenent]

$$= p.q.p.p.q.q\ldots\ldots..p.q \qquad \text{(from equation 2)}$$

$$= [p \text{ x } p \text{ x } p \text{ x } \ldots\ldots\ldots..r \text{ times}] \text{ x } [q \text{ x } q \text{ x } q\ldots.(n\text{-}r) \text{ times}]$$

$$= p^r.q^{n-r} \qquad \ldots\ldots\ldots.(3)$$

But in 'n' trials, the total number of possible ways of obtaining r successes and (n-r) failures is

$$\frac{n!}{r!(n-r)!} = {}^nC_r$$

all of which are mutually disjoint. The probability for each of these ${}^nC_r$ mutually exclusive ways is same as given in (equation 3) viz., p^rq^{n-r}. Hence by the addition theorem of probability, the required probability of getting 'r' successes and consequently (n-r) failures in 'n' trials, in any order what-so-ever is given by :

$$P(X = r) = p^r q^{n-r} + p^r q^{n-r} + \ldots. + p^r q^{n-r} ({}^nC_r \text{ terms})$$

$$= {}^nC_r p^r q^{n-r};\ r = 0, 1, 2\ldots., n$$

Remarks 1 :

Putting r = 0, 1, 2,...., in (table 3.1) we get the probabilities of 0, 1, 2,...,n successes respectively in 'n' trials and these are tabulated in table 3.1 since the probabilities are the successive terms in the binomial expansion $(q + p)^n$, it is called binomial distribution.

Table 3.1 : Binomial Probabilities

r	$p(r) = P(X = r)$
0	${}^nC_0p^oq^n = q^n$
1	${}^nC_1p^1q^{n-1}$
2	${}^nC_2p^2q^{n-2}$
:	:
:	:
:	:
n	${}^nC_np^nq^0 = p^n$

2. **Total probability is unity, i.e., 1 :**

$$\sum_{r=0}^{n} p(r) = p(0) + p(1) + \ldots.. + p(n)$$

$$= q^n + {}^nC_1q^{n-1}p + {}^nC_2q^{n-2}p^2 + \ldots.. + p^n$$

$$= (q + p)^1 = 1$$

$$(\therefore\ p + q = 1)$$

3. The expression for P(x=r) in table 3.1 is known as the probability (mass) function of the Binomial distribution with parameters 'n' and 'p'. The random variable X following the probability law (table 3.1) is called a Binomial variate with parameters 'n' and 'p'.

   The Binomial distribution is completely determined, i.e., all the probabilities can be obtained, if 'n' and 'p' are known. Obviously, 'q' is known when 'p' is given because q = 1 - p.

4. Since the random variable x takes only integral values, Binomial distribution is a discrete probability distribution.

5. For ‘n’ trials, the binomial distribution consists of (n+1) terms, the successive binomial coefficients being,

$$^{n}C_{0},^{n}C_{1},^{n}C_{2},......,^{n}C_{n-1},^{n}C_{n}$$

since $^{n}C_{0}=^{n}C_{n-1}=1$, the first and last coefficient will always be 1. Further since

$$^{n}C_{r}=^{n}C_{n-r}$$

the binomial coefficients will be symmetric. Moreover, we have for all values of x:

$$(1+x)^{n}=^{n}C_{0}x,^{n}C_{1}+^{n}C_{2}x^{2},......,^{n}C_{n}x^{n}$$

Putting X = 1 we get :

$$(1+1)^{n}=^{n}C_{0}+^{n}C_{1}+^{n}C_{2}+,......+^{n}C_{n}\Rightarrow^{n}C_{0}+^{n}C_{1}+^{n}C_{2}+,......+^{n}C_{n}$$

i.e., the sum of binomial coefficients is 2^{n}.

3.2 : Constants of Binomial distribution

r	P(x = r) = p(r)	r.p(r)	r2.p(r)
0	qn	0	0
1	$^{n}C_{1}q^{n-1}p$	1. $^{n}C_{1}q^{n-1}p$	1^{2}. $^{n}C_{1}q^{n-1}p$
2	$^{n}C_{2}q^{n-2}p^{2}$	$^{2.}$ $^{n}C_{2}q^{n-2}p^{2}$	2^{2}. $^{n}C_{2}q^{n-2}p^{2}$
3	$^{n}C_{3}q^{n-3}p^{3}$	3. $^{n}C_{3}q^{n-3}p^{3}$	$3^{3.}$ $^{n}C_{3}q^{n-3}p^{3}$
.	.	.	.
.	.	.	.
.	.	.	.
n	p^{n}	np^{n}	$n^{2}p^{n}$

Mean $=\sum rp(r)=^{n}C_{1}q^{n-1}+2^{n}C_{2}q^{n-2}.p^{2}+3^{n}C_{3}q^{n-3}p^{3},......+np^{n}$

$$=nq^{n-1}p+2.\frac{n(n-1)}{2!}q^{n-2}p^{2}+\frac{3n(n-1)(n-2)}{3!}q^{n-3}p^{3}+...+np^{n}$$

$$= np\left[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2!}q^{n-3}p^2 + \ldots + p^{n-1}\right]$$

$$= np\left[q^{n-1} + {}^{n-1}C_1 q^{n-2}p + {}^{n-2}C_2 q^{n-3}p^2 + \ldots + p^{n-1}\right]$$

$= np\ (q + p)^{n-1}$ [By binomial expansion for positive integer index]

$= np$

$(\therefore p + q = 1)$

Variance = $\sum rp(r) - [\sum rp(r)]^2 = \sum r^2 p(r) - (\text{mean})^2$

$$\sum r^2 p(r) = 1^2 x\, {}^nC_1 q^{n-1}p + 2^2\, {}^nC_2 q^{n-2}p^2 + 3^2\, {}^nC_3 q^{n-3}p^3 + \ldots. + n^2p^n$$

$$= nq^{n-1}p + \frac{4n(n-1)}{2!}q^{n-2}p^2 + \frac{9n(n-1)(n-2)}{3!}q^{n-3}p^3 + \ldots. + n^2p^n$$

$$= np\left[\left\{q^{n-1} + 2(n-1)q^{n-2}p + \frac{3}{2}(n-1)(n-2)q^{n-3}p^2 + \ldots + np^{n-1}\right\}\right.$$

$$= np\left[\left\{q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2}q^{n-3}p^2 + \ldots + 1p^{n-1}\right\}\right.$$

$$\left. + \left\{(n-1)q^{n-2}p + (n-1)(n-2)q^{n-3}p^2 + \ldots + (n-1)p^{n-1}\right\}\right]$$

$$= np\left[\left\{q^{n-1} + (n-1)q^{n-2} + p + \frac{3}{2}(n-1)(n-2)q^{n-3}p^2 + \ldots + np^{n-1}\right\}\right.$$

$$\left. + \left\{(n-1)q^{n-2} + p + (n-1)(n-2)q^{n-3}p^2 + \ldots + (n-1)p^{n-1}\right\}\right]$$

$$= np\left[\left\{(q+p)^{n-1}\right\} + (n-1)p\left\{q^{n-2} + (n-2)q^{n-3}p + \ldots + p^{n-2}\right\}\right]$$

$$= np\left[(q+p)^{n-1} + (n-1)\,p(q+p)^{n-2}\right]$$

$= np\ [1 + (n-1)p]$

$(\therefore p = q = 1)$

substituting in (*) we get

Variance = np[1+np-p] -$(np)^2$ =np[1+np-p-np] = np[1-p]=npq

Hence for the binomial distribution,

$$\text{Mean} = \text{np} \quad \text{..........(4) and} \quad \mu^2 = \sigma^2 = \text{npq} \quad \text{..........(5)}$$

Similarly we can obtain the other constants given below :

$$\mu_3 = \text{npq(q - p)} \quad \text{........(6) and} \quad \mu_4 = \text{npq}[1 + 3\text{pq(n - 2)}] \quad \text{.........(7)}$$

Hence, the moment coefficient of skewness is :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[\text{npq}(q-p)]^2}{(\text{npq})^3} = \frac{(q-p)^2}{\text{npq}} \quad \text{............(8)}$$

and

$$\gamma_1 = +\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{q-p}{\sqrt{\text{npq}}} \quad \text{...........(8a)}$$

Coefficient of Kurtosis is given by :

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\text{npq}[1 + 3\text{pq}(n - 2)]}{(\text{npq})^2}$$

$$= \frac{1 + 3\text{pq}(n - 2)}{\text{npq}}$$

$$\therefore \quad \beta_1 = \frac{3 + 1 - 6\text{pq}}{\text{npq}} \quad \text{........(9)}$$

and

$$\gamma_2 = \beta_2 - 3 = \frac{1 - 6\text{pq}}{\text{npq}} \quad \text{.......(9a)}$$

Remarks :

1. Since q is the probability (of failure), we always have 0 < q <1.

   ∴ Variance = np x q < np = Mean

($\therefore$ 0 < q < 1)

$\Rightarrow$ Varaicne < mean ....(9b)

Hence for the Binomial distribution variance is less than mean.

2. Var (x) = npq = n (1-p) = n(p - p^2) = ϕ (p), say ....(*)

for maximum value of ϕ (p), we should have :

ϕ '(p)=0 and ϕ "(p) = 0 ....(**)

Differentiating (***) w.r.t. 'p', we get :

ϕ "(p) = n(-2) = -2n < 0.

Hence, using (**), we conclude that var (x) = ϕ (p), is maximum when

$$p = \frac{1}{2} \Rightarrow q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

$\therefore$ Maximum Variance = (npq) $= n \text{ x } \frac{1}{2} \text{x} \frac{1}{2} = \frac{n}{4}$

Hence, if X ~ B(n, p), var $\text{Var}(x) \leq \frac{n}{4}$ ....(9c)

i.e., for the binomial distribution with parameter n and p, variance cannot exceed n/4.

3. As $n \to \infty$, from equation (8) and (9), we get

$$\beta_1 \to 0, \gamma_1 \to 0, \beta_2 \to 3 \text{ and } \gamma_2 \to 0$$

4. From equation (8) we see that Binomial division is symmetrical if

$$\beta_1 = 0 \Rightarrow q = p = \frac{1}{2}$$

$(\therefore q + p = 1)$

from equation (8) we observe that, it is postively skewed if

$$\gamma_1 > 0 \Rightarrow q - p > 0 \Rightarrow 1 - 2p > 0$$

$$\Rightarrow 1 > 2p \Rightarrow \frac{1}{2} > p \Rightarrow p < \frac{1}{2}$$

Similarly, it is negatively skewed if

$$\gamma_1 < 0 \Rightarrow q - p < 0 \Rightarrow p > \frac{1}{2}$$

Hence, we arrive at all the following conclusion :

"Binomial distribution is symmetrical if p = q = 0.5. It is postively skewed if p<0.5 and negatively skewed if p > 0.5. Obviously, from equation (8) and (8a), we observe that as 'n' increases, the skewness of the binomial distribution becomes less pronounced, irrespective of the values of p and q.

1.2 Mode of Binomial Distribution:

Mode is the value of x which maximises the probability function. Thus if x = r gives mode then we should have

p(r) > p(r-1) and p(r) >p(r+1) ....(10)

13.2.1 Working Rule to find the Mode of Binomial Distribution:

Let x be a Binomial variate with parameters n and p.

Case (1) : When (n+1) p is an integer

Let (n + 1) p = k (an integer).

In this case the distribution is bi-modal, the two modal values being X = k and X = k - 1.

Thus if n = 9 and p = 0.4, then (n+1) p = 10x0.4=4, which is an integer. Hence, in this case the distribution is bi-modal, the two modal values being 4 and 4-1=3.

Case (2) : When (n+1) p is not an integer

Let (n + 1) p = k_1 + f, where k_1 is the integral part and f is the fractional part of (n+1)p. In this case distribution has a unique mode at X = k_1, the integral part of (n+1)p.

For example: if n = 7 and p = 0.6, then (n+1)p = 8x0.6=4.8

Hence mode = 4, the integral part of 4.8

Remarks : Let np be a whole number. Then

$$(n + 1)p = np + p = (\text{whole number}) + \text{fraction}$$

$\Rightarrow$ (n+1) is not an integer

Hence, in this case, the binomial distribution is unimodal, the unique mode being 'np'.

It np is a whole number (i.e., integer), then the distribution is unimodal and the mean and mode are equal, each being 'np'

Example 1: Ten unbiased coins are tossed simultaneously. Find the probability of obtaining.

i. Exactly 6 heads
ii. At least 8 heads
iii. No head
iv. At least one head
v. Not more than three heads
vi. At least 4 heads

Solution : If p denotes the probability of a head, then $p = q = \frac{1}{2}$. Here n = 10. If the random variable X denotes the number of heads, then by the Binomial probability law, the probability of r heads is given by,

$$p(r) = P(x{=}r) = {}^{n}C_{r}p^{r}.q^{n-r}$$

$$= {}^{10}C_{r}\left(\frac{1}{2}\right)^{r}.\left(\frac{1}{2}\right)^{10-r}$$

$$= {}^{10}C_{r}.\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}{}^{10}C_{r}......(*)$$

i) Required probability : $= p(6) = \frac{1}{1024}$

$$^{10}C_6 = \frac{210}{1024} = \frac{105}{512} \qquad \text{............[from (*)]}$$

ii) Required probability $= P(x \geq 8) = p(8) + p(9) + p(10)$

$$= \frac{1}{1024}\left[{}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right] \qquad \text{[from (*)]}$$

$$= \frac{45+10+1}{1024} = \frac{56}{1024} = \frac{7}{128}$$

iii) Required probability $= P(x = 0) = p(o) = \frac{1}{1024}$

$$^{10}C_0 = \frac{1}{1024} \qquad \text{............[from (x)]}$$

iv) Required probability = P [At least one head]

= 1 - P[no head] = 1 - p(o)

$$= 1 - \frac{1}{1024} = \frac{1023}{1024} \qquad \text{[from part (iii)]}$$

v) Required probability $= P(X \leq 3) = p(0) + p(1) + p(2) + p(3)$

$$= \frac{1}{1024}\left[{}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3\right]$$

$$= \frac{1+10+45+120}{1024} = \frac{176}{1024} = \frac{11}{64}$$

vi) Required probability $= (x \geq 4) = p(4) + p(5) + ... + p(10)$

$$= \frac{1}{1024}\left[{}^{10}C_4 + {}^{10}C_5 + ... + {}^{10}C_{10}\right]$$

Last part can be conveniently done as follows

Required probability $= P(x \geq 4) = 1 - P(x \leq 3)$

$$= 1 - [p(0) + p(1) + p(2) + p(3)] = 1 - \frac{11}{64} = \frac{56}{64} \quad \text{[from part (v)]}$$

Example 2: In a binomial distribution with 6 independent trials, the probabiity of 3 and 4 successes is found to be 0.2457 and 0.0819 respectively. Find the parameters p and q of the binomial distribution.

Solution : X~B(n=6, p) where x denotes the number of successes. Then, by binomial probability law, the probability of r successes is given by :

$$p(r) = P(x = r) = {}^6C_r p^r q^{6-r};\ r = 0, 1, 2, \ldots, 6;\ (q = 1 - p) \qquad \ldots\ldots(*)$$

putting r = 3 and 4 in (*), we get respectively:

$$p(3) = {}^6C_3 p^3 q^2 = 20p^3q^3 = 0.2457 \text{ (given)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (**)$$

$$p(4) = {}^6C_4 p^4 q^2 = 15p^4q^2 = 0.0819 \text{ (given)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (***)$$

$$\left[\therefore {}^6C_3 = \frac{6 \text{ x } 5 \text{ x } 4}{3!} = 20; {}^6C_4 = {}^6C_2 \frac{6 \text{ x } 5}{2} = 15\right]$$

Dividing (***) by (**) we get :

$$\frac{p(4)}{p(3)} = \frac{15p^4q^2}{20p^3q^3} = \frac{0.0819}{0.2457} = \frac{1}{3} \Rightarrow \frac{3}{4} \cdot \frac{p}{q} = \frac{1}{3}$$

$$\therefore \qquad 9p = 4q = 4(1 - p) \Rightarrow 13p = 4 \Rightarrow p = \frac{4}{13}$$

$$\Rightarrow q = 1 - p = 1 - \frac{4}{13} = \frac{9}{13}$$

13.2.2 Fitting of Binomial Distribution :

Suppose a random experiment consists of 'n' trials, satisfying the conditions of binomial distribution and suppose this experiment is repeated N-times. Then the frequency of r successes is given by the formula.

$$N \text{ x } p(r) = N \text{ x } {}^nC_r p^r q^{n-r};\ r = 0, 1, 2, \ldots., n \qquad \ldots\ldots\ldots\ldots(11)$$

Putting r = 0, 1, 2,...., n we get the expected or theoretical frequencies of the binomial distribution, which are given in the table 3.3

No. of successes (r)	Expected or theoretical frequencies N.p(r)
0	$N.q^n$
1	$N.{}^nC_1.q^{n-1}p$
2	$N.{}^nC_2.q^{n-2}p^2$
:	:
:	:
n	$N.p^n$

If p, the probability of success which is constant for each trial is known, then the expected frequencies can be oblained easily as given in the table :3.3. However, if p is not known and if we want to graduate or fit a binomial distribution to a given frequency distribution, we first find the mean of the given frequency distribution by the formula $\bar{x} = \sum fx / \sum f$ and equate it to 'np', which is the meanof the binomial probability distribution. Hence, p can be estimated by the relation.

$$np = \bar{x} \qquad \Rightarrow \qquad p = \frac{\bar{x}}{n} \quad ....(12)$$

then q = 1-p. With these values of p and q, the expected or theoretical binomial frequencies can be obtained by using the formula given in the table: 3.3

Example 1: a) 8 coins are tossed at a time, 256 times. Find the expected frequencies of successes (getting a head) and tabulated the results obtained.

b) Also obtain the values of the mean and standard deviation of the theoretical (fitted distribution.

