Contents

Lesson - 1

QUANTITATIVE DECISION MAKING AN OVERVIEW

Objectives:

After studying this lesson you should be able to

- Importance of quantitative techniques
- Difference between statistics and operations research
- Advantages of quantitaive methods
- Importance of computers in Business Applications

Structure:

1.1 Introduction:

Quantitative studies of management have generally been considered to have originated during the World War II period, when operations research teams were formed to deal with strategic and tactical problems faced by the military. These teams, which often consisted of people with diverse specialties (e.g., Engineers, Mathematicians and Behavioral Scientists) were joined together to solve common problems through the utilization of scientific methods (Anderson, Sweeney and Williams, 1994). After the war, many of these team members continued their research on

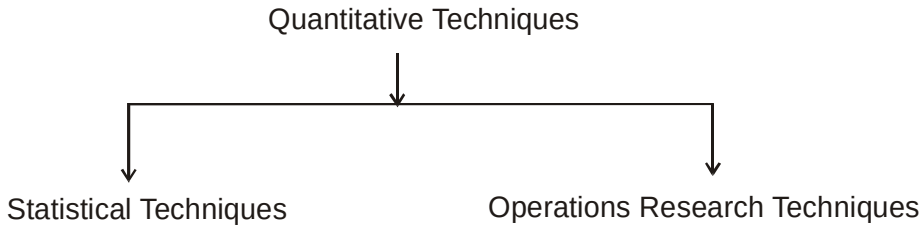
quantitative approaches to decision-making, leading to the growth and use of management science in nonmilitary applications such as manufacturing, health care, engineering projects, transportation and traffic studies, communication, business and educational administration.

Concurrent with these methodological developments, system analysis was developed. It represents one approach to solving problems within the framework of systematic output followed by feedback. Thus, information systems facilitated the advance of computer technology. Numerous software programs were written to develop variants of the post-World War II methodological approaches, allowing for solutions to more complex and larger problems than those requiring only intuitive, simpler solutions (Robbins & DeCenzo, 1995).

The contingency viewpoint or the situational approach is the most recent school of thought about management. In essence, managers who follow this approach can use any other viewpoint, depending on the current circumstances. Managers or administrators should consider the three key contingency variables of environment, technology and people, before making a decision (Hellriegel & Slocum, 1992).

1.2 Meanning of Quantitative Techniques:

Quantitative Techniques refer to the group of statistical and operations research techniques as given below:



These techniques require basic knowledge of certain topics in mathematics. The quantitative approach in decision making is that if the factors that influence the decision can be identified and quantified then it becames easier to resolve the complexity of the tools of quantitative analysis. Quantitative analysis is now extended to several areas of business operations and represents probably the most effective a proaches to handling of some types of decision problems. One can use terms management science, operations research and quantitative analysis interchangeably.

Definition of Terms:

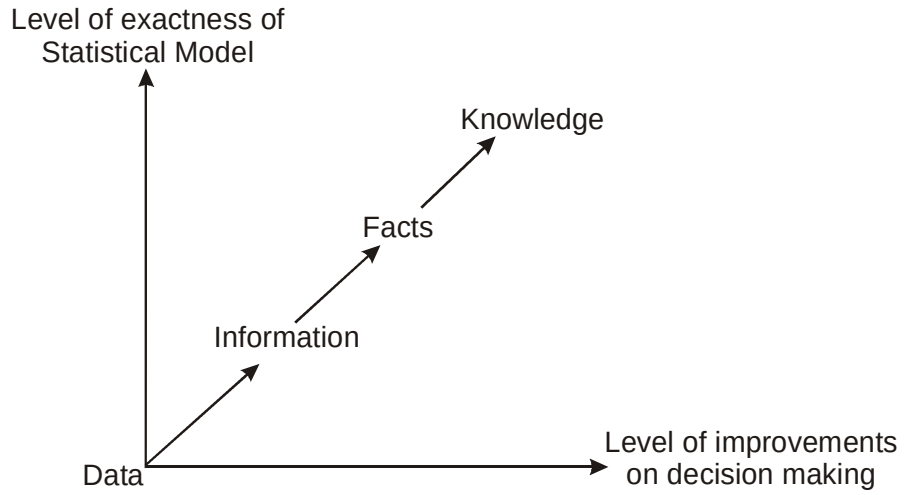
Quatitative Techniques: Quantitative techniques are managerial aids concerned with the development of any appropriate utilization of rational approaches of intervention in human affairs.

Quantitative Techniques: Quantitative techniques are mathematical and statistical models describing a diverse array of variable's relationship and are designed to assist administrators with management problem solving and decission making.

Statistical Model:

Fact becomes knowledge, when it is used in the successful completion of a decision process. Once you have a massive amount of facts integrated as knowledge, then your mind will be superhuman in the same sense that manking with writing is superhuman compared to mankind

before writing. The following figure illustrates the statistical thinking process based on data in constructing statistical models for decision making under uncertainties.



The above figure depicts the fact that as the exactness of a statistical model increases, the level of improvements in decision - making increases. That's why we need statistical data analysis. Statistical data analysis arose from the need to place knowledge on a systematic evidence base. This required a study of the laws of probability, the development of measures of data properties and relationships and so on.

1.3 Statistics and Operations Research:

Statistics can be described in two senses namely.

Plural Sense (Statistical Data): One may count the number of heads and determine the size of the population. In statistics, however, this term implies an aggregate or collection of measurements on a given variable or variables.

Singular sense (Statistical Methods): Statistics are numerical facts capable of analysis and interpretation and the science of statistics is study of principles and methods in collection, presentation, analysis and interpretation of numerical data.

As an illustration, let us suppose that we are interested in known the litaracy level of the people living in Andhra Pradesh. For this we may adopt the following procedures:

(a) Data collection: The following data is received for the given purpose

  (i) Population of the Andhra Pradesh

  (ii) Number of individuals who have completed their education

(b) Organise (or condense) the data: After collecting data procedure is completed the data should be organised in different educational level. This will reduce the bulk of the data.

(c) Presentation: The third step is to present by means of various types of graphs. Data presented in an orderly manner facilitates statistical analysis.

(d) **Analysis:** The fourth step is find average literacy in different groups this information will help to get an understanding of the phenonimean.

(e) Interpretation: This is the last step. One can draw the conclusion based in the analysis which will aid indecision making - a policy decision for improvement ofthe existing situations.

Characteristics of data:

The data must possess the following charateristics:

(i) They must be aggreate of facts. For example single number cannot be used to study.

(ii) They should be effected to a marked extent by multiplicity of causes for example in beauty context the observations recorded are effected by a number of factors.

(iii) They must be enumerated or estimated according to reasonable standard of accuracy. For example blood test is estimated by certain tests on small samples drawn from a human body.

(iv) They must have been collected in systematic manner for a pre determined purpose. Lack or msuse of data about the problem situation leads wrong results.

(v) They must be placed in relation to each other. The data should be comparable otherwise it can not used for relation.

(vi) They must be numerically expressed the data should be in numerically collected otherwise it is not possible to analyse.

Types of Statistical Data:

Primary and Secondary Data:

Primary data is data that you collect yourself using such methods as:

Direct Observation: Lets you focus on details of importance to you; lets you see a system in real rather than theoretical use.

Surveys: Written surveys let you collect considerable quantities of detailed data. You have to either trust the honesty of the people surveyed or build in self - verifying questions.

Interviews: Slow, Expensive and they take people away from their regular jobs, but they allow in - depth questioning and follow-up questions. They also show non-verbal communication such as face-pulling, fidgeting, shrugging, hand gestures, sarcastic expressions that add further meaning to spoken words. e.g., "I think it's a GREAT sytem" could mean vastly different things depending on whether the person was sneering at the time! A problem with interviews is that people might say what they think the interviewer wants to hear; they might avoid being honestly critical in case their jobs or reputation might suffer.

Logs: (e.g., fault logs, error logs, complaint logs, transaction logs). Good empirical, objective data sources (usually, if they are used well). Can yield lots of valuable data about system performance over time under different conditions.

Primary data can be relied on because you know where it came from and what was done to it. It's like cooking something yourself. You konw what went into it.

Secondary Data:

Secondary data analysis can be literally defined as second-hand analysis. It is the analysis of data or information that was either gathered by someone else (e.g. researchers, institutions, other NGOs, etc.) or for some other purpose than the one currently being considered, or often a combination of the two.

If secondary research and data analysis is undertaken with care and diligence it can provide a cost - effective way of gaining a broader understanding of specific phenomena and/or conducting preliminary needs assessments.

Secondary data are also helpful in designing subsequent primary research and as well, can provide a baseline with which to compare your primary data collection results. Therefore, it is always wise to begin any research activity with a review of the secondary data.

Operations Research:

The term operations research (O.R.) was coined during World War II, when the British military management called upon a group of scientists together to apply a scientific approach in the study of military operations to win the battle. The main objective was to allocate scarce resources in an effective manner to various military operations and to the activities within each oepration. The effectiveness of operations research is in military spread interest in it to other government departments and industry.

Due to the availability of faster and flexible computing facilities and the number of qualified O.R. professionals, it is now widely used in military, business, industry, transportation, public health, crime investigation etc.

OR can be regarded as use of mathematical and quantitative techniques to substantiate the decisions being taken. It also takes tools from subjects like mathematics, statistics, engineering, economics, psychology etc., and uses them to know the consequences of possible alternative actions. Making decisions or taking actions is central to all operations.

1.4 Classification of Statistical Methods:

Statistical Methods broadly fall into three categories as follows:

Descriptive Statistics:

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measurs. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Descriptive statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive simple number used to summarize how well a batter is performing in baseball, the number of times at bat (reported to three significant digits). A batter who is hitting 333 is getting a hit one time in every three at bats. One batting 250 is hitting one time in four. the single number

describes a large number of discrete events. Or, consider the scourge of many students the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

Every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether the batter is hitting home runs or singles. It doesn't tell whether she's been in a slump or on a streak. the GPA doesn't tell you whether the student was in difficult courses or easy ones, or whether they were courses in their major field or in other disciplines. Even given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

Inferential Statistics:

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgements of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Perhaps one of the simplest inferential test is used when you want to compare the average performance of two groups on a single measure to see if there is a difference. You might want to know whether eighth - grade boys and girls differ in math test scores or whether a program group differs on the outcome measure from a control group. Whenever you wish to compare the average performance between two groups you should consider the t-test for differences between groups.

Most of the major inferential statistics come from a general family of statistical models known as the General Linear Model. This includes the t-test. Analysis of Variance (ANOVA). Analysis of Covariance (ANCOVA), regression analysis and many of the multivariate methods like factor analysis, multidimensional scaling, cluster anmalysis, discriminant function analysis and so on. Given the importance of the General Linear Model, it's a good idea for any serious reserarcher to become familiar with its workings.

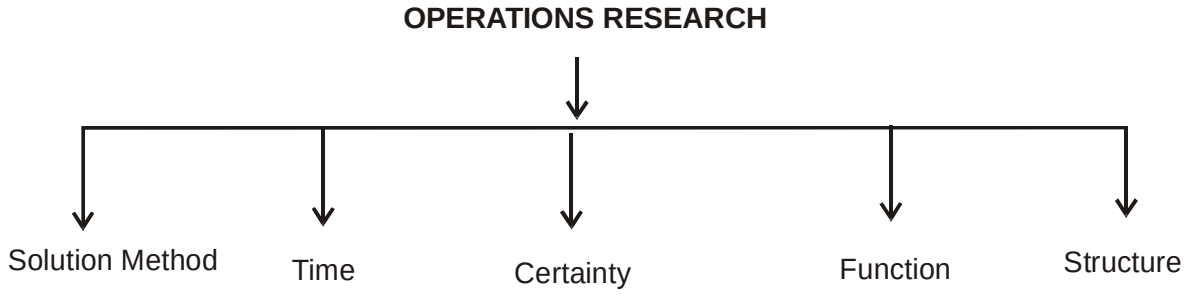
Statistical Decision Theory:

Statistics is a integrated with decision making in areas such as management, public policy, engineering and clinical medicine. Decision theory states the case and in a self - contained, comprehensive way shows how the approach is operational and relevant for real - world decision making under uncertainty.

Starting with an extensive account of the foundations of decision theory, uses concepts of subjective probability and utility.

1.5 Models of operations research:

The models of operations research are classified follows:



These categories are brifely explained.

Solution Method: These solution methods are classified accordeing to the methodology, of which are described as follows:

(i) **Analytical Models:** These models have a specific mathematical structure and thus can be solved by known analytical or mathematical techniques.

(ii) **Simulation Models:** A simulation model is essentially computer assisted experimentation on a mathematical structure of a real time structure in order to study the system under a variety of assumptions.

(iii) **Heuristic Models:** Heuristic models do not claim to find the best solution to the problem.

Time Constraints Methods: These time constraints methods are classified as for their presentation which are described as follows:

(i) **Static Models:** These models do not consider the impact of changes that takes place during the planning horizon.

(ii) **Dynamic Models:** In these models, time is considered as one of the important variables and admit the impact of changes generated by time.

Certainty Models: These models are characterized according to the certainty factor that how much form the model is with respect to fixed variables, values etc...

(i) **Deterministic Models:** Such models assume conditions of complete certainty and perfect knowledge. For examples linear programming, transportation and assignment models are deterministic type of models.