Solution : In the usual notations, we are given :

$n = 8, N = 256$

$p =$ probability of success (head) in a single throw of a coin $= \frac{1}{2}$

$\therefore \quad q = 1 - p = \frac{1}{2}$

Table 3.4 : Expected binomial frequencies

No. of heads	Expected frequencies
0	${}^8C_0 = 1$
1	${}^8C_1 = 8$
2	${}^8C_2 = 28$
3	${}^8C_3 = 56$
4	${}^8C_4 = 70$
5	${}^8C_5 = 56$
6	${}^8C_6 = 28$
7	${}^8C_7 = 8$
8	${}^8C_8 = 1$

Hence, by the binomial probability law, the probability of r successes in a toss of 8 coins is given by :

$$p(r) = {}^nC_r p^r q^{n-r} = {}^8C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{8-r}$$

$$= {}^8C_r \left(\frac{1}{2}\right)^8 = \frac{1}{256} \qquad {}^8C_r \ldots.(*)$$

Hence, in 256 throws of 8 coins, the frequency of r successes is :

$$= f(r) = N - p(r) = 256 \text{ x } \frac{1}{256} {}^8C_r = {}^8C_r; r = 0, 1, \ldots, 8$$

Thus, the expected frequencies are as obtained in table 3.4

b) For the theoretical distribution (Binomial distribution),

$$\text{Mean} = np = 8x\frac{1}{2} = 4;\quad \text{s.d.} = \sqrt{npq} = \sqrt{8x\frac{1}{2}x\frac{1}{2}} = \sqrt{2}$$

$\Rightarrow 1.4142$

13.3 POISSON DISTRIBUTION :

Poisson distribution was derived in 1837 by a French mathematician Simeon D. Poisson (1781-1840).

13.3.1 Conditions of Poisson Distribution :

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., $n \to \infty$

ii) p, the constant probability of success for each trial is indefinitely small i.e., $p \to o$

iii) np = m, (say), is finite

Under the above three conditions the binomial probability function (table 3.1) tends to the probability function of the poisson distribution given below :

$$p(r) = P(x = r) = \frac{e^{-m}.m^{r}}{r!}\text{, r=0, 1, 2, 3........ (1)}$$

where x is the number of successes (occurences of the event), m = np and e = 2.71828 (the base of the system of Natural Logarithms)

r! = r(r-1) (r-2).....x 3 x 2 x 1.

Derivation of equation 1 : We shall obtain the limiting form of the binomial probability function (table 3.1) under the conditions :

$$n \to \infty \text{ and } np = m \Rightarrow p = \frac{m}{n} \text{ and } q = 1 - \frac{m}{n}$$

Probability function of binomial distribution is

$$^{n}C_{r}p^{r}q^{n-r} = \frac{n!}{r!(n-r)!}p^{r}q^{n-r}$$

$$= \frac{n(n-1(n-2).....[n-(r-1)]}{r!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{n!}\frac{n}{n}.\frac{n-1}{n}.\frac{n-2}{n}.....\frac{n-(r-1)}{n}.\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{n!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)......\left(1-\frac{r\text{-}1}{n}\right) x \left(1-\frac{m}{n}\right)^n x \left(1-\frac{m}{n}\right)^{-r}$$

Taking the limit as $n \to \infty$, the limiting form of binomial probability function becomes

$$\frac{m^r}{r!}.\lim_{n\to\infty}\left(1-\frac{1}{n}\right) x \lim_{n\to\infty}\left(1-\frac{2}{n}\right) x.....x \lim_{n\to\infty}\left(1-\frac{r-1}{n}\right) x \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n x \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r}$$

$$\frac{m^r}{r!} x\ (1-0)\ x\ (1-0)\ x...x\ (1-0)\ x \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n x \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r} \quad ....(*)$$

But we know that :

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a \text{ and } \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^A = 1 \qquad ....(2)$$

If A is constant independent of n. substituting these values in (*), we get the limiting form of binomial probability function as

$$\frac{m^r}{r!} x\ 1\ x\ e^{-m} x\ 1 = \frac{e^{-m}m^r}{r!}$$

Hence the probability function of the Poisson distribution is

$$p(r) = P(x = r) = \frac{e^{-m}m^r}{r!} : r = 0,\ 1,\ 2,\ 3,......\text{ as stated in equation (1)}$$

Remarks : 1) Poisson distribution is a discrete probability distribution, since the variable x can take only integral values 0, 1, 2, ... ∞

2) Putting r = 1, 2, 3,,..., in equation (1), we obtain the probabilities of 0, 1, 2, 3, ..., successes respectively, which are given in the table 3.5

Table 3.5 : Poisson Probabilities

No. of Successes (r)	Probability p(r)
0	$\frac{e^{-m}.m^{0}}{0!}=e^{m}$
1	$\frac{e^{-m}.m^{1}}{1!}$
2	$\frac{e^{-m}.m^{2}}{2!}$
3	$\frac{e^{-m}.m^{3}}{3!}$
:	:
:	:

3. Total probability is 1.

$$\sum_{r=0}^{\infty} p(r) = e^{-m} + me^{-m} + \frac{m^2}{2!}e^{-m} + \frac{m^3}{3!}e^{-m} + ...$$

$$= e^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right]$$

$$= e^{-m} \text{ x } e^{m} = e^{-m+m} = e^{0} = 1$$

$$\left[\therefore e^{x} = 1 + x\frac{x^2}{2!} + \frac{x^3}{3!} + ...\right] \quad ....(3)$$

4. If we know m, all the probabilities of the poisson distribution can be obtained. m is, therefore, called the parameter of the poisson distribution.

We obtained above the poisson distribution as a limiting case of the binomial distribution. Below we give the general model underlying poisson distribution.

Let us consider the random experiment with the following conditions :

i. The probability of success (i.e., occurence of an event) in the small time interval (t, t + dt) is m.dt i.e., it is directly proportional to the magnitude of the interval.

ii. The probability of getting more than one successes in this time interval is very small.

iii. The probability of any particular success in the time interval (t, t + dt) is independent of the actual time to and also of all previous successes.

Under the above conditions, it has been established that the probability of r successes in time interval of magnitude t is given by

$$P_r(t) = \frac{e^{-mt}(mt)^r}{r!}, \ r = 0 \ , \ 1, \ 2, \ 3.... \quad (3a)$$

which is the probability function of the poisson distribution with parameter mt.

13.3.2 Utility or Importance of Poisson Distribution :

The conditions under which poisson distribution is obtained as a limiting case of the binomial distribution and also the conditions for the general model underlying poisson distribution suggest that poisson distribution can be used to explain the behaviour of the discrete random variables where the probability of occurence of the event is very small and the total number of possible cases is sufficiently large. As such poisson distribution has found application in a variety of fields such as queuing theory (waiting time problems), Insurance, Physics, Biology, Business, Economics and Industry. Most of the Temporal Distributions (dealing with events which are supposed to occur in equal intervals of time) and the spatial distributions (dealing with events which are supposed to occur in intervals of equal length along a straight line) follow the poisson probability law. We give below some practical situations where poisson distribution can be used :

i. The number of telephone calls arriving at a telephone switch board in unit time (say, per minute).

ii. The number of customers arriving at the super market; say per hour.

iii. The number of defects per unit of manufactured product.

iv. To count the number of radio-active disintegrations of a radio-active element per unit of time (Physics).

v. To count the number of bacteria per unit (Biology)

vi. The number of defective material say, pens, blades etc. in a packing manufactured by a good concern.

vii. The number of suicides reported in a particular day or the number of casualities (pesons dying) due to a rare disease such as heart attach or cancer or snake bite in a year.

viii. The number of accidents taking place per day on a busy road.

ix. The numer of typographical errors per page is a typed material or the number of printing mistakes per page in a book.

(i), (ii), (iv), (vii) and (viii) are examples of temporal distributions and the remaining are examples of spatial distributions.

13.3.3 Constants of Poisson Distribution :

$$\text{Mean} = \sum_{r-0}^{\infty} rp(r)$$

$$= me^{-m} + 2.\frac{m^2e^{-m}}{2!} + 3.\frac{m^3e^{-m}}{3!} + 4.\frac{m^4e^{-m}}{4!} + ...$$

$$= me^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right]$$

$= me^{-m}$ x e^{m} (using equation (3))

$= me^{-m+m} = me^{0}$

$= m \quad (\therefore e^0 = 1)$ ....(4)

Table 3.6

r	p(r)	rp(r)	$r^2p(r)$
0	e^{-m}	0	0
1	me^{-m}	$1.\ me^{-m}$	me^{-m}
2	$\frac{m^2e^{-m}}{2!}$	$2.\ \frac{m^2e^{-m}}{2!}$	$2^2.\ \frac{m^2e^{-m}}{2!}$
3	$\frac{m^3e^{-m}}{3!}$	$3.\ \frac{m^3e^{-m}}{3!}$	$3^2.\ \frac{m^3e^{-m}}{3!}$
4	$\frac{m^4e^{-m}}{4!}$	$4.\ \frac{m^4e^{-m}}{4!}$	$4^2.\ \frac{m^4e^{-m}}{4!}$

$$\text{Variance } = \sum r^2p(r) - \left[\sum rp(r)\right]^2$$

$$= \sum r^2p(r) - (\text{mean})^2$$

$$= \sum r^2p(r) - m^2 \qquad \text{............(*)}$$

$$\sum r^2p(r) = me^{-m} + 2^2.\frac{m^2e^{-m}}{2!} + 3^2.\frac{m^3e^{-m}}{3!} + 4^2.\frac{m^4e^{-m}}{4!} + ...$$

$$= me^{-m}\left[1 + 2m + \frac{3}{2!}m^2 + \frac{4}{3!}m^3 + ...\right]$$

$$= me^{-m}\left[\left\{1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right\} + \left\{m + \frac{2m^2}{2!} + \frac{3m^3}{3!} + ...\right\}\right]$$

$$= me^{-m}\left[\left\{1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right\} + m\left\{1 + m + \frac{m^2}{2!} + \frac{3m^3}{3!} + ...\right\}\right]$$

$$= me^{-m}\left[e^m + me^m\right] = me^{-m}.e^m(1 + m) = m(1 + m)e^0$$

(using equation 3)

= m (1 + m)

substituting in (*) we get

Variance = $m(1+m) - m^2 = m + m^2 - m^2 = m$ ......(5)

Hence for the poisson distribution with parameter m,

we have Mean = Variance = m ....(6)

i.e., mean and variance are equal, each being equal to the parameter m.

Other constants : The moments (about mean) of the poisson distribution are

:

$$\mu_1 = 0\,;\ \mu_2 = \text{variance} = m\,;\ \mu_3 = m;\ \mu_4 = m + 3m^2$$

Hence, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$ and $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$ ....(7)

$$\beta_1 = \frac{\mu_4}{\mu_2^2} = \frac{m+3m^2}{m^2} = 3 + \frac{1}{m} \Rightarrow \gamma_1 = \beta_2 - 3 = \frac{1}{m} \qquad ....(8)$$

13.3.4 Mode of Poisson Distribution :

The poisson distribution has mode at X = r,

if p(r) > p(r - 1) and p(r) >P(r + 1)

Case (1) : When m is an integer : If m is an integer, equal to k, (say), then the poisson distribution is bimodal, the two modes being at the points X = k and X = k - 1

Case (2) : When m is not an integer : If m is not an integer, then the distribution is unimodal, the unique modal value being the integral part of m. For example, if m = 5.6, then mode is 5, the integral part of 5.6.

Example : 1) If the standard deviation of a poisson variance X is $\sqrt{2}$, then the probability that X is strictly positive is

(i) e^2 (ii) e^{-2} (iii) $1 - e^{-\sqrt{2}}$ (iv) None of the these

Solution : Let $X - P(\lambda)$. We know that for poission distribution with parameter λ,

$$\text{Variance} = \lambda = \left(\sqrt{2}\right)^2 = 2 \qquad \left[\therefore \text{s.d.} = \sqrt{2}\,(Given)\right]$$

$$\therefore \qquad P(X = r) = \frac{e^{-\lambda}.\lambda^r}{r!} = \frac{e^{-2}.2^r}{2!}\ ;\ r = 0,\ 1,\ 2, \ldots.\ (*)$$

The probability that X is strictly positive is given by :

$P(X > 0) = 1 - P(X = 0) = 1 - e^{-2}$ (from (*))

Hence, (iv) is the correct answer.

Example : 2) If 5% of the electric bulbs manufactured by a company are defective, use poisson distribution to find the probability that in a sample of 100 bulbs.

i) none is defective

ii) 5 bulbs will be defective (Given : $e^{-5} = 0.007$)

Solution : Here we are given :

n = 100

p = probability of a defective bulb = 5% = 0.05

Since p is small and n is large, we may approximate the given distribution by poisson distribution.

Hence, the parameter m of the poisson distribution is :

m = np = 100 x 0.05 = 5

Let the random variable X denote the number of defective bulbs in a sample of 100. Then (by poisson probability law)

$$P(x = r) = \frac{e^{-m}m^r}{r!} = \frac{e^{-5}5^r}{r!}\ ;\ r = 0,\ 1,\ 2,\ \ldots.\ (*)$$

i) The probability that none of the bulbs is defective is given by :

$P(x = 0) = e^{-5} = 0.007$ (from (*))

ii) The probability of 5 defective bulbs is given by :

$$P(x = 5) = \frac{e^{-5}\ x\ 5^5}{5!} = \frac{0.007\ x\ 625}{24} = \frac{4.375}{24} = 0.1823$$

13.3.5 Fitting of Poisson Distribution :

If we want to fit a poisson distribution to a given frequency distribution, we compute the mean $\overline{x}$ of the given distribution and take it equal to the mean of the fitted (poisson) distribution i.e., we take $m = \overline{x}$. One 'm' is known, the various probabilities of the poisson distribution can be obtained, the general formula being

$$p(r) = P(x = r) = \frac{e^{-m} \text{ x } m^{r}}{r!} ; r = 0, 1, 2, 3, ... \quad (9)$$

If N is the total observed frequency, then the expected or theoretical frequencies of the poisson distribution are given by Nx p(r).

Expected frequencies can be very conveniently computed as explained in the Table 3.7

Table 3.7 : Expected Poisson Frequencies

Value of variable(r)	Probability p(r)	Expected or theoretical poisson frequencies f(r) = N p(r)
0	$p(0) = e^{-m}$	$f(0) = Np(0) = Ne^{-m}$
1	$p(1) = me^{-m} = mp(0)$	$f(1) = m.Np(0) - mf(0)$
2	$p(2) = \frac{m^2 e^{-m}}{2!} = \frac{m}{2} me^{-m} = \frac{m}{2} p(1)$	$f(2) = \frac{m}{2}.Np(1) = \frac{m}{2} f(1)$
3	$p(3) = \frac{m^3 e^{-m}}{3!} = \frac{m}{3} \frac{m^2 e^{-m}}{2!} = \frac{m}{3} p(2)$	$f(3) = \frac{m}{3}.Np(2) = \frac{m}{2} f(2)$
4	$p(4) = \frac{m^4 e^{-m}}{4!} = \frac{m}{4} \frac{m^3 e^{-m}}{3!} = \frac{m}{4} p(3)$	$f(4) = \frac{m}{4}.Np(3) = \frac{m}{4} f(3)$

Example : 1) Fit a poisson distribution to the following data and calculate the theoretical frequencies :

x :	0	1	2	3	4
f :	123	59	14	3	1

Solution :

x	0	1	2	3	4	
f	123	59	14	3	1	$\sum f = 200$
fx	0	59	28	9	4	$\sum fx = 100$

$$\therefore \qquad x = \frac{\sum fx}{\sum f} = \frac{100}{200} = 0.5$$

Thus the mean (m) of the theoretical (poisson) distribution is $m = \bar{x} = 0.5$. By poisson probability law, the theoretical frequencies are given by :

$$f(r) = Np(r) = 200.\frac{e^{-m}m^{r}}{r!}\ ; r = 0, 1, 2, 3.....$$

$$f(0) = Np(0) = 200 \text{ x } e^{-m} = 200 \text{ x } e^{-0.5} = 200 \text{ x } 0.6065 = 121.3$$

Table 3.8 : Computation of Expected Frequencies

x	Expected poisson frequencies N.p(n)	
0	Np(0) = 121.3	= 121
1	Np(1) = Np(0) x m = 121.3 x 01.5 = 60.65	= 61
2	$np(2) = Np(1) \text{ x} \frac{m}{2} = \frac{60.65 \text{ x } 0.5}{2} = 15.3125$	= 15
3	$Np(3) = Np(2) \text{ x} \frac{m}{3} = \frac{15.3125 \text{ x } 0.5}{3} = 2.552$	= 3
4	$Np(4) = Np(3) \text{ x} \frac{m}{4} = \frac{2.552 \text{ x } 0.5}{4} = 0.32$	= 0
	Total	200

The standard normal probability curve is symmetric about the line z = 0.