(ii) **Probabilistic (or) Stochastic Models:** These type of models usually handle such situations in which the consequences or pay off of managerial actions cannot be predicted with certainty.

Based on function models: These models of operations research can also be categarised as for the utilities which are discussed as follows:

(i) **Descriptive Models:** A descriptive model simply describe some aspects of a situation based on observations, survey questionnaire results, or other available data. The result of an openion poll represents a descriptive model.

(ii) Predictive Models: Such models can answer "what if" type of questions i.e., they can make predictions regarding certain events.

(iii) Prescriptive (normative) Models: Finally, when a predictive model has been repeatedly successful, it can be used to prescribe a source of action. For example, linear programming is a prescriptive (or) normative model because it prescribes what the managers ought to do.

Based on structures: These model are classified on the basis of their structures. They are classified as follows:

(i) Physical Models: These models are the model which are scaled in size (enlarging or reducing the size) by its original size to analyse and produce the best results.

(ii) Symbolic Models: The symbolic or mathematical models is one which employs a set of mathematical symbols to present the decision variables of the system.

1.6 Various Statistical Methods:

Measures of Central Tendency: The most commonly used measure of central tendency is the mean. To compute the mean, you add up all the numbers and divide by how many numbers and divide by how many numbers there are. It's not the average nor a halfway point, but a kind of center that balances high numbers with low numbers. For this reason, It's most often reported along with some simple measure of dispersion, such as the range, which is expressed as the lowest and highest number.

The median is the number that falls in the middle of a range of numbers. It's not the average it's the halfway point. There are always just as many numbers above the median as below it. In cases where there is an even set of numbers, you average the two middle numbers. the median is best suited for data that are ordinal, or ranked. It is also useful when you have extremely low or high scores.

The mode is the most frequently occuring number in a list of numbers. It's the closest thing to what people mean when they say something is average or typical. The mode doesn't even have to be a number. It will be a category when the data are nominal or qualitative. The mode is useful when you have a highly skewed set of numbers, mostly low or mostly high. You can also have two modes (bimodal distribution) when one group of scores are mostly low and the other group is mostly high, with few in the middle.

Measures of dispersion: In data analysis the purpose of statistically computing a measure of dispersion is to discover the extent to which scores differ, cluster, or spread from around a measure of central tendecy. The most commonly used measure of dispersion is the standard deviation. You first compute the variance, which is calculated by subtracting the mean from each number, squaring it and dividing the grand total (Sum of Squares) by how many numbers there are. The square root of the variance is the standard deviation.

The standard deviation is important for many reasons. One reason is that, once you know the standard deviation, you can standardize by it. Standardization is the process of converting raw scores into what are called standard scores, which allow you to better compare groups of different sizes.Standardization isn't required for data analysis but it becomes useful when you want to compare different sub groups in your sample, or between groups in different studies. A standard score is called a z-score (not to be confused with a z-test), and is calculated by subtracting the

mean from each and every number and dividing by the standard deviation. Once you have converted your data into standard scores, you can then use probability tables that exist for estimating the likelihood that a certain raw score will appear in the population. This is an example of using a descriptive statistic (standard deviation) for inferential purposes.

Correlation: The most commonly used relational statistic is correlation, and it's a measure of the strength of some relationship between two variables, not causality. Interpretation of a correlation coefficient does not even allow the slightest hint of causality. The most a researcher can say is that the variables share something in common; that is, are related in some way. The more two things have something in common, the more strongly they are related. There can also be negative relations, but the important quality of correlation coefficients is not their sign, but their absolute value. A correlation of - 58 is stronger than a correlation of 43, even though with the former, the relationship is negative. The following table lists the interpretations for various correlation coefficients:

0.8 to 1.0	Very Strong
0.6 to 0.8	Strong
0.4 to 0.6	Moderate
0.2 to 0.4	Weak
0.0 to 0.2	Very Weak

If you square the pearson correlation coefficient, you get the coefficient of determination, symbolized by the large letter R. It is the amount of variance accounted for in one variable by the other. Large R can also be computed by using the statistical technique of regression, but in that situation, it's interpreted as the amount of variance.

Regression: Regression is the closest thing to estimating causality in data analysis, and that's because it predicts how much the numbers "fit" a projected straight line. There are also advanced regression techniques for curvilinear estimation. The most common form of regression, however, is linear regression, and the least squares method to find an equation that best fits a line representing what is called the regression of y on x. Instead of findingthe perfect number, however, one is interested in finding the perfect line, such that there is one and only one line (represented by equation) that perfectly represents, or fits the data, regardless of how scattered the data points. The slope of the line (equation) provides information about predicted directionality, and the estimated coefficients (or beta weights) for x and y (independent and dependent variables) indicate the power of the relationshop.

Sampling is the process of selecting a small number of elements from a larger defined target group of elemetns such that the information gathered from the small group will allow judgments to be made about the larger groups.

Simple random sampling is a method of probability sampling in which every unit has an equal non zero chance of being selected.

Probability:

- Simple random sampling
- Systematic random sampling

- Stratified random sampling
- Cluster Sampling

Non Probability:

- Convenience sampling
- Judgement sampling
- Quota sampling
- Snowball sampling

A time series is a set of ordered observations on a quantitaive characteristic of a phenomenon at equally spaced time points. One of the main goals of time series analysis is to forecast future values of the series. A trend is a regular, slowly evolving change in the series level. Changes that can be modeled by low order polynomials. In Time-Series Models we presume to know nothing about the causality that effects the variable we are trying to forecast. Instead, we examine the past behavior of a time series in order to infer something about its future behavior. The method used to produce a forecast may involve the use of a simple deterministic model such as a linear extrapolation or the use of a complex stochastic model for adaptive forecasting.

Index Numbers: Index Number is an indicator of the trend. It is a specialised type of average. It measures the central tendency of the time series or spatial series. In other words index numbers are intended to show variation in magnitude which are not susceptible to direct measurement. Thus index number performs a function to that of an average. As Bowley puts it "index numbers are used to measure the changes in some quantity which we can not observe directly."

General price level is an imaginary concept, and is effected by a number of causes whose absolute effects cannot be measured correctly, so at least to have an idea in such cases relative changes in the price level of different commodities can be measured with the help of index numbers. index numbers is a relative measure of the central tendency of a group of items.

1.7 Advantages of Quantitative Approach to Management:

The role played by statistics with regard to various aspects of quantitative data in various fields. The followingare some advantages of statistics.

1. **Definiteness:** Numerical expressions are usually more convincing than general expression. Thus one of the most important functions of statistics is to present general statements in a precise and definite form. Statistics presents facts in a precise and definite form and helps proper comprehension of what is stated.