13.5 LESSON END EXERCISE :

Q1. Explain the binomial distribution properties ?

Q2. A coin is tossed six times. What is the probability of obtaining four or more heads ?

Ans. 0.344

Q3. The following data show the number of seeds germinating out of 10 on damp filter for 80 set of seeds.

Fit on binomial distribution of this data :

X:	0	1	2	3	4	5	6	7	8	9	10
Y:	6	20	28	12	8	6	0	0	0	0	0

Ans = 80.1

Q4. The following mistakes per page were observed in a book :

No. of mistakes per page	0	1	2	3	4
No. of times mistakes occurred	211	90	19	5	0

Fill a poisson distribution to the given data.

SUGGESTED FOR FURTHER READING :

- Das, N.G. (2012), statistical methods, Tata McGraw Hill Education Private Limited, New Delhi.
- Douglas et.al (2008), Statistical Techniques in Business and Economis, Tata McGraw, Hill Publishing Company Limited, New Delhi.
- Yamane, T. (1990), Statistics : An introductory analysis, Harper and Row, New York.
- Rao, C.R. (1965), Linear Statistical INference and Applications, Willey and Sons.
- Goon, A.M. M.K. Gupta and B. Dasgupta (1993), Fundamentals of statistics Vol. I, The World Press, Calcutta

****

M.A. Economics **Semester - IInd** **Lesson - 14**

Course No. 203 **Unit - III**

NORMAL DISTRIBUTIONS, CHI-SQUARE AND F-DISTRIBUTION

STRUCTURE :

14.1 NORMAL DISTRIBUTION:

The distributions discussed so far, viz., binomial distribution and poisson distribution, are discrete probability distributions, since the variables under study were discrete random variables. Normal distribution is a continuous probability distributions which arise when the underlyig variable is a continuous one. Normal probability distribution or commonly called the normal distribution is one of the most important continuous theoretical distribution in statistics. Most of the data relating to economic and business statistics or even in social and physical sciences conform to this distribution. It was first discovered by English Mathematician De-Moivre (1667-1754) in 1733 who obtained the mathematical equation for this distribution while dealing with problems arising inthe game of chance. Normal distribution is also known as Gaussian distribution (gaussian Law of Errors) after Karl Friedrich Gauss (1777-1855) who used this distribution to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Today, normal probability model is one of the most important probability models in statistical analysis.

14.1.1 Equation of Normal Probability Curve :

If x is a continuous random variable following normal probability distribution with mean μ and standard deviation σ, then its probability density function (p.d.f) is given by

$$\text{p(x)} = \frac{1}{\sqrt{2\pi.\sigma}} e^{-\frac{1}{2}} \frac{-(x-\mu)^2}{\sigma}, -\infty < x < \infty \qquad \text{.........(a)}$$

$$\text{p(x)} = \frac{1}{\sqrt{2\pi.\sigma}} e \frac{-(x-\mu)^2}{2\sigma^2}, -\infty < x < \infty \qquad \text{.......(1a)}$$

where π and e are the constants given by

$$\pi = \frac{22}{7}, \sqrt{2\pi} = 2.5066$$

and e = 2.71828 (the base of the system of natural logarithms)

Remark : The mean μ and standard deviation σ are called the parameters of the normal distribution.

14.1.2 Standard Normal Distribution :

If x is a random variable with mean μ and standard deviation σ, then the random varable z defined as follows :

$$Z = \frac{X - E(x)}{\sigma_n} = \frac{X - \mu}{\sigma} \qquad \text{............(2)}$$

is called the standard normal variate (S.N.V.). We have :

$$E(Z) = E\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}E(x-\mu) \qquad [\therefore \Sigma(cx) = cE(x)]$$

$$= \frac{1}{\sigma}[E(x) - E(\mu)] = \frac{1}{\sigma^2}(\mu - \mu) = 0$$

$$\text{Var }(z) = \text{var } = \left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{ Var }(x - \mu)$$

$$[\therefore \text{Var}(cx) = c^2\text{Var}(x)]$$

$$= \frac{1}{\sigma^2}.\text{Var }(x)$$

$$\therefore \qquad \text{Var }(z) = \frac{1}{\sigma^2}.\sigma^2 = 1$$

Therefore, the standard variate (S.N.V.) Z has mean 0 and standard deviation 1.

Hence the probability density function (p.d.f.) of S.N.V.Z. is given by :

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, -\infty < z < \infty \qquad \text{.......... (3)}$$

(Taking x = z, μ = 0 and σ = 1 in equation 1)

This gives the height (ordinate) of standard normal curve at the point z.

The standard normal probability curve is symmetric about the line z = 0.

Why do we need standard normal distribution ?

A normal distribution is characterized by two parameter (constant)

i) The mean (μ) whose position can be located anywhere on the x-axis and

ii) The standard deviation (σ) which determines the spread of its bell shape curve along the x-axis.

If we want to construct tables to compute the areas, say, $P(a \le x \le b)$ under any normal probability curve $N(\mu, \sigma^2)$, we need to construct an infinite number of tables for different combinations of the values of μ and σ, which is practically impossible. In practice, we standardize the variable x to obtain the z-scores $Z = \left(\frac{x - \mu}{\sigma}\right)$ and then construct the table for the areas under the standard normal probability curve. Hence, by rescaling the normal distribution axis, any normal distribution can be converted into standard normal distribution with mean 0 and SD = 1. Consequently we need only one table of areas under standard normal curve. Thus, for any normal distribution with mean μ and sd σ, this table can be used for obtaining the areas under standard normal curve for almost any interval along the z-axis.

14.1.3 Properties of Normal Distribution :

The normal probability curve with mean μ and standard deviation σ is given by

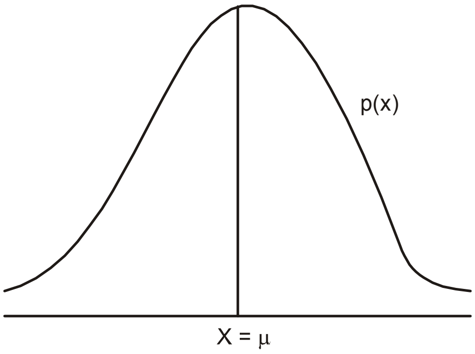
$$p(x) = \frac{1}{\sqrt{2\pi}.\sigma}.e^{-(x-\mu^2)/2\sigma 2}, -\infty < x < \infty \qquad \text{........ (*)}$$

The standard normal probability curve is given by the equation :

$$\phi(z) = \frac{1}{\sqrt{2\pi}}.e^{-z^2/2}, -\infty < x < \infty \qquad \text{............(**)}$$

It has the following properties :

1. The graph of p(x) is the famous bell shaped curve as showin in the Fig. The top of the bell is directly above the mean (μ)



2. The curve is symmetrical about the line $x - \mu, (z = 0)$, i.e.,... it has the same shape on either side of the line $x - \mu$ (or $z = 0$).

   This is because the equatin of the curve $\phi(z)$ remains unchanged if we change z to -z.

3. Since the distribution is symmetrical, mean, median and mode coincide. Thus,

   $$\text{Mean} = \text{Median} = \text{Mode} = \mu$$

4. Since Mean = Median = μ, the ordinate at $X = \mu$, $(z = 0)$ divides the whole area into two eqyal parts. Further, since total area under normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate at $X = \mu$ (or $z = 0$) is 0.5

5. Also, by virtue of symmetry, the quartiles are equidistant from median (μ), i.e.,

   $$Q_3 - Md = Md - Q_1 \Rightarrow Q_1 + Q_3 = 2\ Md = 2\mu \qquad \text{..........} (4)$$

6. Since the distribution is symmetrical, the moment coefficient of skewness is given by :

   $$\beta_1 = 0 \Rightarrow \gamma_1 = 0 \qquad \text{..........}(5)$$

7. The coefficient of kurtosis is given by :

   $$\beta_2 = 3 \Rightarrow \gamma_2 = 0 \qquad \text{.........}(6)$$

8. No portion of the curve lies below the x-axis, since p(x) being the probability can never be negative.

9. Theoretically, the range of the distribution is from $-\infty$ to ∞. But practially, Range = 60°.

10. As x increases numerically [i.e., on either side of $x = \mu$], the value of p(x) decreases rapidly, the maximum probability occuring at $x = \mu$ and is given by (put $x = \mu$ in (*)]

$$[p(x)]_{\max} = \frac{1}{\sqrt{2\pi}.\sigma}$$

Thus, maximum value of p(x) is inversely proportional to the standard deviation. For large value of σ, p(x) decreases, i.e., the curve tends to flatten out and for small values of σ, p(x) increases, i.e., the curve has a sharp peak.

Figure : 3.5 (a), 3.5 (b) and 3.5(c) give normal probability curves with moderate, large and small values of σ respectively.

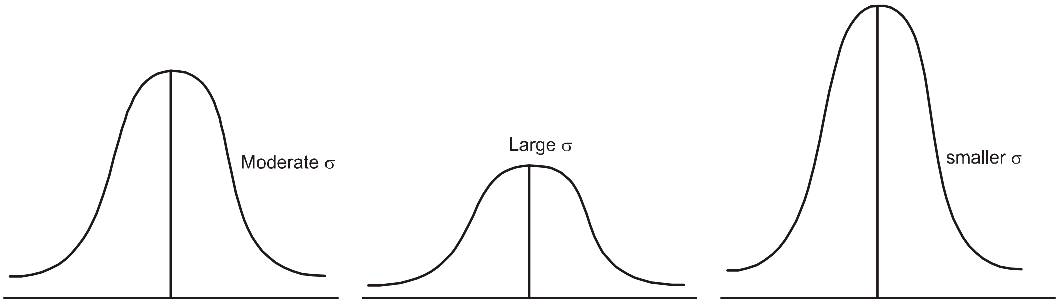


Fig. 3.5(a) Fig. 3.5(b) Fig. 3.5©

11. Distribution is unimodal, the only mode occurring at $x = \mu$.

12. Since the distribution is symmetrical, all moments of odd order about the mean are zero. Thus

$$\mu_{2n+1} = 0;; (n = 0, 1, 2, ...) \quad ....(7)$$

i.e. $$\mu_1 = \mu_3 = \mu_5 = ... = 0 \quad ....(7a)$$

13. The moments (about mean) of even order are given by :

$$\mu_{2n} = 1.3.5...(2n-1)\sigma^{2n}, (n = 1, 2, 3, ...) \quad ....(8)$$

Putting n = 1 and 2 we get

$$\mu_2 = \sigma^2 \text{ and } \mu_4 = 1.3\sigma^4 = 3\sigma^4 \qquad ....(9)$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3 \qquad ....(9a)$$

14. X-axis is an asymptote tothe curve, i.e., for numerically large value of X on either side of the line (x = μ .), the curve becomes parallel to the X-axis and is supposed to meet it at infinity.

15. A linear combination of independent normal variates is also a normal variate. If X_1, X_2,....X_n are independent normal variates with means μ_1, μ_2, ..., μ_n and standard deviation σ_1, σ_2,..., σ_n respectively, then their linear combination.

$$a_1x_1 + a_2x_2+,....+a_nx_n \qquad ....(10)$$

where a_1, a_2,...a_n are cosntants is also a normal variate with

$$\text{Mean} = a_1\mu_1 + a_2\mu_2....+ a_n\mu_n$$

and $\quad$ $\text{Variance} = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + ...+ a_n^2\sigma_n^2$ $\quad$ }........(10a)

In particular, if we take a_1, a_2,...a_n = 1 in equation (10) then we get :

"X_1 + X_2 +....+ X_n is a normal variate with mean $\mu_1 + \mu_2 + ... + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$

Thus, the sum of independent normal variates is also a normal variate. This is known as the 'Reproductive or Additive Property' of the normal distribution.

If we take $a_1 = a_2 = 1$ and $a_3 = a_4 = ... = a_n = 0$, then we have equation (10) and (10a)

$X_1 + X_2$ is a normal variate with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

Further if we take $a_1 = 1$ and $a_2 = 1$ and $a_4 = a_4 =...= a_n = 0$, in (equation 10) and (quation 10a), we get;

"X_1 - X_2 is a normal variate with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

Hence, the sum as well as teh difference of independent normal variates is a normal variate.

16. Mean Deviation (M.D.) about mean as median or mode,

($\therefore$ M = Md = Mo) is given by :

$$\text{M.D.} = \sqrt{\frac{2}{\pi}}.\sigma = 07979\sigma = \frac{4}{5}\sigma \qquad ....(11)$$

17. Quartiles are given (in terms of μ and σ) by :

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma \qquad ....(12)$$

18. Quartile Deviation (Q.D.) is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = 0.6745\sigma = \frac{2}{3}\sigma \qquad \text{(from equation 12 and 13)}$$

Also

$$\text{Q.D.} = \frac{2}{3}\sigma = \frac{4}{6}\sigma = \frac{5}{6}\text{x}\frac{5}{6}\sigma = \frac{5}{6}\sigma \text{ M.D...............(from equation 11)}$$

$$\therefore \qquad \text{Q.D.} = \frac{5}{6}\text{ M.D.} \qquad ....(14)$$

19. We have (approximately) :

$$\text{Q.D.} : \text{M.D.} : \text{S.D.} :: \frac{2}{3}\sigma; \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1$$

$$\Rightarrow \qquad \text{Q.D. : M.D. : S.D. : : 10 : 12 : 15}$$

20. From equatio (11) and equation (14) we also have

4 S.D. = 5 M.D. = 6 Q.D. ....(16)

21. Points of inflexion of the normal curve are at $X = \mu \pm \sigma$ i.e., they are equidistant from mean at a distance of σ and are given by :

$$\left[\text{x} = \mu \pm \sigma, \text{p(x)} = \frac{1}{\sigma.\sqrt{2\pi}} e^{-1/2} \right]$$

22. Area property : One of the most fundamental properties of the normal probability curve is the area property. The area under the normal probability curve between the ordinates at $x = \mu - \sigma$ and $x = \mu + \sigma$ is 0.6826. In other words, the range $\mu \pm 6$ covers 68.26% of the observations. The area under

the normal probability curve between the ordinates at $x = \mu - 3\sigma$ and $x = \mu + 3$ is 0.9973 i.e., the range $\mu \pm 3\sigma$ covers 99.73% of the observations. Hence, for practical purposes, the range $\mu \pm 3\sigma$ covers the entire area, which is 1 (or all the observations).

The standard normal variate corresponding to X is $Z = \dfrac{x - \mu}{\sigma}$

when $X = \mu + \sigma, z = \dfrac{\mu + \sigma - \mu}{\sigma} = 1;$

when $X = \mu - \sigma, z = \dfrac{\mu + \sigma - \mu}{\sigma} = 1;$

when $X = \mu + 2\sigma, z = \dfrac{\mu + 2\sigma - \mu}{\sigma} = 2;$

when $X = \mu - 2\sigma, z = \dfrac{\mu - 2\sigma - \mu}{\sigma} = -2;$

when $X = \mu + 3\sigma, z = \dfrac{\mu + 3\sigma - \mu}{\sigma} = 3;$

when $X = \mu - 3\sigma, z = \dfrac{\mu - 3\sigma - \mu}{\sigma} = -3;$

Hence the area under the standard normal probability curve

i) Between the ordiantes at $Z = \pm 1$ is 0.6826

ii) Between the ordinates at $Z = \pm 2$ is 0.9544

iii) Between the ordinates at $Z = \pm 3$ is 0.9973

These areas are exhibited in the Fig. 3.6

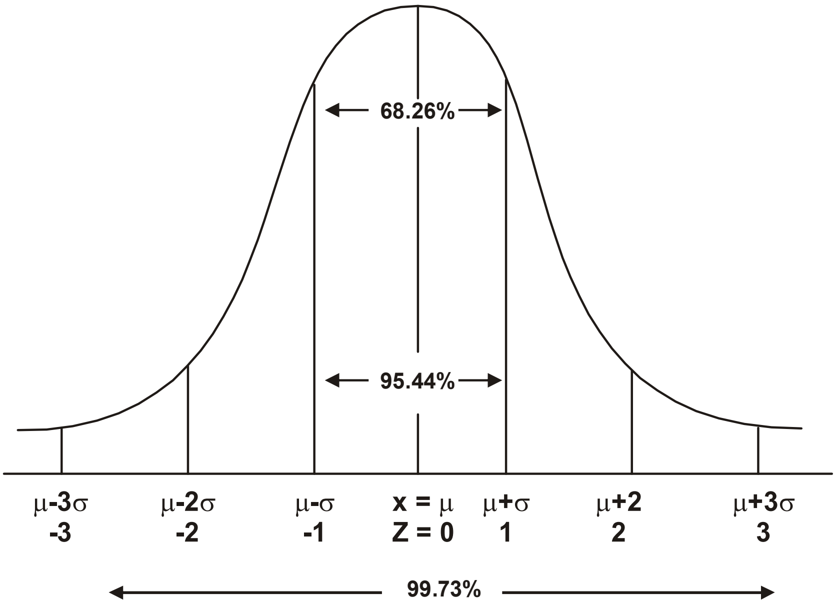


Fig. 3.6 : Areas Under Normal Probability Curve

The table 3.9 gives the areas under the standard normal probability curve for some important values of Z :

Table 3.9 : Areas under standard normal curve

Distance from the mean ordiantes in terms of $\pm\sigma$	Area under the curve
Z = ± 0.6745	50% = 0.50
Z = ± 1.00	68.26% = 0.6826
Z = ± 1.96	95% = 0.95
Z = ± 2.0	95.44% = .9544
Z = ± 2.58	99% = 0.99
Z = ± 3.0	99.73% = 0.9973

14.1.4 Importance of Normal Distribution :

Normal distribution has occupied a very important role in statistics. We enumerate below some of its important applications.