2. **Condensation:** Not only statistics presents facts in a definite form but it also helps in condensing mass of data into few significant figures. The main function of statistics is to simplify the huge mass of data as it is not possible to make any analysis from the unorganised data. In a way statistics does a great service by reducing the raw data to totals or averages. Statistical methods present a meaningful overall information from the mass of data which enables us to know the salient features of the data. The raw data may be translated into graphs, charts (or) frequency distribution to make it meaningful.

3. **Comparison:** Statistics facilitates comparison between two or more variables or objects or series relating to different times and places. Unconnected figures have no meaning unless they are being placed in relation to the other figures of the same type.

For making comparison, statistics provides measures such as rate, average, percentages, graphs, diagrams, etc... Thus by furnishing suitable device for comparison of data statistics enables a better appreciation of the significance of a series of figures.

4. **Relationship:** Another function of statistics is to establish relationship between two or more variables relating to different fields.

   For studying relationships statistics provides measures such as correlation regression coefficient of associations etc.

5. **Inference:** Statistics provides methods which are employed for drawing inferences about the nature of parent population. In most of the enquires the characteristics of the population cannot be studied exhaustively. Statistical inference has become the most important branch of statistics. It saves the investigator from the botheration of collecting huge mass of data and results in saving time, money & human energy.

6. **Prediction:** Forecasting of future events has become an importance function of statistics. Plans and policies of organisations are formulated well in advance of the time of their implementation. A knowledge of future trends is very useful in framing suitable policies and plans. Statistical methods are employed for forecasting demand, production, population, exports etc...

   Tools like regression, time series analysis, growth curves etc... are used for making forcasts.

7. **Testing Hypothesis:** Statistical methods are very useful in formulating and testing various types of hypothesis and to develop new theories. In the field of social sciences hypothesis formulation and their testing is very important. In fact in most of the researches in social sciences some hypotheses are formulated and by collecting, analysing and interpreting empirical data and validity of the hypothesis is tested. The hypothesis will be either accepted or rejected. The acceptance of the hypothesis leads to the formulation of new theories.

8. **Formulation of Policies:** Statistics provide the basic material for framing suitable policies. For example it may be necessary to decide how much pulses should be imported in 2009. The decision would depend on the expected pulses production in the country and the likely demand for pulses in 2009. In the absence of the informations regarding the estimated domestic output and demand, the decision on imports cannot be made with reasonable accuracy.

9. **Addition of knowledge:** The science of statistics enlarges human experience and knowledge by making it easier for man to understand, describe and measure the effects of his action on the action of others. Many fields of knowledge would have ever remained unknown to mankind but for the efficient and refined techniques and sound methodology provided by the science of statistics. It has provided such a master key to the mankind that we can use it any where and can study any problem in its right perspective and on the right line.

1.8 Quantitative Techniques in Business and Management:

Some of the areas where statistics can be used are as follows:

1. **Finance, Budgeting and Investments:**

   Credit policy analysis

   Cash flow analysis

   Divided policies

   Investment portfolios

2. **Marketing:**

   Product selection, timing, etc....

   Advertising media, budget allocation.

   Number of salesman required.

   Selection of product mix.

3. **Purchasing, Procurement and Exploration:**

   Optimal buying and reordering

   Replacement policies

4. **Production Management:**

   Location and size of warehouses, factories, retail outlets, etc....

   Distribution policy

   Loading and unloading facilities for trucks etc....

   Production scheduling

   Optimum product mix

   Project scheduling and allocation of resourcess

5. **Personnel Management:**

   Selection of suitable personnel

   Recruitment of employees

   Assignment of jobs

   Skills balancing

6. **Research and Development**

   Project selection

   Control of R & D projects

   Reliability and alternative design

1.9 Uses of Computers:

The Computer is one of the wonders of modern technology.

Banks	:	All financial transactions are done by computer software. They provide security speed and convenience.
Travel	:	One can book air tickets or railway tickets and make hotel reservations online.
Telecommunications	:	Software is widely used here. Also all mobile phones have software embedded in them.
ATM machines	:	The computer software authenticates the user and dispenses cash.
Washing Machines	:	They operate using software
Microwave Oven	:	They are operated by software
Planning and Scheduling	:	Software can be used to store contact information, generating plans, scheduling appointments and deadlines.
Business	:	Corporate computing has revolutionized the business environment with the computer - assistance of many time - consuming complex tasks such as accounting, inventory control, customer databases, shipping control and financial analysis.

Today's desktop computers are versatile and relatively inexpensive. They have replaced the typewriter and in many cases the personal assistant for word processing and organizational services. Networking of all the company's desktop computers allows access to and storage of the full range of company resources and information, reducing paperwork and increasing the productive flow of information. The links included herein relate to business and corporate computing. One can predict future trends of business using artificial intelligence software. Software is used in major stock markets. One can do trading online. There are fully automated factories running on software.

It is used in each and every aspect of human life. They will spearhead the human quest of eradicating social problems like illiteracy and poverty. This revolutionary technology is indeed a boon to the human race. May computers continue to shower their blessings to us.

1.10 Summary:

One can constand about quantitative methods to improve the ability as well as literary skills, so that he can present numerical data and information which requires analysis and interpretation and take quick decision in net only. Business but also any other fields. This lesson teaches to various quantitave approach and tells how to solve it. There are powerful computers, techniques using these techniques you can expleve new and more sophisticated methods of data analysis.

1.11 Exercises:

1. What are the meaning of Quantitative Techiniqes?
2. What is difference between statistics and operations research.
3. Describe the descriptive statistics and inferential statistics.
4. Explain the operations research models?
5. What are statistical methods?
6. Describe the advatages of Quantitative Methods?
7. Explain how Quantitative Techniques used in Business Management?

1.12 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics for Business, Economics and Social Sciences, Mc Graw - Hill ; New York.
2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.
3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof. K. CHANDAN

Lesson - 2

FUNCTIONS AND PROGRESSIONS

Objectives:

After studying this lesson you should be able to understand and appreciate:

- The need to identy or define the relationships tha t exists among business variables
- How to define functional relationships
- The various types of functional relationships
- The use of graph to decipt functional relationships
- The managerial applicability and use of functional relationships in diverse fields.
- The progressions and their applications

Structure:

2.1 Introduction:

Far decision problems which use mathematical tools, the first requirement is to identify or farmally define all significant interactions or relationships among primary factors relevant to the

problem. These relations usually are stated in the form of an equation (or set of equations) or inequetions. Such type of simplified mathematical relationships help the decision - maker in under standing (any) complex management problems. For example, the decision - maker knows that demand of an item is not only related to price of that item but also the price of the substitutes thus if he can define specific mathematical relationship (also called model) that exists, then the demand of the item in the near future can be forecasted. The main objective of this unit is to study mathematical realtionships (or functions) in the context of managereal problems.