1. If X is a normal variate with mean μ and variance σ^2, then we have proved that .

   $$P(\mu - 3\sigma < x < \mu + 3\sigma) = P(-3 < z < 3) = 0.9973$$

   $\Rightarrow$ $[\ |\ Z\ |\ > 3] = 1 - 0.9973 = 0.0027$

   Thus, the probability of standard normal variate going outside the limits ± 3 is practically zero. In other words, in all probability, we should expect a standard normal variate to lie between the limits ± 3. This property of the normal distribution forms the basis of entire large sample theory :

2. Most of the discrete probability distribution (e.g. binomial and poisson distribution) tend to normal distribution as n, the number of trials increases. For large values of n, computations of probability for discrete distributions becomes quite tedious and time consuming. In such cases, normal approximation can be used with great ease and convenience.

3. Almost all the exact sampling distributions, e.g., student's t-distribution, snedecor's F-distribution, Fisher's z-distribution and chi-square distribution conform to normal distribution for large degrees of freedom (i.e., as).

4. The whole theory of exact sample (small sample) tests, viz., t, f, x^2 tests, etc. is based onthe fundamental assumption that the parent population from which the samples have been drawn follows normal distribution.

5. Perhaps, one of the most important applications of the normal distributions is inherent in one of the most fundamental theorems in the theory of statistics, viz., the central limit theorem which may be stated as follows :

   "If x_1, x_2, ...x_n are 'n' independent random variables following any distribution, then under certain very general conditions, their sum is asymptotically normally distributed, i.e., follows normal distribution "

   An immediate consequences of this theorem is the following result :

   "If x_1, x2, ...x_n is a random sample of size 'n' from any population with mean and variance , then the sample mean.

is asymptotically normal (as) with mean and variance ”

6. Normal distribution is used in statistical quality control in industry for the setting of control limits for the construction of control charts.

Example :

1) Suppose the waist measurement w of 800 girls are normally distributed with mean 66 cms, and standard deviation 5 cms. Find the number of N of girls with waist -

   i) between 65 and 70 cms;

   ii) greater than or equal to 72 cms.

Solution : W : Waist measurements (in cms) of girls.

We are given , where cms and cms

w(in cms)	65	70	72
(standard Normal variate)			

i) The probability that a girl has waist between 65cms and 70cms is given by :

= (By symmetry)

= 0.0793 + 0.2881 = 0.3674 (from normal tables)

Hence in a group of 800 girls, the expected number of girls with waists between 65cms and 70cms is 800 x 0.3674 = 293.92 = 294

ii) The probability that a girl has waist greater than or equal to 72cms is given by

Hence, in a group of 800 girls the expected number of girls with waist greater then or equal to 72 cms is : 800 x 0.1151 = 92.08 = 92

14.2 CHI-SQURE DISTRIBUTION :

The square of standard normal variate is known as chi-square distribution with 1 degree of freedom

The square of standard normal variate is known as chi-square distribution with 1 degree of freedom

14.2.1 General Form

14.2.2 Probability density function of

14.2.3 Application of Chi-squre

Application of Chi-squre X^2 :- It is based on X^2 distribution and is a parametric test. It is a technique through the use of which it is possible for all researcher to

(i) Test the hypothetical value of the population variance

(ii) Test the goodness of fit

(iii) Test the independence of attributes

(iv) Test the homogeneity of independent estimates of the population variance.

14.3 F-DISTRIBUTION

14.3.1 Definition

If X and Y are two independent chi-square variates with v_1 and v_2 degrees of freedom

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (v_1,v_2) degrees of freedom.

14.3.2 Probability Function

The probability function of F distribution is given as

$$f(F) = \frac{(\frac{V1}{V2})^{V\frac{1}{2}}}{B(\frac{V1}{2},\frac{V2}{2})} . \frac{F^{V\frac{1}{2}-1}}{(1+\frac{V1}{V2}F)^{\frac{V1+V2}{2}}} ; 0 \leq F \leq \infty$$

14.3.3 Application of F-test

F-test is based on F-distribution and is used for the following :

(i) This test is used in the context of ANOVA for judging the significance of more than two sample mean. In it test statistic F is calculated and compared with its probable value for accepting or rejecting the null hyputhesis.

(ii) To test the significance of an observed multiple correlation co-efficient.

(iii) To test the significance of an observed sample correlation ratio.

(iv) To test the significance of equality of two population variances.

14.4 LESSON END EXERCISE

Q1. Explain the condition for normality.

Q2. Discuss the properties of Normal Distribution.

Q3. Define Chi-squre Distribution. Give the application of X^2 (Chi-squre).

Q4. Differentiate between Chi-squre and F-Distribution.

Q5. Difine F-Distribution and the probality function of F-Distribution.

SUGGESTED FOR FURTHER READING :

♦ Das, N.G. (2012), statistical methods, Tata McGraw Hill Education Private Limited, New Delhi.

♦ Douglas et.al (2008), Statistical Techniques in Business and Economis, Tata McGraw, Hill Publishing Company Limited, New Delhi.

♦ Gupta, S. C., Kapoor, V.K. (200), Fundamental of Mathematical Statistics, Sultan Chand and Sons Publisher House, New Delhi

♦ Gupta, S.P (2011), Statistical methods, Sultan Chand and Sons Educational Publisher, New Delhi..

....

M.A. Economics	Semester - IInd	Lesson - 15
Course No. 203		Unit - IV

CONCEPT OF AN ESTIMATOR AND ITS SAMPLING DISTRIBUTION, PROPERTIES OF A GOOD ESTIMATOR, ESTIMATING MEAN, PROPORTIONS, VARIANCE OF POPULATIONS FROM SAMPLES.

CHAPTER HIGHLIGHTS :

This chapter contains the concept, types and methods of an estimator. It also explains the properties of a good estimator.

CHAPTER OUTLINES :

15.1 INTRODUCTION :

When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences or generalisation about that population from the information embodied in the sample. Statistical estimation or briefly estimation is concerned with the methods by which population characteristics are estimated from sample information. It may be pointed out that the true value of a parameter is an unknown constant that can be correctly ascertained only by an exhaustive study of the population. However, it is ordinarily too expensive or it is infeasible to enumerate complete populations to obtain the required information. In case of finite populations, the cost of complete censuses may be prohibitive and in case of infinite population complete enumerations are impossible.

15.2 CONCEPT OF AN ESTIMATOR :

Estimator as a function of sample values. This function may simply be , i.e. sum of the sample values divided by the size of the sample, or it may be $(IIx)^{1/n}$ i.e.

the product of all sample values to the n-th root or some other. The symbol II (capital pi) is used to indicate product of values, just as indicates sum of these values. To be more precise we have

The function of sample values simply implies the way the sample values are combined to obtain the estimate. This function of sample values is also called statistic.

It will be good at this stage to distinguish between an 'Estimator' and an 'Estimate'. As noted above the estimator can notes the way the sample values are combined. It is the procedure. An estimate, on the other hand means the numerical value of the estimator for a given sample. The discrepancy between an estimator and the true parameter value is called the sampling error.

Sampling error of an estimator Estimator - the true parameter value

With respect to estimating a parameter the following two types of estimates are possible :

1. Point estimates
2. Interval estimates

15.2.1 Types of an estimate : Point Estimates :

A point estimate is a single number which is used as an estimate of the unknown population parameter. In point estimation, the estimated value is given by a single quantity, which is a function of sample observations. This function is called the estimator of the parameter and the value of estimator in a particular sample is called an estimate. Given a certain population we are considering estimation of some population parameter with the help of an estimator based on a sample. This is called point estimation because we provide a single numerical estimate of the parameter. We have seen that any estimator will have a sampling distribution because its value will be affected by sampling fluctuations. For example, the population mean (m) may be estimated either by sample A.M.() or by sample median or by sample G.M. or by sample H.M. or by sample mode. Similarly the population variance () may be estimated either by sample variance.

or by the sample mean deviation

where the two bars || indicate that absolute deviations are considered.

15.2.3 Interval estimates :

In interval estimation an interval within which the parameter is expected to lie is given by using two quantities based on sample values. This is known as confidence interval and the two quantities which are used to specify the interval are known as confidence limits. Interval estimation where we provide an interval which will cover the parameter value with specified probability. For example suppose we want to estimate the total yield of wheat in the next crop. In this case instead of providing a precise numerical estimate of the total yield it might be sufficient. If we can provide limits between which the total yield will lie with specified probability. An interval estimate of a population parameter is a statement of two values between which it is estimated that the parameter lies. An interval estimate would always be specified by two values i.e., the lower one and the upper one. In more technical terms, interval estimation refers to the estimation of a parameter by a random interval called the confidence interval whose end points Land U and L<U are functions of the observed random variables such that the probability that the inequality L< <U is satisfied in terms of pre-determined number, I- . L and U are called the confidence limits and are the random end points of interval estimate. Since in an interval estimate, we determine an interval of plausible values, hence the name interval estimation. Thus, on the basis of sample study. If we estimate the average income of the people living in a village as Rs. 875 it will be a point estimate. On the other hand, if we say that the average income could like between Rs. 800 and Rs. 950, it will be an interval estimate. Let $x_1, x_2,....,x_n$ be a random sample from a population of a known mathematical from which involves

an unknown parameter being included in the interval (t_1, t_2) has a given value say c.

such an interval, when it exists, is called a confidence interval for . The two quantities t_1 and t_2 which serve as the lower and upper limits of the interval are known as confidence limits. The probability (c) with which the confidence interval will include the true value of the parameter is known as confidence coefficient of the interval. The significance of confidence limits is that if many dependent random samples are drawn from the same population and the confidence interval is calculated from each sample, then the parameter will actually be included in the intervals in c proportions of cases in the long run. Thus the estimate of the parameter is stated as an interval with a specified degree of confidence.

The calculation of confidence limits is based on the knowledge of sampling distribution of an appropriate statistic. Suppose, we have a random sample of size n from a normal population , where the variance is known. It is required to find 95% confidence limits for the unknown parameter . We know that the sample mean follows normal distribution with mean and variance and so

has a standard normal distribution. Since 95% of the area under the standard normal curve lies between the ordinates at , we have

i.e. in 95% of cases the following inequalities hold separating out we get

the interval is known as the 95% confidence interval for and the 95% confidence limits are

Again, 99% of area under the standard normal curve lies between the ordinates at and 99.73% of the area lies between . Hence proceeding exactly in the same manner, the 99% confidence limits for are

Infact, using values from the normal probability integral table, confidence limits corresponding to any specified percentage can be obtained. These are exact confidence limits.

15.3 APPROXIMATE CONFIDENCE LIMITS (LARGE SAMPLE) ANY DISTRIBUTION :

1. For mean :

   95% confidence limits =

   99% confidence limits =

   Almost sure limits =

2. For proportion p :

   95% confidence limits =

99% confidence limits =

Almost sure limits =

3. For difference of means

95% confidence limits =

99% confidence limits =

Almost sure limits =

4. For difference of proportions (p_1 - p_2)

95% confidence limit =

99% confidence limit =

Almost sure limit =

Example :

A random sample of 100 ball bearings selected from a shipment of 2000 ball bearing has an average diameter of 0.354 inch with a S.D. = 0.048 inch. Find 95% confidence interval for the average diameter of these 2000 ball bearings.

Solution :

If a random sample of large size n is drawn without replacement from a finite population of size N, then the 95% confidence limits for the population mean are , where denotes the sample mean and

denoting the standard deviation (S.D.) of the population. Here,

Sample size (n) = 100

Sample mean ( ) = 0.354

Population size (N) = 2000

Sample S.D. (5) = 0.048

Since is not known, an approximate value of S.E. is obtained on replace the population S.D. () by the sample S.D. (S)

approximately

= 0.0047

The 95% confidence limits for the population mean are

=0.354 1.96 x 0.0047

= 0.354 0.0092

= 0.3632 and 0.3448

Thus, the 95% confidence interva is 0.3448 to 0.3632 inch.

Example :

A random sample of size 10 was drawn from a normal population with an unknown mean and a variance of 44.1 (inch)^2. If the observations are (inchs) : 65, 71, 80, 76, 78, 82, 68, 72, 65 and 81, obtain the 95% confidence interval for the population mean.

Solution :

We are given n = 10

= 73.8

Since the population S.D. is known using formula , 95%

confidence limits for are given by

= 77.9 and 69.7

The 95% confidence interval for is therefore 69.7 to 77.9 inches.

Example :

The standard deviation of a random sample of size 12 drawn from a normal population is 5.5. Calculate the 95% confidence limits for the standard deviation ( ) in the population. (Given $x^2_{0.975} = 3.82$ and $x^2_{0.025} = 21.92$ for 11 degree of freedom)

Solution :

Here n = 12 and the sample S.D. (s) = 5.5. Substituting the values in formula

, the 95% confidence

interval for is

or

i.e.

The 95% confidence limits for the population s.d. ( ) are 4.1 and 9.7

15.4 SAMPLING DISTRIBUTION OF AN ESTIMATOR :

If the population has a large number of individual units we will have a large number of possible samples and corresponding estimates of the population parameter. The estimates obtained from different samples will provide the sampling distribution of the estimator. Let us illustrate this point with the help of the following examples.

Example :

Suppose the population consists of three individual units and the variate values are :

$x_1 = 8,\ \ x_2 = 10, x_3 = 12$

As we can see the population mean is

and population variance is

=

=

Now suppose we would like to estimate the populatino mean (m) from a random sample of size two and we decide to use the sample as its estimator. Then the following results will be obtained :

Sample No.	Sample Values	Sample Mean
1	8, 10	9
2	8, 12	10
3	10, 12	11

The totality of estimates (values of) obtained from the different samples constitute the sampling distribution of . Similarly, if we decide to use the sample variance.

to be an estimate of we would be get the following results.

Sample No.	Sample Values	Values of s^2
1	8, 10	1.00
2	8, 12	4.00
3	10, 12	1.00

and this constitutes the sampling distribution of s^2.

We may consider estimating m by the sample median. Then we will compute the median for each sample and totality of sample medians obtained from different samples will provide the sampling distribution of the median. Similarly, we may obtain the distribution of any other estimator of m or

15.5 PROPERTIES OF A GOOD ESTIMATOR :

A distinction is made between an estimate and an estimator. The numerical value of the sample mean is said to be an estimate of the population mean figure. On the other hand, the statistical measure used, that is the method of estimation, is referred to as an estimator. For example the sample mean, is an estimator of the populationmean.

A good estimator, as common sense dictates is said to close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties.

155.5.1 Unbiasedness :

An estimator is said to be unbiased if its expected value is identical with the population parameter being estimated. That is if is an unbiased estimate of , then we must have . Many estimators are "Asymptotically unbiased" in the sense that the biases reduce to practically insignificant value zero when n becomes sufficiently large. The estimator s^2 is an example. It should be noted that bias in estimation is not necessarily undesirable. It may turn out to be an asset in some situations.

A statistic t is said to be an unbiased estimator of a parameter , if the expected value of t is .

Otherwise, the estimator is said to be ‘biased’. The bias of a statistic in estimating is given as.

Bias =

Let x_1, x_2.....x_n be a random sample drawn from a population with mean and variance . Then

Sample mean

Sample variance

The sample mean is an unbiased estimator of the population mean , because

The sample variance s^2 is a biased estimator of the population variance because

An unbiased estimator of the population variance is given by

The distribution between s^2 and s^2 in which only the denominators are different. S^2 is the variance of the sample observations, but s^2 is the ‘unbiased estimator’ of the variance ( ) in the population.

Example :

The following observations constitute a random sample from an unknown population. Estimate the mean and standard deviation of the population. Also, find the estimate of standard error of sample mean.

14, 19, 17, 20, 25

Solution :

The unbiased estimators of the population mean ( ) and the population variance ( ) are and respectively. Here n = 5 and

= = 19.

$= (-5)^2 + 0^2 + (-2)^2 + 1^2 + 6^2 = 66$

= 16.5

= 4.06

The estimates of and are 19 and 4.06 respectively.

The standard error of sample mean is . But as is not known it is estimated by s.

Estimate of

$$= 1.82$$

15.5.2 Consistency :

If an estimator, say , approaches the parameter closer and closer as the sample size n increases, is said to be a consistent estimator of . Stating somewhat more regorously, the estimator is said to be a consistent estimator of if, as n approaches infinity the probabiliyt 1 that will differ from the parameter by not more than an arbitrary small constant.

The sample mean is an unbiased estimator of no matter what from the population distribution assumes, while the sample median is an unbiased estimate of only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of in terms of both unbiasedness and consistency.