2.2 Defenitions:

1. Variable: A variable is something whose magnitude can change, something that can take different values. For example price, profit, sales, cost and so on. Since each variable can assume various values, it is represented by a symbol instead of a specific number. We may represent demand by D, revenue by R, cost by C and so on.

We come across situations in which one variable is "dependent" or related to another for example, demand is related to price of a product. Mathematically, it can be expressed as

$$D = f(p)$$

where f stands for function, D is dependent variable and p is independent variable and is read as "D is a function of p". Variables can be classified in a number of ways. For example, a variable can be discrete (2 houses, 3 machines etc...) or continious (temperature, height etc...)

2. Constant: A quantity that remains fixed in the context of a given problem or situation is called a constant.

3. Parameter: An absolute constant such as $\sqrt{2}$, π, e etc. retains the same value in all problems where as an arbitrary constant or parameter retains the same value throughout any particular problem but may assume different values in different problems, such as wage rates of different category of labourers in an industrial unit.

4. Function: A function is a relationship in which value of a dependent variable are determined from the values of one or more independent variables. For example, demand (D) of a commodity is related to its price (p). It can be mathematically expressed as

$$D = f(p)$$

where f stands for function, D is the dependent variable and p is independent variable and is read as "D is a function of p". In this example demand not only depends on price, but also upon income and price of the substitutes. Hence

$$D = f(p, y, ps)$$

where y denotes income, ps denotes price of the substitute.

A function is sometimes defined as a rule of correspondence between variables. The set of values given to independent variable is called the "domain of the function" while the corresponding set of values of the dependent variable is called the range of the function.

Some of the examples of functions:

i) The distance (d) covered is a function of time (T) and speed (s)

$$d = g(T, s)$$

ii) Sales volume (V) of the commodity is a function of price (p)

$$V = h(p)$$

iii) Total inventory cost (T) is a function of order quantity (Q)

$$T = f(Q)$$

iv) The volume of a sphere (V) is a function of its radius (r)

$$V = f(r)$$

2.3 Types of Functions:

In this section some different types of functions are introduced which are particularly useful in calculus.

1) Linear Function: A linear function is one in which the power of independent variable is 1, the general expression of linear function having only one independent variable is

$$y = f(x) = a x + b$$

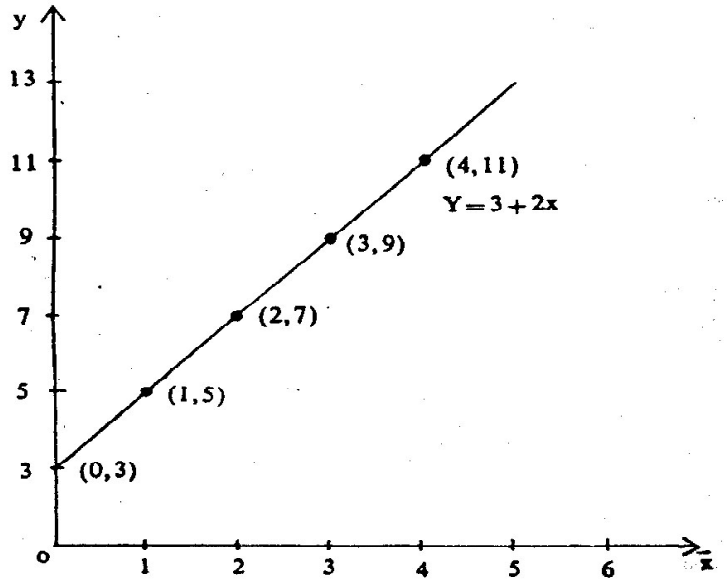
where a and b are given real numbers and x is an independent variable taking all numerical values in an interval. A linear function with one variable can always be graphed in a two dimensional plane. This graph can always be plotted by giving different values of x and calculating corresponding values of y. The graph of such function is always a straight line.

Example: Plot the graph of the function, $y = 3 + 2x$

For plotting the graph of the given function, assigning various values to x and then calculating the corresponding values of y as shown in the table given below:

x	0	1	2	3	4	5
y	3	5	7	9	11	13

The graph of the given function is shown in Figure 1



A function with more than one independent variable is defined, in general form as

$y = f(x_1, x_2, \cdots, x_n) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$ where $a_0, a_1, a_2, \cdots, a_n$ are given real numbers and $x_1, x_2, \cdots, x_n$ are independent variables taking all numerical values in the intervals. Such functions are also called linear multivaribale functions.

2. Polynomial Functions:

A function of the form

$$y = f(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

where a_i's $(i = 0, 1, 2, \cdots, n)$ are real numbers $a_0 \neq 0$ and n is a positive integer is called a polynomial of degree n.

If n = 1, then the polynomial function is of degree 1 and is called a linear function. Then the function is

$$y = a_0x + a_1 \quad , \quad a_0 \neq 0$$

If $n = 1$, then the polynomial function is of degree 2 and is called a quadratic function. Then the function can be of the form

$$y = ax^2 + bx + c$$

where, a, b, c are real numbers.

3. Absolute Value Function:

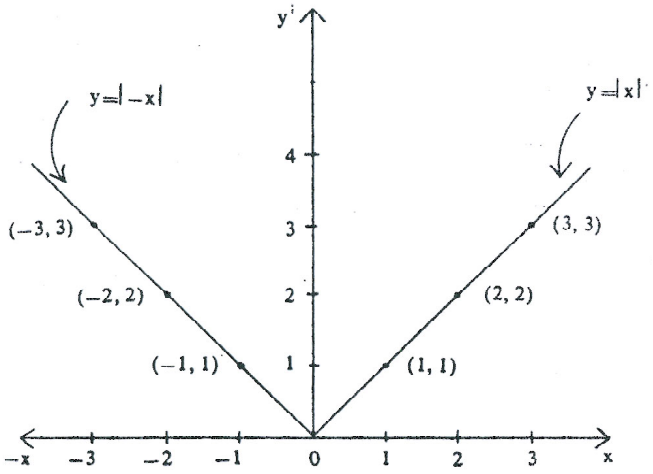
$f(x) = |x|$ then $f(x)$ is known as absolute value function. $|x|$ is known as absolute value of x. By absolute value we mean that whether x is positive or negative, the absolute value will always be positive. Thus $|-5| = 5$, $|6| = 6$

Example: Plot a graph of the function $y = |x|$

Plotting the graph of the function $y = |x|$, assigning various values to x and then calculating the corresponding values of y, is shown in the table below:

x	. . .	- 3	- 2	- 1	0	1	2	3	. . .
y	. . .	3	2	1	0	1	2	3	. . .

The graph of the given function is shown in figure 2.



4. Inverse Function: Take the function $y = f(x)$. Then the value of y, can be uniquely determined for given values of x as per the functional relationship. Sometimes, it is required to consider x as a function of y, so that for given values of y, the value of x can be uniquely determined as per the functional relationship. This is called the inverse function and is also denoted by $x = f^{-1}(y)$.

For example consider the linear function

$$y = ax + b$$

$$\Rightarrow x = \frac{y-b}{a} = \frac{y}{a} - \frac{b}{a} = cy + d$$

where $C = \frac{1}{a}$, $d = -\frac{b}{a}$

This is also linear function and is denoted by $x = f^{-1}(y)$

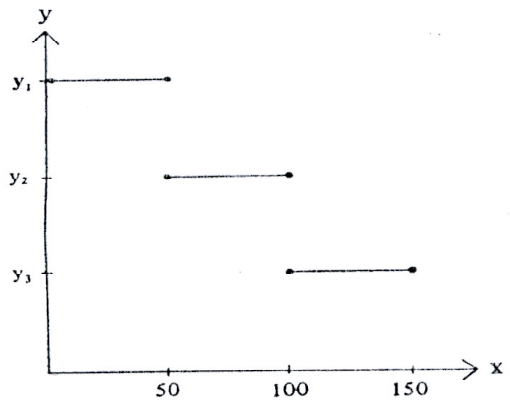
5. Step Function: For a real number x, the step function of x is defined as follows

$[x] = n$ for $n \le x < n+1$

This can also be represented as under

$$y = f(x) = \begin{cases} y_1 & \text{if } a \le x < b \\ y_2 & \text{if } b \le x < c \\ y_3 & \text{if } c \le x < d \end{cases}$$

where $y_3 < y_2 < y_1$ and the graph of the function is given below:



Example:

i) If $x = 2 \cdot 7$ and $2 < 2 \cdot 7 < 3$, then $[x] = 2$

ii) If $x = 0 \cdot 8$ and $0 < 0 \cdot 8 < 1$, then $[x] = 0$

iii) If $x = 5$ and $5 < x < 6$, then $[x] = 5$

iv) If $x = -1 \cdot 2$ and $-2 < -1 \cdot 2 < -1$ then $[x] = -2$

v) If $x = -0 \cdot 3$ and $-1 < -0 \cdot 3 < 0$ then $[x] = -1$

6. Algebraic and Transcendental Functions: Functions can also be classified with respect to the mathematical operations (addition, subtraction, multiplication, division, powers and

roots) involved in the functional relationship between dependent variable and independent variable(s). When only finite number of terms are involved in a functional relationship and variables are effected only by the mathematical operations, then the function is called and "algebraic function", otherwise transcendental function.

The following functions are algebraic functions of x.

i) $y = 2x^3 + 5x^2 - 3x + 9$

ii) $y = \sqrt{x} + \frac{1}{x^2}$

iii) $y = x^3 - \frac{1}{\sqrt{x}} + 2$

7. Exponential Function: If the independent variable in any functional relationship appears as an exponent (or power), then that functional relationship is called exponential function, such as

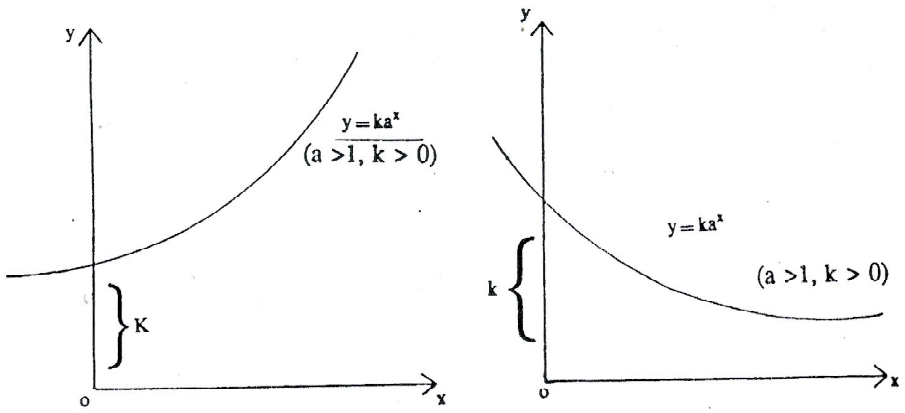
i) $y = a^x$, $a \neq 1$

ii) $y = ka^x$, $a \neq 1$

iii) $y = ka^{bx}$, $a \neq 1$

iv) $y = ke^{bx}$

Such functions are useful for describing sharp increase or decrease in the value of dependent variable. For example, the exponential function $y = ka^x$ curve rises to the right for $a > 1,\ k > 0$ and falls to the left for $a < 1,\ k > 0$ as shown in the figures



8. Logarithmic Function:

A logarithmic function is expressed as

$$y = \log_a x$$

where $a > 0$ and $a \neq 1$ is the base and x is positive real number. It is read as y is the log to the base a of x^1.

This can also be written as $x = a^y$ and can observe that the inverse of an exponential function is a logarithmic function. The two most widely used bases for logarithms are 10 and $e\,(= 2 \cdot 7182)$

i) Common logarithm : It is the logarithm to the base 10 of a number x. It is written as $\log_{10} x \cdot$ If $y = \log_{10} x,$ then $x = 10^y$.

ii) Natural logarithm: It is the logarithm to the base e of a number x. It is written as $\log_e x$ or $\ln x$. when no base is mentioned, it will be understood that the base is e.

Some important properties of the logarithmic function $y = \log_e x$ are as follows:

i) $\log 1 = 0$

ii) $\log e = 1$

iii) $\log_m (x\,y) = \log_m x + \log_m y$

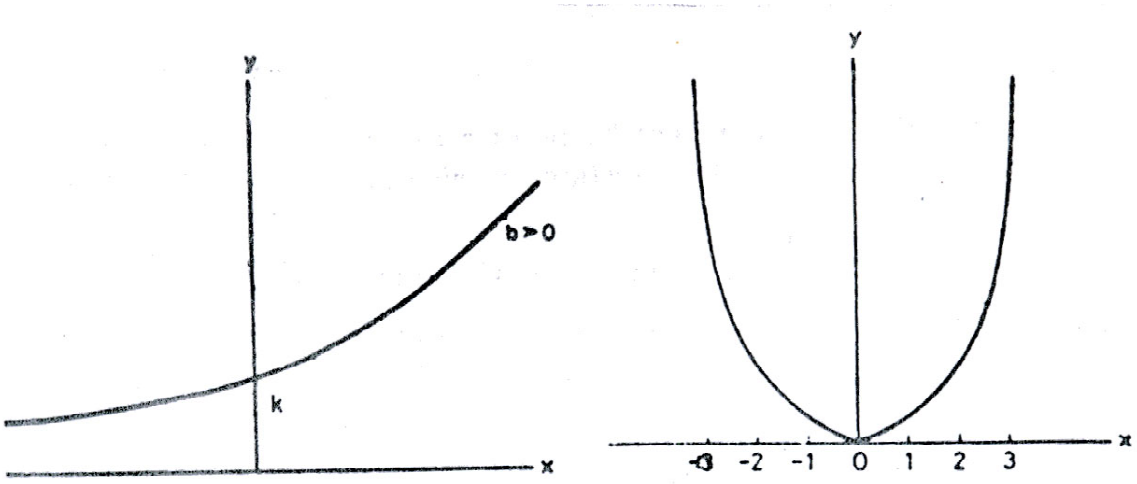
iv) $\log_m \left(\frac{x}{y}\right) = \log_m x - \log_m y$