In case of large samples consistency is a desirable property for an estimator to posses. However in small samples, consistency is of little importance unless the limit of probability defining consistency is reached even with a relatively small size of the sample.

15.5.3 Efficiency and Minimum Variance :

Of two consistent estimators for the same parameter, the statistic with the smaller sampling variance is said to be "more efficient". Thus if t and t' are both consistent estimators of and var (t) < var (t') then t is 'more efficient' than t' in estimating because it is grouped more closely around and will on the average deviate less from .

If a consistent estimator exists whose sampling variance is less than that of any other consistent estimator, it is said to be 'most efficient', and it provides a standard for the measurement of "efficiency" of a statistic. If V_0 be the variance of the most efficient estimator and V be the variance of any other estimator, then the efficiency of the estimator is defined as

Efficiency =

obviously, the measure of efficiency cannot exceed 1. In sampling from a normal population N , both the sample mean and the sample median are consistent estimators of , but

, Var (median) =

(for large n). Since is smaller than var (median), mean is more efficient than median in estimating the parameter . It can be shown that the sample mean is the most efficient estimator.

Hence.

Efficiency of median = = 0.64 approx

A statistic t which has the minimum variance among all estimators of is called the minimum variance (mv) estimator. A statistc which is unbiased and has also the minimum variance (i.e., most efficient) is said to be the minimum variance unbiased estimator (MVUE). The variance of MVUE is often given by the R.H.S. of

In sampling from a normal population N the sample mean is the minimum variance unbiased estimator for the parament .

Let x_1, x_2....x_n be a random sample and

$T = a_1x_1 + a_2x_2 + ... + a_nx_n$

where $a_1, a_2.....a_n$ are constants. If the linear function T is an unbiased estimator of a parameter and also has the minimum variance, it is said to be the best linear unbiased estimator (BLUE). The sample mean ▯ is the best linear unbiased estimator of the population mean ▯.

Example :

If T_1, T_2, T_3 are independent, unbiased estimates of ▯ and all have the same variance which of the following unbiased estimates of ▯ would you prefer?

$$(T_1 + T_2 + T_3)/4, (2T_1 + T_2 + 2T_3)/5, (T_1 + T_2 + T_3)/3$$

Solution :

The last expression, viz. $(T_1 + T_2 + T_3)/3$ would be the most preferable, because it is the mean of T_1, T_2, T_3 and hence the best linear unbiased estimator, i.e., has the minimum variance among all the linear functions which may be proposed as unbiased estimator of ▯.

15.5.4 Sufficiency :

An estimator is said to be sufficient if it conveys as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exists, it is absolutely unnecessary to consider any other estimator, a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is being utilised.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two important methods are the least square method and the method of maximum likelihood.

15.6 METHODS OF POINT ESTIMATION :

15.6.1 Method of Maximum Likelihood :

This is a convenient methof for finding an estimator which satisfied most of the criteria discussed earlier. Let $x_1, x_2,....x_n$ be a random sample from a population with p.m.f. (for discrete case) or p.d.f. (for continuous case) f (x,▯), where ▯ is the parameter. Then the joint distribution of the sample observations viz.

$L = f(x_1, ▯), f(x_2, ▯), \ldots.., f(x_n, ▯)$

is called the likelihood function of the sample.

The method of maximum likelihood consists in choosing as an estimator of ▯ that statistic, which when substituted for ▯, maximises the likelihood function L. Such a statistic is called a maximum likelihood estimator (m.l.e.) we shall denote the m.l.e. of ▯ by the symbol $▯_0$.

since Log L is maximum when L is maximum, in practice the m.l.e. of ▯ is obtained by maximising log L. This is achieved by differentiating log L partially with respect to ▯ and using the two relations

▯

15.6.2 Properties of Maximum Likelihood Estimator :

1. The maximum likelihood estimator (M.L.E.) tends to be distributed normally for large samples.
2. The M.L.E. is invariant under functional transformations. This means that if 7 is an M.L.E., of ▯ and g(▯) is a function of ▯, then g(T) is the M.L.E. of g(▯).
3. The M.L.E. is consistent, most efficient, an also sufficient provided a sufficient estimator exists.
4. The M.L.E. is not necessarily unbiased. But when the M.L.E. is biased, by a slight modification, it can be converted into an unbiased estimator.

Example :

A tossed a biased coin 50 times and got head 20 times, while B tossed it 90 times and got 40 heads. Find the maximum likelihood estimate of probability of getting head when the coin is tossed.

Solution :

Let P be the unknown probability of obtaining a head. Using binomial distribution.

Probability of 20 heads in 50 tosses

Probability of 40 heads in 90 tosses

The likelihood function is given by the product of these probabilities :-

Hence,

The maximum likelihood estimate P_0 is therefore obtained by solving

This gives

15.6.3 Method of Moments :

The method of moments consists in equating the first few moments of the population with the corresponding moments of the sample i.e. setting

Where and . Since the parameters enter into the population moments, these relations when solved for the parameters give the estimates by the method of moments of course, this method is applicable only when the population moments exist. The method is generally applied for fitting theoreticaldistribution to observed data.

Example :

Estimate the parameter P of the binomial distribution by the method of moments (when n is known).

Solution :

For the binomial distribution . Also . Setting .

We have . Thus

i.e. the estimated value of P is given by the sample mean divided by the parameter n (known).

15.7 ESTIMATING THE MEAN :

The Finite Population Correction (FPC) factor is used to reduce the standard error by a value equal to . When developing confidence interval estimates for population parameters, the FPC factor is used when samples are selected without replacement. Thus, the confidence interval estimate for the mean is calculated as in equation.

..........(1)

To illustrate the finite population correction factor, refers to the confidence interval estimate for the mean developed for Saxon Plumbing Company. Suppose that in this month there are 5,000 sales invoices. Using =\$110.27, S = \$28.95, N =5,000, n = 100 and with 95% confidence, t_{99} = 1.9842. From equation (1)

= 110.27 5.744 (0.99)

= 110.27 5.69

Example :

A sample of 30 insulators were selected. Suppose a population of 300 insulators were produce by the company. Set up a 95% confidence interval estimate of the population mean.

Solution :

Using the finite population correction factor, with = 1,723.4 pounds, S=89.55, n = 30, N = 300, and t_{29} = 2.0452 (For 95% confidence)

= 1723.4 33.44 (0.9503)

= 1723.4 31.776

Here, because 10% of the population is to be sampled, the FPC factor has a small effect on the confidence interval estimate.

15.8 ESTIMATING THE PROPORTION :

In sampling without replacement, the confidence interval estimate of the proportion is defined in equation :

.... (2)

To illustrate the use of the finite population correction factor when developing a confidence interval estimate of the population proportion, consider again the estimate developed for Saxon Home Improvement Company. For these data, N = 5000, n = 100, s = 10/100 =- 0.10 and with 95%, confidence, Z = 1.96. Using equation (2)

=

= 0.10 (1.96) (0.03) (0.99)

= 0.10 0.582

In this case, because the sample is a very small fraction of the population, the FPC factor has virtually no effect on the confidence interval estimate.

15.9 ESTIMATING A POPULATION PROPORTION :

We use the sample of 1007 trials and consider the sample proportion = 85% as the best point estimate of the population proportion. Since we have no indication of how good our best estimate is, instead of using a single value, 0.85, we may use a range of values (or interval) that is likely to contain the true value of the population proportion. This is called a confidence interval.

With a confidence interval is associated a degree of confidence. The degree of confidence tells us the percentage of times that the confidence interval actually does contain the population parameter (e.g. proportion or mean), assuming that the estimation process is repeated a large number of times.

15.9.1 Assumptions :

1. The sample is a simple random sample.
2. The normal distribution can be used to approximate the distribution of sample proportions because and are both satisfied.
3. The conditions for a binomial distribution are satisfied. That is there is a fixed number of trials, the trials are independent, there are two categories of

outcomes, and the probabilities remain constant for each trial

15.9.2 Point Estimate of Population Proportion :

where

= p - hat is the sample proportion

x = sample of size n

15.9.3 Standard Deviation of Sample Proportions :

where = sample proportion

= p - hat is the sample proportion

= q - hat is the percent of those in the sample who do not have the quality under discussion.

Example :

The 0.95 (or 95%) degree of confidence interval estimate of the population proportion p is 0.826 < p < 0.874.

Solution :

It is incorrect to say that there is 95% chance that the true population proportion will fall between 0.826 and 0.874 because p is a constant, not a random variable. p has already occurred we just don't know what it is.

15.10 Let us sum up :

In this chapter we shall develop the technique which enables us to generalise the results of the sample of the population; to find how far these generalisations are valid and also to estimate the population parameters along with the degree of confidence. The answers to these and many other related problems are provided by a very important branch of statistics, known as the statistical inference. Estimation of population parameters like mean, variance, proportion correlation coefficient etc. from the corresponding sample statistics is one of the vary important problems of population parameters is imperative in making business decisions. In the estimation of parameters and also in the testing of hypothesis, the sampling distribution of a statistic and its standard error play a very important role.

15.11 LESSON END EXERCISE :

Q1. Find the estimates of and in the normal population by the method of moments ?

Ans. , S

Q2. A sample of 6500 screws is taken from a large consignment and 75 are found to be defective. Estimate the percentage of defectives in the consignment and assign limits with in which the percentage lies ?

Ans. 8.45% and 16.55%

Q3. 10 life insurance polices in a sample of 200 takes out of 50,000 were found to be insured for less than Rs. 5000. How many policies can be reasonably expected to be insured for less than Rs. 5,000 in the whole lot of 95% confidence level ?

Ans. 0.080 and 0.020

Q4. Distinguish between point estimation and interval estimation ?

Q5. Explain the concept of confidence interval confidence limits and confidence coefficient ?

****

M.A. Economics	Semester - IInd	Lesson - 16
Course No. 203		Unit - IV

TESTING OF HYPOTHESIS : FORMULATION OF STATISTICAL HYPOTESES - NULL AND ALTERNATIVE HYPOTHESIS, CONFIDENCE, INTERVAL AND LEVEL OF SIGNIFICANCE, ERRORS OF TYPES II AND I

CHAPTER HIGHLIGHTS :

This chapter discuss about the formulation of statistical hypothesis, level of significance and critical region. It also explain the type of error I and II and two tailed and one tailed test.

CHAPTER OUTLINES :

16.1 INTRODUCTION :

Any statement or assertion about a statistical population or the values of its parameters is called a statistical hypothesis. There are two types of hypothesis - simple and composite. A statistical hypothesis which does not specify the population completely i.e. either the form of probability distribution or some parameters remain unknown is called a composite hypothesis. On the other hand a statistical hypothesis which specifies the population completely i.e. the probability distribution and all parameters are known is called a simple hypothesis. A hypothesis is a supposition made as a basis for reasoning. According to Prof. Morris Hamburg, "A hypothesis in statistics is simply a quantitative statement about a population."

There can be several types of hypothesis. For example, a coin may be tossed 200 times and we may get heads 80 times and tails 120 times. We may now be interested in testing the hypothesis that the coin is unbiased. To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110Ib. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115Ib. Similarly, we may be interested in testing the hypothesis that the variable in the population are uncorrelated. The validity of a hypothesis will be tested by analysing the sample. The procedure which enables us to decide whether a certain hypothesis is true or not is called test of significance or test of hypothesis. A test of hypothesis is a procedure which specified a set of "rules of decision" whether to 'accept' or 'reject' the hypothesis under consideration.

16.2 NULL AND ALTERNATIVE HYPOTHESIS :

A statistical hypothesis which is set up and whose validity is tested for possible rejection on the basis of sample observations is called a null hypothesis. It is denoted by H_0 and tested against alternatives. Tests of hypothesis deal with rejection or acceptance of null hypothesis only. For example, the null hypothesis may be that the population mean is 40. We write

$$H_o(\blacksquare = 40)$$

The null hypothesis is a very useful tool in testing the significance of difference. In its simplest form the hypothesis asserts that there is no real difference in the sample and the population in the particular matter under consideration and that the difference found is accidental and unimportant arising out of fluctuations of sampling. The null hypothesis is akin to the legal principle that a man is innocent until he is proved guilty. It constitutes a challenge; and the function of the experiment is to give the facts a chance to refute this challenge. For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that "extra coaching has not benefited the students." The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicates that the differences are due to chance.

A statistical hypothesis which differs from the null hypothesis is called an alternative hypothesis and is denoted by H_1. The alternative hypothesis is not tested, but its acceptance depends on the rejection of the null hypothesis. Alternative hypothesis contradicts the null hypothesis. The choice of an appropriate critical region depends on the type of alternative hypothesis, viz., whether both sided, one-sided or specified alternative. The null hypothesis is tested against an alternative hypothesis which is the above case, may be either that the population mean is not 40, or that it s greater than 40, or that it is less than 40, i.e., any one of

The sample is then analysed to decide whether to 'reject' or not to reject the null hypothesis H_o. For this purpose, we choose a suitable statistic called "Test statistic" and its sampling distribution, assuming that H_o is really true. The 'observed value' of the statistic in the sample will in general be different from the 'expected value' because of sampling fluctuations. If the difference between them is large, the null hypothesis H_o is rejected, and we doubt the validity of our assumption. If the difference is not large, H_o is not rejected, and the difference may be considered to have arisen solely due to fluctuations of sampling. It is therefore necessary to decide how much of difference is tolerable before we are able to conclude that the null hypothesis is acceptable.

As against the null hypothesis the alternative hypothesis specified those values that the researcher believes to hold true and of course, he hopes that the sample data lead to acceptance of this hypothesis as true. The alternative hypothesis may embrace the whole range of values rather than single point.

16.3 LEVEL OF SIGNIFICANCE AND CRITICAL REGION :

The maximum probability with which a true null hypothesis is rejected is known as level of significance of the test, and is denoted by ▢. In framing decision rules, the level of significance is arbitrarily chosen in advance depending on the consequences of a statistical decision Customarily 5% or 1% level of significance is taken, although other levels such as 2% or ½% is also used. The level of significance ▢ is used to indicate the upper limit of the probability of committing type I error i.e. the size of critical region.

The decision about rejection or otherwise of the null hypothesis is based on probability considerations. Assuming the null hypothesis to be true, we calculate the probability of obtaining a difference equal to or greater than the observed difference. If this probability is found to be small, say less than 0.05, the conclusion is that the observed value of the statistic is rather unusual and has risen because the underlying assumption is not true i.e. null hypothesis. We say that the observed difference is significant at 5% level and hence the ‘null hypothesis is rejected at 5% level of significance. If, however, this probability is not very small, say more than 0.05, the observed difference cannot be considered to be unusual and is attributed to sampling fluctuations only. The difference is, now said to be not significant at 5% level, and we conclude that there is no reason to reject the null hypothesis at 5% level of significance. It has become customary to use 5% and 1% levels of significance, although other levels, such as 2% or 0.5% may also be used.

Suppose the sampling distribution of the statistic is a normal distribution.

Since the area under normal curve outside the ordinates at mean 1.96(s.d.) is only 5% the probability that the observed value of the statistic differs from the expected value of 1.96 times the S.E. or more is 0.05, and the probability of a larger difference will be still smaller. If therefore,

is either greater than 1.96 or less than -1.96 i.e. numerically greater than 1.96, the null hypothesis H_o is rejected at 5% level of significance. The set of values i.e. constitutes what is called the critical region for the test. The set of values of the test statistic which lead to rejection of the null hypothesis is called critical region of the test. The probability with which a true null hypothesis is rejected by the test is often referred to as "Size" of the Critical Region. Geometrically, a sample x_1, x_2...x_n, of size n is looked upon as just a point x, called sample point, with in the region of all possible samples, called the sample space (W). The critical region is then defined as a subset (w) of those sample points which lead to the rejection of the null hypothesis. Similarly since the area outside mean 2.58 (s.d.) is only 1%. H_o is rejected at 1% level of significances, If Z numerically exceeds 2.58 i.e., the critical region at 1% level.

Using the sampling distribution of an appropriate test statistic we are thus able to establish the maximum difference at a specified level between the observed and expected values that is consistent with the null hypothesis H_o. The set of values of the test statistic corresponding to this difference which lead to the acceptance of H_o is called region of acceptances. Conversely the set of values of the test statistic leading to the rejection of H_o is referred to as region of rejection or "Critical Region" of the test. The value of the statistic which lies at the boundary of the regions of acceptance and rejection is called critical value. When the null hypothesis is true, the probability of observed value often test statistic falling in the critical region is often called the "Size of Critical Region".

Size of critical region Level of significance

However, for a continuous population, the critical region is so determined that its size equals the level of significance ().

16.4 STEPS IN TEST OF SIGNIFICANCE :

1. Set up the "Null Hypothesis" H_o and the "Alternative hypothesis" H_1 on the basis of the given problem. The null hypothesis usually specified the values of some parameters involved in the population:

   The type of alternative hypothesis determines whether to use a two-tailed or one-tailed test.

2. State the appropriate "test statistic" T and also its sampling distribution, when the null hypothesis is true. In large sample tests the statistic /S.E. (T), which approximately follows standard normal distribution, is often used. In small sample tests, the population is assumed to be normal and various test statistics are used which follow standard normal, chi-square, t or f distribution exactly.

3. Select the "level of significance" of the test, if it is not specified in the given problem. This represents the maximum probability of committing a type I error, i.e of making a wrong decision by the test procedure when in fact the null hypothesis is true. Usually a 5% or 1% level of significane is used (If nothing is mentioned, use 5% level).

4. Find the "Critical region" of the test at the chosen level of significance. This represents the set of values of the test statistics which lead to rejection of the null hypothesis. The critical region always appears in one or both tails of the distribution, depending on whether the alternative hypothesis is one-sided or both sided. The area in the tails (called size of the critical region) must be equal to the level of significance . For a one-tailed test, appears in one tail and for a two tailed test /2 appears in each tail of the distribution. The critical region is

   when

when

when

Where is the value of T such that the area to its right is .

5. Compute the value of the test statistic T on the basis of sample data the null hypothesis. In large sample tests. If some parameters remain unknown they should be estimated from the sample.

6. If the computed value of test statistic T lies in the critical region, "reject H_o"; otherwise "do not reject H_o". The decision regarding rejection or otherwise of H_o is made after a comparison of the computed value of T with the critical value.

7. Write the conclusion inplain non-technical language. If H_o is rejected the interpretation is : "The data are not consistent with the assumption that the null hypothesis is true and hence H_o is not tenable". If H_o is not rejected, "the data cannot provide any evidence against the null hypothesis and hence H_o may be accepted to the true". The conclusion should preferably be given in the words stated in the problem.

16.5 TYPE I AND TYPE II ERRORS :

The procedure of testing statistical hypothesis does not guarantee that all decisions are perfectly accurate. At times, the tests may lead to erroneous conclusions. This is so because the decision is taken on the basis of sample values, which are themselves fluctuating and depend purely on chance. The errors in statistical decisions are of two types :

16.5.1 Type I Error :

This is the error committed in rejecting a null hypothesis by the test when it is really true. The critical region is so determined that the probability of type I error does not exceed the level of significance of committing a type I error is denoted by a (pronounced as alpha), where

= Probability (Type I error)

= Probability (Rejecting H_o/H□ is true)

= Probability that the test statistic lies in the critical region,

assuming □

The probability of Type I error must not exceed the level of significance (□) of the test.

Probability of the Type I error □ Level of significance

16.5.2 Type II Error :

This is the error committed in accpeting a null hypothesis by the test when it is really false. The probability of type II error depends on the specified value of the alternative hypothesis, and is used in evaluating the efficiency of a test. The probability of committing a Type II error is denoted by □ (pronounced as beta), where

□ = Probability (Type II error)

= Probability (Not rejecting or accepting H_o/H□ is false)

= Probability that the test statistic lies in the region of acceptance

assuming □

The probability of Type I error is necessary for constructing a test of significance. It is in fact the 'size of the critical region". The probability of Type II error is used to measure the "power" of the test in detecting falsity of the null hypothesis.

The distinction between these two types of errors can be made by an example. Assume that the difference between two population means is actually zero. If our test of significance when applied to the sample means lead us to believe that difference in population means is significant, we make a type I error. On the other hand, suppose there is true difference between the two population means. Now If our test of significance leads to the judgement "not significant", we commit a Type II error. we thus find ourselves in the situation which is described by the following table :

	Accept H_o	Reject H▨
H_o is true	Correct Decision	Type I Error
H_o is false	Type II error	Correct decision

while testing hypothesis the aim is to reduce both the types of error, i.e. Type I and Type II. But due to fixed sample size, it is not possible to control both the errors simultaneously. There is a trade-off between these types of errors : the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. In order to get a low ▨, we will have to put up with a high ▨. To deal with this trade-off in business situations, managers decide the appropriate level of significance by examining the costs or penalities attached to both types of errors.

It is more dangerous to accept a false hypothesis (Type II error) than to reject a correct one (Type I error). Hence we keep the probability of committing Type I error at a certain level, called the level of significance. The level of significance (also known as the size of the rejection region or size of the critical region or simply size of the test) is traditionally denoted by the Green letter ▨. In most statistical tests, the level of significance is generally fixed at 5%. This means that the probability of accepting a true hypothesis is 95%.

16.6 TWO-TAILED AND ONE TAILED TESTS :

While testing hypothesis we often talk of two-tailed tests and one-tailed tests. A two-tailed test of hypothesis will reject the null hypothesis, if the sample statistic is significantly higher than or lower than the hypothesized population parameter. Thus in a two-tail test the rejection region is located in both the tails. If we are testing a hypothesis at 5% level of significance, the size of the acceptance region on each side of the mean wold be 0.475 and the size of the rejection region is 0.025. If we consult the table of areas under the normal curve we find that an area of 0.475 corresponds to 1.96 standard errors on each side ▨H, the hypothetical mean, and this equals the size of the acceptance region. If the sample mean falls nto this area, the hypothesis is accepted. If the sample mean falls into the area beyond 1.96 standard error, the hypothesis is rejected, because it falls into the rejection region. The acceptance and rejection regions for testing hypothesis, at the 0.05 level of significance, are given for a two-tail test.

It would be clear from the diagram that in a two-tail test, rejection regions are located in both tails.

Suppose we want to reduce the risk of committing an error of Type I. This is done by reducing the size of the rejection region. For this a hypothesis may be treated at the 0.01 level of significance which means that the probability of rejecting a true hypothesis is 1%.

If we consult the table of areas under the normal curve. We find that an acceptance region of 0.495 is equal to 2.58 standard error from . The acceptance and rejection regions at 0.01 level of significance are given in the diagram on below.

It will be clear from the above figure that we decrease the size of rejection region, we increase the probability of accepting our hypothesis. As =0.01 the probability of rejecting a true hypothesis is 1%.

A two-tailed test is appropriate when the null hypothesis is and the alternative hypothesis is . For example if we are interested in testing the null hypothesis that the average income per household is Rs. 1,000 against the alternative hypothesis that it is not Rs. 1000 the rejection region would lie on both sides because we would reject the null hypothesis if the mean income in the sample is either too far above Rs. 1000 or too far below Rs. 1000. Hence we are using a two-tail test, we can apply one-tailed test also. One tailed test is so called because the rejection will be located in only one tail which may be either left or right depending upon the alternatgive hypothesis formulated. For example, if we are interested in testing a hypothesis that the average income per household is greater than Rs. 1000 against the alternative hypothesis that the income is Rs.

1000, we will place all the alpha risk on the right side of one theoretical sampling distribution and the test will be one sided rights test. On the other hand, if we are testing the hypothesis that the average income per household is Rs. 1000 against alternative that the income is less than Rs. 1000 or less, the alpha risk is on the left side of theoretical sampling distributon and the test willbe one sided left tail test.

To sum up if we want to test the hypothesis that the population has a specified mean, say, then the null hypothesis would be :

and the alternative hypothesis could be :

1.

2.

3.

The alternative hypothesis in (1) is known as a two-tailed alternative and the alternatives in (2) and (3) are known as right-tailed and left-tailed alternatives. Accordingly the corresponding tests of significance are called two-tailed, right-tailed and left-tailed tests respectively.

16.7 POWER OF A HYPOTHESIS TEST :

The null hypothesis is accepted when the observed value of test statistic lies the critical region as determined by the test procedure. Suppose that the true value of is not , but another value i.e. a specified alternatives hypothesis is true. Type II error is committed if H_o is not rejected i.e. the test statistic lies outside the critical region. Hence the probability of Type II error is a function of , because is assumed to be true.

If denotes the probability of Type II error, when is true, the complementary probability is called power of the test against the specified alternative .

Power = 1 - Probability of Type II error

= Probability of rejecting H_o when H_1 is true

Obviously, we would like a test to be as 'powerful' as possible for all critical regions of the same size. Treated as a function of , the expression is called Power Function of the test for against . The curve obtained by plotting against all possible values of , is known as power curve.

Example :

The fraction of defective items in a large lot is P. To test the null hypothesis H_o : P = 0.2, one consider the number of defectives in a sample of 8 items and accept the hypothesis. If , and rejects he hypothesis otherwise. What is the probability of type I error of this test ? What is the probability of type II error corresponding to P = 0.1 ?

Solution :

We are going to test whether the fraction (P) of defective in the lot is 0.2 or not.

Null hypothesis H_o (P=0.2). Alternatives hypothesis H_1(P 0.2). The test procedure is as follows :

1. Take a random sample of 8 items from the lot.
2. Count the number of defectives in the sample.
3. Accept H_o, if the number of defectives found in the sample is 6 or less , and conclude that fraction of defectives in the lot may be 0.2

Reject H_o, if the number of defectives actually obtained in the sample is 7 or 8.

Conclusion : The fraction of defectives in the lot is not 0.2, i.e. H_o is not tenable. It may be seen that the number of defectives(f) in the sample is a random variable, which follows Binomial distribution with parameter and P. The probability of r defectives is .

Probability of Type I error

= Probability of rejecting H_o, when H_o is true.

= Probability of 7 or 8 defectives, when P = 0.2

=

=0.0000,8448

Probability of Type II error

= Probability of accepting H_o, when a specified H_1 is true.

= Probability of 6 or less defectives, When P = 0.1

= 1-Probability of 7 or 8 defectives, When P = 0.1

= 0.9990,9927

16.8 LET US SUM UP :

Statistics which helps us in arriving at the criterion for such decisions is known as testing of hypothesis. A statistical hypothesis is some assumption or statement, whcih may or may not be true, about a population or equivalently about the probability distributon characterising the given population, which we want to test on the basis of the evidence from the random sample. The random selection of the samples from the given population makes the tests of significance valid for us. The null hypothesis consists of only a single parameter value and is usually simple while alternative hypothesis is usually compsite. One tailed or two-tailed test is to be applied entirely on the nature of the alternative hypothesis.

16.9 LESSON END EXERCISE :

Q1. What is meant by a test of a null hypothesis ? What are the type I and type II errors ?

Q2. Explainthe terms :

1. Test of significance

2. Critical region

3. Level of significance

Q3. A professor is concerned with the effectiveness of a given teaching technique. What null hypothesis is the testing if he is committing a type I error when he erroneously concludes that the particular teaching technique is effective ?

Q4. Explain the concept of power of a hypothesis test with the help of example?

Q5. Explain the steps in test of significance ?

****

M.A. Economics	Semester - IInd	Lesson - 17
Course No. 203		Unit - IV

HYPOTHESIS TESTING BASED ON CHI-SQUARE TEST AND GOODNESS OF FIT

CHAPTER HIGHLIGHTS

This chapter explains the characteristics, uses and applications of chi-square test.

CHAPTER OUTLINES :

17.1 INTRODUCTION :

The X^2 test (chi-square test) is one of the simplest and most widely used non-parametric tests in statistical work. The symbol X^2 test was the Greek letter chi. The X^2 test was first used by Karl Pearson in the year 1990. The quantity X^2 describes the magnitude of the discrepancy between theory and observation.

.....(1)

Where O refers to the observed frequencies and E refers to the expected frequencies. To determine the value of X^2, steps are required :

1. Calculate the expected frequencies. In general the expected frequency for any cell can be calculated from the following equation :

   E = Expected frequency

   RT = The row total for the row containing the cell

   CT = The column total for the column containing the cell

   N = The total number of observations

2. Take the difference between observed and expected frequencies and obtain te square of these differences i.e., obtain the values of $(O-E)^2$.

3. Divide the values of $(O-E)^2$ obtained the step (2) by the respective expected frequency and obtan the total . This gives the value of X^2 which can range from zero to infinity. If X^2 is zero it means that the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater shall be the value of X^2.

   The calculated value of X^2 is compared with the table of X^2 for given degrees of freedom at a certain specified level of significance. If at the stated level (generally 5% level is selected), the calculated value of X^2 is more than the table value of X^2, the difference between theory and observation is considered to be significant i.e., it could not have arisen due to fluctuations of simple sampling. If, on th other hand, the calculated value of X^2 is less than the table value, the difference between theory and observation is not considered as significant i.e., it is regarded as due to fluctuations of simple sampling and hence ignored. The computed value of X^2 is a random variable which takes on different values from sample to sample.

17.2 DEGRES OF FREEDOM :

Degrees of freedom we mean the number of classes to which the values can be assigned arbitrarily or at will without violating the restrictions or limitations placed. For example, if there are 10 classes and we want our frequencies to be distributed in such a manner that the number of cases, the mean and the standard deviation agree with the original distribution, we have three constraint and so three degrees of freedom are lost. Hence in this case the degrees of freedom will be 10-3-7. Thus the number of degrees of freedom is obtained by subtracting from the number of classes the number of degrees of freedom lost is fitting. Symbolically, the degrees of freedom are denoted by the symbol v (pronounced nu) or by d.f. and are obtained as follows :

$$V = n-k$$

Where k refers to the number of independent of independent constraints.

In a contingency table the degrees of freedom are calculated in a slightly different manner. The marginal total or frequencies place the limit on our choice of selecting cell frequencies. The cell frequencies of all columns but one (c-1) and of all rows but one (r-1) can be assigned arbitrarily and so the number of degrees of freedom for all the cell frequencies = (c-1) (r-1) where c refers to column and r refers to rows. Thus in a 2x2 table the degree of freedom = (2-1) (2-1)=1.

17.3 CHARACTERISTICS OF CHI-SQUARE TEST:

1. Mean = n, where n is the number of degrees of freedom (d.f.) of chi-square distribution.
2. The chi-square curve is positively skew and starting from O extends to infinity on the right.
3. If x and y are independent chi-square variates with d.f. n_1 and n_2 respectively, then their sum (x+y) also follows chi-square distribution with d.f. $(n_1 + n_2)$.
4. When the d.f.n. is large approximately follows the standard normal distribution.

17.4 SIMPLIFIED FORMULA FOR (2X2) TABLE :

Suppose, the contingency table has only 2 rows and 2 columns with the four cell frequencies a, b, c, d, as shown in table.

			Total
	a	b	R_1
	c	d	R_2
Total	C_1	C_2	N

For simplicity we write R_1, R_2 to denote the row totals and C_1, C_2 to denote the column totals, i.e.

$$R_1 = a + b,\ R_2 = c + b,\ C_1 = a + c,\ \ C_2 = b + d$$

$$N = a + b + c + d = R_1 + R_2 = C_1 + \ C_2$$

In this case, formula (1) reduces to

............(2)

with degrees of freedom = (2-1) (2-1) = 1

17.5 YATES CORRECTION :

For a (2x2) table, there is only one degree of freedom i.e. only one of the four cell frequencies can be arbitarily given, if the row and colum totals should remain fixed. It is therefore, necessary to make a correction to formula (2), so that its approximation to the continuous chi-square distribution can be improved.

If ad>bc, reduce a and d by and increase b and c by ;

If ad < bc, increase a and d by and reduce b and c by

This is known as yates correction for continuity.

.... (3)

It should be noted that Yate's correction can be applied only for 1 d.f.

Example :

In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers; while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys ? Use X^2 test Value of X^2 for 1 d.f. at 5% level is 3.84

Solution :

The data are shown in the following (2x2) table :

Intelligence of Son	Skill of father		Total
	Skilled	Unskilled	
Intelligenct	40	35	75
Unintelligent	40	85	125
Total	80	120	200

Null hypothesis is that the two attributes “skill of father” and “Intelligence of Son” are independent, and the alternative hypothesis is that they are not independent. On the hypothesis of independence, the test statistic [(formula (3)] follows X^2 distribution with 1 d.f.

Since the observed value of the statistic 8.02 is greater than the tabulated value 3.84 (given), it is significant. We therefore, reject the null hypothesis at 5% level of significant and conclude that the attributes are not independent, i.e, the data support the alternative hypothesis that ‘skilled fathers have intelligent boys’.

17.6 USES OF X^2 TEST :

The chi-square distribution is used in both large sample and small sample tests. It is mainly used in :

1. Test for goodness of fit
2. Test for independence of attributes

3. Test of Homogeneity

17.6.1 Test for goodness of fit :

The test, devised by Karl Pearson, is used to decide whether the observations are in good agreement with a hypothetical distribution, i.e. whether the sample may be supposed to have arisen from a specified population. The observed frequencies (fo) of different classes are compared with the expected frequencies (fe) by the test statistic :-

This is called "Pearsonian chi-square" or "goodness-of-fit chi-square". When the null hypothesis, viz.

H_o (Date are in agreement with the hypothetical population) is true, the statistic approximately follows chi-square distribution with (k-1) degrees of freedom, where k is the number of classes. If the observed value of the statistic exceeds the tabulated value of X^2 at a given level, the null hypothesis is rejected.

Example :

A dice was thrown 60 times with the following results :

Face	1	2	3	4	5	6	Total
Frequency	6	10	8	13	11	12	60

Are the data consistent with the hypothesis that the dice is unbiased ?

(Given $X^2_{0.01}$= 15.09 for 5 degree of freedom)

Solution :

Null hypothesis is that the dice is unbiased. Then the probability of each face is and the expected frequency is for each.

Observed frequency (fo)	6	10	8	13	11	12
Expected frequency (fe)	10	10	10	10	10	10
$(fo\text{-}fe)^2$	16	0	4	9	1	4

There are 6 classes, degree of freedom (d.f.) = (6-1) = 5

Since the observed value of X^2 (viz. 3.4) is less than the tabulated value 15.09 at 1% for 5 degree of freedom, we cannot reject the null hypothesis at 1% level of significance. The conclusion is that the data are in agreement with the hypothesis of an unbiased dice.

17.6.2 Test of independence :

With the help of X^2 test we can find out whether two or more attributes are associated or not. Suppose we have N observations classified according to some attributes. We may ask whether the attributes are related or independent. Thus, we can find out whether qumme is effective in controlling fever or not, whether there is any association between marriage and failure, or eye colour of husband and wife. In order to test whether or not the attributes are associated we take the null hypothesis that there is no association in the attributes under study or in other words, the two attributes are independent. If the calculated value of X^2 is less than the table value at a certain level of significance (generally 5% level), we say that the results of the experiment provide no evidence for doubting the hypothesis or, in other words the hypothesis that the attributes are not associated holds good. Other hand, if the calculated value of X^2 is greater than the table value at a certain level of significance, we say that the results of the experiment do not support the hypothesis or in other words, the attributes are associated. It should be noted that X^2 is not a measure of the degree or form of relationship, it only tells us whether two principles of classification are or are not significantly related, without reference to any assumptions concerning the form of relationship.

17.6.3 Test of Homogeneity :

The X^2 test of homogeneity is an extension of the chi-square test of independence. Tests of homogeneity are designed to determine whether two or more independent random samples are drawn from the same population or from different populations. Instead of one sample as we use with independence problem we shall now have 2 or more samples. For example, we may be interested in finding out whether or not university students of various levels i.e. undergradute, postgraduate, Ph.D, feel the same in regard to the amount of work required by their professors, i.e., to much work, right amount of work required by their professors, i.e., to much work, right amount of work or too little work. We shall take the hypothesis that the three samples come from the same population; that is, the three classifications are homogeneous in so far as the opinion of three different groups of students about the amount of work required by their professors is concerned. This also means there exists no difference in opinion among the three classes of people on the issue.

The same testing statistic used for tests of independence is used for tests of homogeneity. These two types of tests are however, different in a number of ways. First they are associated with different kinds of problems. Tests of independence are concerned with the problem of whether one attribute is indpendent of another, while tests of homogeneity are concerned with whether different samples come from the same population. Secondly, the former involvesa single sample taken from one population but the latter involves two or more independent samples one from each of the possible populations.

17.7 CHI-SQUARE TEST FOR SPECIFIED VALUE OF POPULATION VARIANCE :

When we want to test if a random sample x_1, x_2,....,x_n has been drawn from a normal population with mean and a specified variance , the statistic

Where S the standard deviation of the sample follows chi-square distribution with (n-1) degrees of freedom. By comparing the calculated value of X^2 with the tabulated value (n-1) d.f. at certain level of significance we may accept or reject the null hypothesis. It should be noted that this test can be applied only if the population is normal.

Example :

A random sample of size 25 from a population gives the sample standard deviation to be 8.5. Test the hypothesis that the population standard deviation is 10.

Solution :

We set the null hypothesis H_o that the population standard deviation is 10 i.e.,

We are given

$$n = 25, S = 8.5$$

We know that

$$V = n - 1$$
$$= (25 - 1)$$
$$= 24$$

For

$$V = 24, X^2_{0.05} = 36.415$$

Since, the calculated value of X^2 is less than the table value, the hypothesis holds true, i.e., the population standard deviation may be 10.

17.8 LIMITATIONS OF CHI-SQUARE TEST :

1. Frequencies of non-occurence should not be ommitted for bonomial or multinomial events.

2. When data from questionaires and similar devices are analysed, the reader should be careful that he does not set up the tables incorrectly.

3. The formula presented for X^2 statistics is in terms of frequencies. Hence the attempt should not be made to compute on the basis of proportions or other derived measures.

17.9 LET US SUM UP:

The square of a standard normal variable is called a chi-square variate with 1 degree of freedom. If any cell frequency in 2x2 table is less than 5, then for the application of X^2 test it has to be pooled with the prceding or succeeding frequency so that total is greater than 5. This results in the loss of 1 degree of freedom. On the other hand any cell frequency in 2x2 table is less than 5, we apply the yates correction for continuity. Chi-square test of goodness of fit and is used to test if the deviation between observation and theory may be attributed to chance or if it is really due to the adequacy of the theory to fit the observed data. Chi-square test can be used if the N, the total frequency, should be reasonably large and the sample observations should be independent.

17.10LESS END EXERCISE :

1. Describe the important characteristics of chi-square test ?

Q2. In his experiments on pea-breeding. Mendel obtained the following frequencies of seeds : Round and yelllow - 315; wrinkled and yellow 101; Round and Green - 108; wrinkled and green - 32; Total - 556. Theory predicts that the frequencies should be in the proportions 9:3:3:1. Exame the correspondence between theory and observations. (Given that 5% value of X^2 for 3 d.f. is 7.815)

Ans 0.51

Q3. Define pearsonian X^2 and discuss its uses in testing of hypothesis ?

Q4. 200 digits from 0 to 9 are taken at random from a page of a certain random from number table. The frequency distribution of the digits is given :

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	13	21	16	25	22	20	21	25

Can this table be regarded as random ? (Given $X^2_{0.059} = 16.92$)

Ans 6.3

Q5. A random sample of 500 students were classified according to economic condition of their family and also according to merit, as shown below :

Merit	Economic Condition			Total
	Rich	Middle Class	Poor	
Meritorious	42	137	61	240
Not-Meritorious	58	113	89	260
Total	100	250	150	500

Test whether the two attributes merit and economic condition are associated or not, (Given $X^2_{0.05} = 5.99$ and $X^2_{0.01} = 9.21$ for 2 d.f.)

Ans. 9.30

****

M.A. Economics	Semester - IInd	Lesson - 18
Course No. 203		Unit - IV

HYPOTHESIS TESTING BASED ON T-TEST

CHAPTER HIGHLIGHTS :

This chapter explain the characteristics and application of t-test.

CHAPTER OUTLINES :

18.1 INTRODUCTION :

Theoretical work on t-distribution was done by W.S. Gosset (1876-1937) in the early 1900. The t-distribution is commonly called student's t-distribution or simply students distribution. The t-distribution is used when sample size is 30 or less and the population standard deviation is unknown.

The "t-statistic" is defined as :

Where

The t-distribution has been derived mathematically under the assumption of a normally distributed population.

It has the following form :

where

C = a constant required to make the area under the curve equal to unity

V = n-1, the number of degrees of freedom

18.2 CHARACTERISTICS :

1. Mean = 0,
2. The t-curve is symmetrical about 0, extending from to . It has zero skewness and positive kurtosis i.e.,
3. When the d.f.n. is large, the t-distribution can be approximated by the standard normal distribution.

18.3 APPLICATION OF THE T-DISTRIBUTION :

The following are some of the examples to illustrate the way in which the 'students' distribution is generally used to test the significance of the various results obtained from small samples.

18.3.1 To test the significance of the mean of a random sample :

In determining whether the mean of a sample drawn from a normal population deviates significantly from a stated value when variance of the population is unknown we calculate the statistic

where

= the mean of the sample

= the actual or hypothetical mean of the population

n = the sample size

s = the standard deviation of the sample

or

Where d = deviation from the assumed mean. If the calculated value of |t| exceeds $t_{0.05}$' we say that the difference between and is significant at 5% level, if it exceeds $t_{0.01}$, the difference is said to be significant at 1% level. If $|t|<t_{0.05}$' we conclude that the difference between and is not significant and hence the sample mght have been drawn from a population with mean = .

Example :

The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. A random sample of 6 such bulbs gave the following values.

Life in months 24 26 30 20 20 18

Can you regard the producer's claim to be valid at 1% of significance ?

Solution :

Let us take the hypothesis that there is no significant difference in the mean life of bulbs in the sample and that of the population.

x	=x	x^2
24	+1	1
+26	+3	9
30	+7	49
20	-3	9
20	-3	9
18	-5	25

= 4.517

v = n-1

= 6-1

= 5

For v = 5, $t_{0.01}$ = 4.032

The calculated value of t is less than the table value. The hypothesis is accepted. Hence the producer's claim is not valid at 1% of significance.

18.3.2 Testing Difference Between Means of Two Samples :

Independent Samples

Given two independent random samples of size n_1 and n_2 with and and standard deviations S_1 and S_2. We may be interested in testing the hypothess that the sample come from the same normal population.

where

= mean of the first sample

= mean of the second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

S = combined standard deviation

The value of S is calculated by the following formula

When the actual means are in fraction the deviations should be taken from assumed means. In such a case the combined standard deviation is obtained by applying the following formula

Where

A_1 = assumed mean of the first sample

A_2 = assumed mean of the second sample

= actual mean of the first sample

= actual mean of the second sample

$(n_1 + n_2 - 2)$ = The degree of freedom

When we are given the number of observations and standard deviation of the two samples, the pooled estimate of standard deviation can be obtained as follows:

If the calculated value of t be > $t_{0.05}$ ($t_{0.01}$), the difference between the sample means is said to be significant at 5% (1%) level of significance otherwise the data are said to be consistent with the hypothesis.

Example :

Two types of drugs were used on 5 and 7 patients for reducing their weight. Drug A was imported and Drug B indigenous. The decrease in the weight after using the drugs for six months was as follows :

Drug A :	10	12	13	11	14		
Drug B :	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs ? If not, which drug should you buy.

Solution :

Let us take the hypothesis that there is not significant difference in the efficacy of the two drugs.

x_1			x_2		
10	-2	4	8	-3	9
12	0	0	9	-2	4
13	+1	1	12	+1	1
11	-1	1	14	+3	9
14	+2	4	15	+4	16
			10	-1	1
			9	-2	4

$= 0.735$

$V = n_1 + n_2 - 2$

$= 5 + 7 - 2$

$= 10$

For $V = 10$, $t_{0.05} =$ 2.228

The calculated value of t is less than the table value, the hypothesis is accepted. Hence there is no significance in the efficacy of two drugs. Since drug B efficacyof two drugs. Since drug B is indigenous and there is no difference in the efficacy of imported and indigenous drug, we should buy indigenous drug i.e. B.

18.3.3 Testing Difference Between Mean of two samples :

Dependent Samples

The two samples may consist of pairs of observations made on the same object, individual or more generally on the same selected population elements. When samples are dependent they comprise the same number of elementary units.

or

where = the mean of the differences

S = the standard deviation of the differences

The value of S is calculated as follows :

or

n-1 = degree of freedom

Example :

To verify whether a course in accounting improved performance, a similar test was given to 12 participant both before and after the course. The original marks recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course the marks were in the same order. 53, 38, 69, 57, 46, 39, 73 48, 73, 74, 60 and 78 was the course useful ?

Solution :

Let us take the hypothesis that there is no difference in the marks obtained before and after the course i.e., the course has not been useful.

Participants	Before (Ist Test)	After (2nd Test)	(2nd-Ist Test) d	d^2
A	44	53	+9	81
B	40	38	-2	4
C	61	69	+8	64
D	52	57	+5	25
E	32	46	+14	196
F	44	39	-5	25
G	70	73	+3	9
H	41	48	+7	49
I	67	73	+6	36
J	72	74	+2	4
K	53	60	+7	49
L	72	78	+6	36

= 5.03

= 3.443

V = n-1 = 12-1 = 11

For V = 11, $t_{0.05}$ = 2.201

The calculaed volume of t is greater than the table value. The hypothesis is rejected. Hence the course has been useful.

18.3.4 Testing the Significance of an observed correlation coefficient :

Given a random sample from a bivariate normal population. If we are to test the hypothesis that the correlation coefficient of the population is zero i.e., the variables in the population are uncorrelated.

Hence

t is based on (n - 2) degree of freedom. If the calculated value of the t exceeds $t_{0.05}$ for (n-2), d.f., we say that the value of r is significant at 5% level. If $t < t_{0.05}$ the data are consistent with the hypothesis of an uncorrelated population.

Example :

A study of the hights of 18 pairs of husbands and their wives in a factory shows that the coefficient correlation is 0.52. Apply t-test to find whether correlation is significant.

Solution :

Let us take the hypothesis that there is no significant difference in the sample correlation and correlation in the population.

$$r = 0.52,\ n = 18$$

$$V = (n - 2) = (18 - 2) = 16$$

For

$$V = 16,\ t_{0.05} = 2.12$$

The calculated value of t is greater than the table value. The hypothesis is rejected. The given value of r is significant.

18.4 LIMITATIONS :

1. T-test is not justified for small samples.
2. If it is not a random sample, then the assumption that the observations are statistically independent is not justified and the conclusion based on the t-test may not be correct.

18.5 LET US SUM UP :

If sample size n is small, then the distributions of these standardised statistics are far from normality and consequently normal test cannot be applied.

To deal with small samples, new techniques and tests of significance known as exact sample tests have been developed. It is very difficult to make a clearcut distinction between small samples and large samples. T-test for significance of single mean, population variance being unknown. It is an observed sample correlation coefficient.

18.6 LESSON END EXERCISE :

Q1. The life time of electric bulbs for a random sample of 10 from a large consignement gave the following data :

Item	1	2	3	4	5	6	7	8	9	10
Life in '000 hours	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can me accept the hypothesis that the average life time of bulbs in 4000 hours.

Ans. 2.148

Q2. A drug is given to 10 patient and the increments in their blood pressure were recorded to be 3, 6, -2, 4 -3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has not effect on change of blood pressue ? (5% value of t for d.f. = 2.26)

Ans. 2

Q3. Is a correlation coefficient of 0.5 significant if obtained from a random sample of 11 pairs of value from a normal population ? (Use t-test)

Ans. 1.732

Q4. The hights of six randomly chosen solidiers are in inches : 76, 70, 68, 69, 69 and 68. Those of 6 randomly chosen sailors are 68, 64, 65, 69, 72, 64. Discuss the light that that these data throw on the suggestions that soldiers are on the average, taller than sailors. Use t-test.

Ans. 1.66

Q5. State the assumptions underlying student's t-test and limitations also ?

****

M.A. Economics **Semester - IInd** **Lesson - 19**
Course No. 203 **Unit - IV**

HYPOTHESIS BASED ON Z-TEST

CHAPTER HIGHLIGHTS :

This chapter highlights two main applications of z-test.

CHAPTER OUTLINES :

19.1 Introduction

19.2 Test for specified correlation coefficient

19.3 To test the significance of the difference between two independent correlation coefficients.

19.4 Let us sum up

19.5 Lesson end exercise

19.1 INTRODUCTION :

Prof. Fisher has given a method of testing the significance of the correlation coefficient in small samples. According to this method the coefficient of correlation is transformed into Z and hence the name Z-transformation. The statistic Z given by Prof. Fisher is used to test :

1. Test for a specified value of and find confidence limits for P.
2. To test the significance of the difference between two independent correlation coefficients.

   For testing whether r differs significantly from zero, the t-test is preferable.

19.2 TEST FOR A SPECIFIED CORRELATION COEFFICIENT:

A random sample of n pairs of observations from a bivariate normal population shows a correlation coefficient r. It is required to test the hypothesis that the population correlation coefficient P has a specified value. H_o ($p = p_o$)

Using fisher's z-transforation and when H_o is true, the test statistic

follows standard normal distribution approximately, where is the value of

from when $p = p_0$

Fisher's z-transformation uses logarithm to the base e, for practical calculations we may change the base to 10

, since $\log_e 10 = 2.30$ approx.

approximately

Example :

A random sample of 28 pairs of observations shows a correlation coefficient of 0.74. Is it reasonable to believe that the sample comes from a bivariate normal population with correlation coefficient 0.6 ?

Solution :

Null hypothesis H_o ($p = 0.6$)

Alternative hypothesis H_1 (p 0.6)

Using fisher's Z - transformation and H_o is true (here r =

0.74, $p_o = 0.6$, n = 28)

$= 1.15\ (\log 174 - \log 26)$

$= 1.15\ (2.2405 - 1.4150)$

$1.15(0.8255) = 0.949$

$= 1.15 \log 4 = 1.15\ (0.6021) = 0.692$

The test statistic is , which follows standard normal distribution under H_o.

$U = (0.949 - 0.692)$

$= 1.285$

Since here $|U|<1.96$, the observed value of the test statistic is not siginificant at 5% level. We therefore, cannot reject H_o and conclude that the sample might come from a bivariate normal population with correlation coefficient 0.6.

19.3 TO TEST THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO INDEPENDENT CORRELATION COEFFICIENT :

To test the significance of two correlation coefficients derived from two separate sample we have to compare the difference of the two corresponding values of Z with the standard error of that difference - remembering that the standard error of the difference of two statistical quantities in the square root of the sum of their variances.

where

or

or

If the absolute value of this statistic is greater than 1.96, the difference will be significant at 5% level.

Example :

The following data give sample sizes and correlation coefficients. Test the significance of the difference between two values using fisher's z-transformation.

Sample Size	Value of r
5	0.870
12	0.560

Solution :

Let us take the hypothesis that the samples are drawn from the same population.

where

$r_1 = 0.87$

= 1.1513 log 14.385

1.1513 x 1.1579 = 1.333

= 1.1513 log 3.545

1.1513 x 0.5496 = 0.633

$Z_1 - Z_2$ 1.333 - 0.633 = 0.7

= 0.782

Since the difference is less than 2.58 S.E. (1% level) the experiment provides no evidence against the hypothesis that the samples are drawn from the sample population.

19.4 LET US SUM UP :

In sampling from a bivariate normal population in which the variables are correlated, i.e., p ▨ 0, the distribution of statistic t-test is not normal even for large samples. In such cases, z-transformation will used. In z-test correlation coefficient in the population has a specified value. We may conclude that the samples have been drawn from the sample population.

19.5 LESSON END EXERCISE :

Q1. A correlation coefficient of 0.2 is discovered in a sample of 28 pairs. Use z-test to find out if this is significantly different from zero ?

Ans. 1.015

Q2. The correlation coefficients 0.45 and 0.70 were obtained from two independent random samples of 19 and 28 pairs of observations respectively drawn from bivariate normal populations. Do these results support the hypothesis that the correlation coefficients in the two populuations are equal ?

Ans. -1.19

Q3. Explain fisher's z-transformation of correlation coefficient and indicate its uses in test of significance ?

Q4. From a sample of 19 pairs of observations the correatlion is 0.5 and the corresponding population value is 0.3. Is the difference significant ?

Ans. 0.96

Q5. The correlation coefficients 0.45 and 0.70 were obtained from two independent random samples of 19 and 28 pairs of observations respectively drawn from bivariate normal populations. Do these results support the hypothesis that the correlation coefficients in the two populations are equal ?

Ans. -1.19

***

M.A. Economics Semester - IInd Lesson - 20

Course No. 203 Unit - IV

HYPOTHESIS TESTING BASED ON F-TEST

CHAPTER HIGHLIGHTS :

This chapter contain the characteristics, assumptions and applications of f-test or the variance ratio test.

CHAPTER OUTLINES :

20.1 Introduction

20.2 Characteristics of the F-test

20.3 Assumptions of the F-test

20.4 Applications of the F-test

20.5 Let us sum up

20.6 Lesson end exercise

20.1 INTRODUCTION :

A random variable is said to follow f distribution with degree of freedom (n_1, n_2). If is p.d.f. is of the form

Where k is the constant. The distribution was discovered by G.W. snedecor and named F is honour of the distinguished mathematical statistician Sir R.A. Fisher. The percentage points of F-distribution with d.f. (n_1, n_2) are denoted by

$F_p(n_1, n_2)$ of briefly F_p if the d.f. are understood. The lower percentage points are given by

i.e. the lower percentage point is the reciprocal of the upper percentage point with the order of d.f. reversed. The object of the F-test is to find out whether the two independent estimates of population variance differe significantly, or whether the two samples may be regarded as drawn from the normal populations having the same variance.

, where

It should be noted is always the larger estimage of variance, i.e.

or

$v_1 = n_1 - 1$ and $v_2 = n_2 - 1$

V_1 = Degrees of freedom for sample having large variance.

V_2 = Degrees of freedom for sample having smaller variance.

The calculated value of F is compared with the table value for v_1 and v_2 at 5% or 1% level of significance. If calculated value of F is greater than the table value then the F ratio is considered significant and the null hypothesis is rejected. On the other hand, if the calculated value of F is less than the table value the null hypothesis is accepted and it is inferred that both the samples have come from the population having some variance.

Since F test is based on the ratio of two variances, it is also known as the variance ratio test.

20.2 CHARACTERISTICS OF THE F-TEST :

1.

   provided they exist and are positive

2. The F-curve is postively skew and starting from 0 extends to infinity.

20.3 ASSUMPTIONS OF THE F-TEST :

The F-test is based on the following assumptions :

1. **Normality** : The values in each group are normally distributed.
2. **Independence of error :** It states that the error should be independent for each value.
3. **Homogeneity :** The variance within each group should be equal for all group

. This assumption is needed in order to combine or pool the variances with in the groups into a single 'within groups' source of variation.

20.4 APPLICATIONS OF F-TEST :

Example :

Two random sources were drawn from two normal populations and their values are :

A:	66	67	75	76	82	84	88	90	92	
B:	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance (F=3.36) at 5% level for $V_1 = 10$ and $V_2 = 8$.

Solution :

Let us take the hypothess that the two populations have the same variance.

A			B		
66	-14	196	64	-19	361
67	-13	169	66	-17	289
75	-5	25	74	-9	81
76	-4	16	78	-5	25
82	+2	4	82	-1	1
84	+4	16	85	+2	4
88	+8	64	87	+4	16
90	+10	100	92	+9	81
92	+12	144	93	+10	100
			95	+12	144
			97	+14	196

= 1.415

For $V_1 = 10$ and $V_2 = 8$, $F_{0.05} = 3.36$

The calcuated value of F is less than the table value. The hypothesis is accepted. Hence it may be calculated that the two populations have the same variance.

Example :

The standard deviations calculated from two random samples of size 9 and 13 are 2.1 and 1.8 respectively. May the samples be regarded as drawn from normal populations with the same s.d.? (The 5% value of F from tables with d.f. 8 and 12 is $F_{0.05} = 2.85$)

Solution :

Here $n_1 = 9$, $n_2 = 13$, $S_1 = 2.1$ and $S_2 = 1.8$. The unbiased estimates of the variances and in the two populations are respectively.

= 4.96

$$= 3.51$$

We have to test against . The value of the statistic is

$$= 1.41$$

Degrees of freedom are (8, 12). Since the observed value of F, viz. 1.41, is less than the 5% tabulated value viz 2.85 corresponding to d.f. (8, 12), we cannot reject the null hypothesis at 5% level of significance. The conclusion is that the population s.d.s. may be equal.

20.5 LET US SUM UP:

F - statistics is the ratio of two independent chi-square variates divided by their respective degrees of freedom. The sampling distribution of F-statistic does not involve any population parameter and depends only on the degrees of freedom n_1 and n_2. A statistic F following snedecor's F-distribution with (n_1, n_2) d.f. will be denoted by $F \sim F(n_1, n_2)$. F-test for testing the significance of an observed sample multiple correlation. It is an observed sample correlation ratio. Its testing the linearity of regression.

20.6 LESSON END EXERCISE :

Q1. Two random samples are drawn from two populations and the following results were obtained.

Sample I	16	17	18	19	20	21	22	24	26	27		
Sample II	19	22	23	25	26	28	29	30	31	32	35	36

Find the variances of the two samples and test whether the two populations have the same variance.

Ans. 1.94

Q2. In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at 5% level.

You are given that a 5% level, critical value of F for $V_1 = 7$ and $V_2=9$ degrees of freedom 3.29 and for $V_1 = 8$ and $V_2 = 10$ degrees of freedom, its value is 3.07 ?

Ans. 1.195

Q3. The standard deviations calculated from two independent random samples from two independent random samples of sized 9 and 10 were found to be 2.4 and 1.8 respectively. Can the samples be regarded to have been drawn from equally variable normal population? (Given that 5% value of F from the tables is 3.23 for degrees of freedom 8 and 9 respectively)

Ans. 1.8

Q4. Two samples are drawn from two normal populations from the following data test whether the two samples have the same variance at 5% level.

Sample I	60	65	71	74	76	82	85	87		
Sample II	61	66	67	85	78	63	85	86	88	91

Ans. 1.468

Q5. The following data present the yields in quintals of common ten sub-divisions of equal area of two agricultural plots :

Plot 1	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot 2	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5

Test whether two samples taken from two random population have the same variance (5% point of F for $V_1 = 9$ and $V_2 = 9$ is 3.18 ?

Ans. 2.63

SUGGESTED FOR FURTHER READING :

- Gupta S.C. (2011), Fundamental of Statistics, Himalaya Publisher House, New Delhi.
- Gupta, S.P. (2011), Statistical Methods, Sultan Chand and Sons Educational Publisher, New Delhi

- Das, N.G. (2012), Statistical Methods Tata McGraw Hill Education Private Limited, New Delhi.
- Rice (2007) Mathematical Statistics and Data Analayiss, Brooks/Cole Cengage Learning India
- Do uglas et.al (2008)., Statistical Techniques in Business and Economic Tata McGraw-Hill Publishing Company Limited, New Delhi.

****

M.A. Economics **Semester - IInd** **Lesson - 21**
Course No. 203 **Unit - IV**

ANALYSIS OF VARIANCE (ANOVA)

STRUCTURE :

21.1 ANALYSIS OF VARIANCE (ANOVA)

The term ANOVA "Analysis of Variance" was introduced by Prof. R.A. Fisher in 1920. It is defined as separation of variation ascribable to one group of causes from the variance ascribable to other group. It consist in the estimation of the amount of variation due to each of the independent factor (i.e. causes) separately and then comparing these estimates due to assignable factors or causes with the estimate due to chance factors or causes.

There are two types of causes.

(i) Assignable causes

(ii) Chance causes / Experimental error

The variation due to assignable causes can be defined and measured whereas the variation due to chance causes is beyond the control of human and cannot be traced separately.

21.1.1 Assumptions of ANOVA Test

(i) The observations are independent

(ii) Parent population from which observations are taken is normal.

(iii) Various treatment and environmental effect are additive in nature.

Application or utility of ANOVA :-

The analysis of variance is a powerful statistical tool for testing of significance. The test of significance based on t-test is an adequate procedure only for testing the significance of difference between two sample means in a situation when we have three or more sample to be considered at a time and alternative procedure is needed for testing the hypothesis that all the samples are drawn from same population i.e., they have the same mean.

Example :- Five fertilizers are applied to four plots of wheat and yield of wheat on each of the plot is given. We may be interested in finding out whether the effect of these fertilizers on the yield is significantly different.

The answer to this problem is provided by technique known as Analysis of variance or ANOVA the basic purpose of ANOVA is to test the homogeneity of several means.

One Way ANOVA :-

Let us suppose that N observations yif where

Assumptions :-

(i) The observations are independent

(ii) Parent population from which observations are taken is normal.

(iii) Various treatment and environmental effect are additive in nature.

Independently identically distributed random variable.

Statistical Analysis of One-way ANOVA :-

Null hypothesis is

e r

under the one way ANOA, we observed only one factor.

The technique involves the following steps

(i) Obtain the mean of each sample

where K are the samples

(ii) Workout the mean of sample means as follows :

(iii) Taking the deviation of the sample mean, from the mean of sample means and calculate the square of such deviations which may be multiplied by the no of items in the corresponding sample and thus obtain their total this is called as sum of square for variance between the samples.

(iv) Divide the SS b/w sample by the degree of freedom to obtain variance or mean square between samples

Where, K-1 is the degree of freedom b/w samples.

(v) Obtain the deviation of the value of the sample item from all the sample from the corresponding mean of samples and calculate the square of such deviation and their obtain their total. This total is known as sum of squares for variance within sample.

(vi) Divide the SS within by degree of freedom to obtain the variance or mean square within sample

Where n-k is the degree of freedom within sample n is the total no. of items in all sample i.e. $n_1+n_2+....+n_k$
K is the no. of samples

(vii) The sum of square of deviation from total variance can also be work out by adding the sum of deviations when the deviations from an individual item in all been taken from the mean of sample means.

It should be equal to the total of the result of step 3 and step 5 explained above.
i.e. SS for total variance = SS b/w sample + SS within sample
The degree of freedom for within and between must add up to degree of freedom for total variance
$(n-1) = (k-1) + (n-k)$

(viii) Finally F-Ratio works as under :-

This F-Ratio works as the test statistic follows F-distribution with [(k-1),(n-k)] degree of freedom

Thumb Rule :- If F Cal > F tab, we reject Ho i.e., it is significant we reject the null hypothesis that all the pop mean or the effect of all the treatments are the same at the given level of significance when the computed value of F-Ratio is greater than that of the critical value which is greater than that of the tabulated value.

ANOVA table

Following are the per hectare production of 3 varieties of wheat each grow in 4 plots and state if the variety differs significantly

Plot	Variety of wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution

We wish to test Ho (null hypothesis)

Ho : He wheat varieties are same or effect of different varieties on wheat production are the same

Vs. H1 : All varieties are not the same

First, we calculate the mean of each sample

$$\overline{x}_1=\frac{6+7+3+8}{4}=6$$

$$\overline{x}_2=\frac{5+5+3+7}{4}=5$$

$$\overline{x}_3=\frac{5+4+3+4}{4}=4$$

Now, $\overline{\overline{x}}=\frac{\overline{x}_1+\overline{x}_2+\overline{x}_3}{k}$

$$=\frac{6+5+4}{3}$$

$$=\frac{15}{3}$$

$$=3$$

We will now find the sum of square between & within samples

ANOVA table

Therefore, this analysis supports the null hypothesis of no difference in means. Thus, we conclude that the difference in wheat output due to varieties is significant and is just a matter of chance.

Two way classification or ANOVA or one observation per cell :-

Suppose n observations are classified into k - categories say A1,A2,….,Ak according to some criteria and into h categories say B1,B2,…..Bh according to some criteria, b having kh combinations (Ai,Bj) where i = 1,n…,k, j=1,n…,h often called cells. This scheme of classification according to two factor or criteria is called two way classification and its analysis is called two-way ANOVA.
In two way classification, the value of the response variable are affected by two factors.

Example :- The Agricultural output may be classified on the basis of different varieties of seed and also on the basis of different varieties of fertilizer used.

Assumptions :-

(i) The observations are independent
(ii) Parent population from which observations is taken is normal.
(iii) Various treatment and environmental effects are additive in nature.

Statistical Analysis of Two-way ANOVA :-

The various steps involve in two-way ANOVA are as follows :
(i) We wish to test that
I HOC : the variety of the first factor has the same effect

Vs Hic : The variety of the first factor are significantly different.

II Hor : The variety of second factor has the same effect

HIR : The variety of second factor are significantly different

(iii) Take the total observations in all the sample and called it "T"

(iv) Workout the correction factor as under

$$\text{C.F.}=\frac{\text{T}^2}{n}=\frac{\text{T}^2}{kh}$$

(v) Find out the square of all the items value one by one and takes its total, subtract the correction factor from this total to obtain the sum of square or deviation from total variance

(vi) Take the total of different column and obtain sum of each column total and divide such square of each column by the no. of items in the concerning column and taking the total of the result thus obtained Finally, subtract the correction factor to obtain the sum of square of deviation for variance between columns or sum of square between coloumns.

(vii) Take the total of different row and obtain sum of each row total and divide such square value of each row by the number of items in the concerning rows and taking the total of the result thus obtained. Finally subtract the correction factor to obtain the sum of square of deviation for variance between rows or sum of squares between rows.

(viii) Sum of squares of deviation from residual or error variance can be obtained as

SSE = TSS-SSC-SSR

(ix) Degree of freedom is obtained as

SSE = TSS-SSC-SSR

(c-1)(r-1) = (cr-1)-(c-1)-(r-1)

Where c = no of coloumns

R = no of rows

(x) Formation of ANOVA Table

For hypothesis I, the test statistic is

$$Fc=\frac{S^2c}{S^2e} \sim F\,[(c-1),(c-1)(r-1)]$$

For hypothesis II, the test statistic is

Thus, the mean square error provide the basis for the F-Ratio concerning variation between coloumn treatment and between row treatment.

The mean square error is always due to the fluctuation of sampling and hence serve as the basis for the significance test. But, the F-Ratio are compared with their corresponding tabulated value for given degree of freedom at a specified level of significance as usual and if its found that the calculated F-Ratio concerning variation between coloumn is greater than its tabulated value, then the difference among the column mean is considered significant. Similarly F-ratio concerning variation between rows can be interpreted

Example : Set up ANOVA table for the following two way designing result

Per acre production of wheat

Variety of fertilizers	Variety of seeds		
	A	B	C
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Also, state whether the variety difference are significant at 5% level of significance

Sol:- Here 1. H_{oc} : variety of seeds are same

Vs H_{ic} : variety of seeds are not same

2. H_{or} : variety of fertilizers are same

H_{ir} : variety of fertilizers are not same

Variety of fertilizers	Variety of seeds			Ri	Ri2	
	A	B	C			
W	6	5	5	16	256	
X	7	5	4	16	256	
Y	3	3	3	9	81	
Z	8	2	4	19	361	
Cj	24	20	16	60		
Cj^2	576	400	256	1232	954	Total

Correction factor

Source of variation	Degree of freedom	Sum of Squares	Mean Sum of Squares	F-Ratio
SSC	8	2	8/2 = 4	4/1 = 4
SSR	18	3	18/3 = 6	6/1 = 6
SSE	6	6	6/6 = 1	
TSS	32	11		

Tabulated F

F tab (2,6) (0.05) = 5.14

F tab (3,6) (0.05) = 4.76

So, we accept the Null hypothesis HOC and conclude that the variety of seeds are same

And Fcal (R) > F tab (R)
i.e. 6 > 4.76
So, we reject the null hypothesis HOR and conclude that the variety of fertilizers are not same.

REFERENCES :

- Gupta S.C. and Kapoor, V.K. (2000), Fundamental of Mathematical Statistics, Sultan Chand & Sons, Publisher House, New Delhi.
- Gupta, S.P. (2011), Statistical Methods, Sultan Chand and Sons Educational Publisher, New Delhi
- Gupta S.C. and Kapoor, V.K. (2000), Fundamental of Applied Statistics, Sultan Chand & Sons, Publisher House, New Delhi.
- Kothari, C.R.(2000), Research Methology Methods & Techniques.