v) $\log_m x^n = n \log_m x$

vi) $\log_e 10 = \frac{1}{\log_{10} e}$

vii) $\log_e a = (\log_e 10)(\log_{10} a) = \frac{\log_{10} a}{\log_{10} e}$

viii) Logarithm of zero and negative number is not defined.



9. Functions of two or more variables: Functions in which independent variables are more than one are known as multivariate functions. For example

$y = f(x_1, x_2, x_3) = 2x_1 + 3x_1x_2 - 4x_2x_3 + x_3^2$ is a multivariate function since it has three independent variables. It is difficult to plot the above function because it requires four dimensional space. In general, a function of n variables will require $(n+1)$ dimensional space for plotting the curve.

2.4 Solution of Functions:

The values of x at which $f(x)$ becomes zero are known as zeros of the function $f(x)$. The zeros of the function are also known as roots. Thus if a polynomial can be expressed as

$$f(x) = (x-a)(x-b)(x-r)(x-s)\cdots = 0$$

Then $a, b, r, s, \cdots$ will be zeros or roots of the polynomial. In case of $f(x) = ax + b$ the root of the equation is given by $a + bx = 0$ or $x = -\frac{a}{b}$.

Hence the line crosses the x - axis at the point $x = -\frac{a}{b}$.

Example 1:

Given $f(x) = 2x - 5$. Then the zero of $f(x)$ is $x = \frac{5}{2}$

2. Given $y = \frac{2}{3}x + 1$ then root of the equation is $\frac{2}{3}x + 1 = 0 \Rightarrow \frac{2}{3}x = -1 \Rightarrow x = -\frac{3}{2}$

2.5 Managerial Applications:

Linear functions are applied in business and management. We present relationships among profit, total revenue, total cost, variable cost and fixed cost by applying the concept of linear functions and the solution of the equation.

The profit a firm makes on its product is the difference between the amount it receives from sales (its revenue) and its cost

2.5.1 Break - Even Analysis:

In break - even analysis, we first determine the break - even point. The break even point is that level of output at which revenue equals cost. If C is the cost of production and x is the output, then the cost function can be expressed as

$$C = a + bx$$

where 'a' is fixed cost and b is variable cost / unit. At x level of output, the revenue - function will be

$R = px$, where 'p' is sales price / unit

The profit function, therefore is given by

$$P = R - C$$

$$= Px - (a + bx)$$

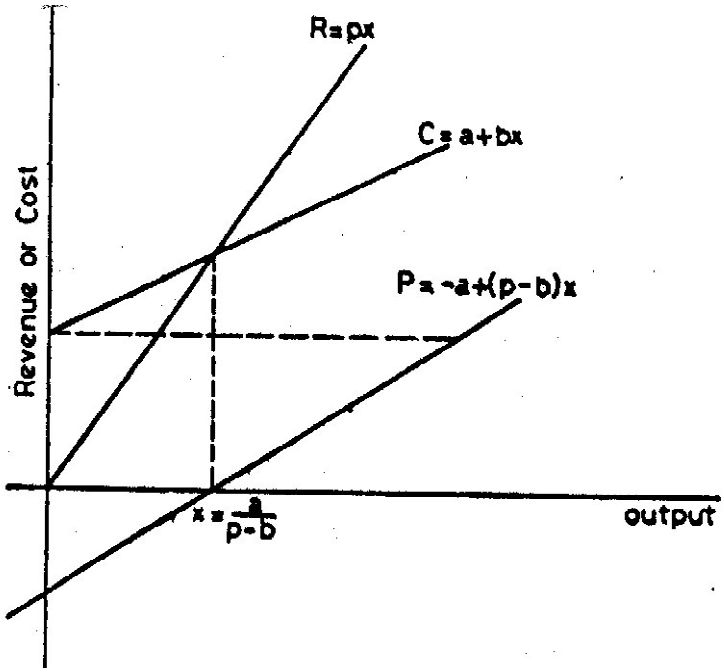
$$= -a + (P - b)x$$

Thus, the profit function is also a linear function, for the break - even point, we should find zeros of the profit function, because at the break - even point, the profit is zero. Hence

$$-a + (P - b)x = 0$$

$$x = \frac{a}{P - b}$$

This is can be graphically presented as in the following figure.



Construction of the cost function for an inventory model. There are two types of cost involved in purchasing raw materials and stocking them. They are ordering cost and inventory holding cost or carrying cost.

If the annual requirement is R units and Q is the quantity which we are likely to order, then the quantity R will be procured in $\frac{R}{Q}$ orders. If the ordering cost is Rs. S per order, then the total ordering cost $= \frac{R}{Q} \times S$

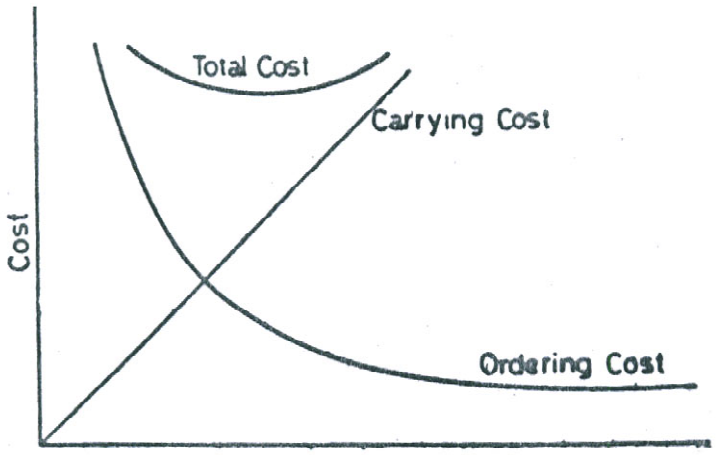
The average inventory at any point of time will be $\frac{Q}{2}$

Hence the inventory carrying cost $= \frac{Q}{2} \times C \times I$

Where I is the carrying cost expressed as a percentage value of average inventory and C is the cost per unit. Hence, the cost - function will be

$$K(Q) = \frac{R}{Q} \times S + \frac{Q}{2} \times C \times I$$

This is represented as in the following figure:



Example 1:

Calculate the break - even point from the following data. Sales price per unit = Rs. 15

Cost Data	(a)	Fixed Cost	Rs.
		Plant Maintenance	15,000
		Salaries	40,000
		Depreciation	1,00,000
		Rent	8,000
		Manufacturing Expenses	12,000
		Advertising	5,000
		Administrative Expenses	20,000
	(b)	Variable cost per unit	
		Labour	3.00
		Materials	5.00
		Sales Commission	2.00
			10 . 00

Solution:

If x is the break - even sales interms of number of units sold, then

Revenue Function = p x = 15 x

Cost Function = a + b x = 2,00,000 + 10 x

Since a = fixed cost = 2,00,000

b = variable cost per unit = 10

At break - even point, Revenue = Cost

Hence $15x = 2,00,000 + 10x$

or $5x = 2,00,000$

Therefore, $x = 40,000$ units

If the firm produces less than 40,000 units, it will be making losses.

2.6 Sequence and Series:

Sequence:

If for every positive integer n, there corresponds a number a_n such that a_n is related to n by some rule, then the terms $a_1, a_2, \cdots, a_n, \cdots$ are said to form a sequence.

A sequence is denoted by bracketing its n^{th} terms i.e., (a_n) or $\{a_n\}$.

Example of a few sequences are:

i) If $a_n = n^2$ then sequence $\{a_n\}$ is 1, 4, 9, 16, $\cdots a_n, \cdots$

ii) If $a_n = 1/n$ then sequence $\{a_n\}$ is $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots, \frac{1}{n}, \cdots$

iii) If $a_n = \frac{n^2}{n+1}$ then sequence $\{a_n\}$ is $\frac{1}{2}, \frac{4}{3}, \frac{9}{4}, \cdots, \frac{n^2}{n+1}, \cdots$

The concept of sequence is very useful in fincance some of the major areas where it plays a vital role are: 'instalment buying', 'simple and compound interest problems', 'annuities and their present values', mortgage paymetns and so on.

Series:

A series is formed by connecting the terms of a sequences with plus or minus sign. Thus if a_n is the n^{th} term of a series sequence, then

$$a_1 + a_2 + \cdots + a_n$$

is the given series of n terms.

2.7 Arithmetic Progression (A P):

A progression is a sequence whose successive terms indicate the growth or progess of some characteristics. An arithmetic progression is a sequence whose term increases or decreases by a constant number called 'common difference' of an A.P. and is denoted by d. in other words, each term of the arithmetic progression after the first is obtained by adding a constant d to the preceding term. The standard form of an A.P. is written as

$a, a+d, a+2d, a+3d, \cdots$

where 'a' is called the first term. Thus the corresponding standard form of an arithmetic seris becomes $a + (a + d) + (a + 2d) + (a + 3d) + \cdots$

Example 5:

Suppose we invest Rs. 100 at a simple interest of 15% per annum for 5 years. The amount at the end of each year is given by

115, 130, 145, 160, 175.

This forms an arithmetic progression.

The n^{th} term of an A.P.:

The n^{th} term of an A.P. is also called the general term of the standard A.P. It is given by

$$T_n = a + (n-1)\, d \;;\; n = 1, 2, 3, \cdots$$

Sum of the first n terms of an A.P.:

Consider the first n terms of an A.P.

$$a,\ a + d,\ a+2d, \cdots, a+(n-1)\, d$$

The sum, s_n of these terms is given by

$$S_n = a + (a + d) + (a + 2d) + (a + 3d) + \cdots + (a + (n-1)d)$$

$$= (a + a + \cdots + a) + d\{1 + 2 + 3 + \cdots + (n-1)\}$$

$$= n \cdot a + d\left\{\frac{n(n-1)}{2}\right\}$$ (using formula for the sum of first (n - 1) natural numbers)

$$= \frac{n}{2}\{2a + (n-1)\, d\}$$

Example 6:

Suppose Mr. X repays a loan of Rs. 3250 by paying Rs. 20 in the first month and then increases the payment by Rs. 15 every month. How long will be take to clear his loan?

Solution:

Since Mr X increases the monthly payment by a constant amount, Rs. 15 every month, therefore d = 15 and first month instalment is a = Rs. 20. This forms an A.P. Now if the entire amount be paid in n monthly instalments, then we have

$$S_n = \frac{n}{2}\left\{2a + (n-1)\,d\right\}$$

or $$3250 = \frac{n}{2}\left\{2 \times 20 + (n-1)\,15\right\}$$

$$6500 = n\left\{25 + 15\,n\right\}$$

$$15\,n^2 + 25n - 6500 = 0$$

This is a quadratic equation in n. Thus to find the values of n which satisfy this equation, we shall apply the following formula as discussed before.

$$n = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-25 \pm \sqrt{(25)^2 - 4 \times 15 \times (-6500)}}{2 \times 15}$$

$$= \frac{-25 \pm 625}{30}$$

$$= 20 \quad \text{or} \quad -21{\cdot}66$$

The value, n = - 21.66 is meaningless as n is positive integer. Hence Mr. X will pay the entire amount in 20 months.

2.8 Geomentric Progression:

A Geometric Progression (GP) is a sequence whose each terms increases or decreases by a constant ratio called "common ratio or G.P." is obtianed after the first by multiplying the preceding term by a constant r. The standard form of a G.P. is written as $a, ar, ar^2, \ldots$ where 'a' is called the first term. Thus the corresponding geometric series in standard form becomes $a + ar + ar^2 + \ldots$

Example 7:

Suppose we invest Rs. 100 at a compound interest of 12% per annum for three years. The amount at the end of each year is calculated as follows:

(i) Interest at the end of first year $= 100 \times \frac{12}{100} =$ Rs. 12

Amount at the end of first year = Principal + Interest

$$= 100 + 100\left(12/100\right)$$

$$= 100\left(1 + \frac{12}{100}\right)$$

This shows that the principal of Rs. 100 becomes Rs. $100\left(1 + \frac{12}{100}\right)$ at the end of first year.

(ii) Amount at the end of second year = (Principal at the beginning of second year)

$$\left\{1 + \frac{12}{100}\right\}$$

$$= 100\left\{1 + \frac{12}{100}\right\}\left\{1 + \frac{12}{100}\right\}$$

$$= 100\left\{1 + \frac{12}{100}\right\}^2$$

(iii) Amount at the end of third year $= 100\left\{1 + \frac{12}{100}\right\}^2\left\{1 + \frac{12}{100}\right\}$

$$= 100\left\{1 + \frac{12}{100}\right\}^3$$

Thus, the progression giving the amount at the end of each year is

$$100\left\{1 + \frac{12}{100}\right\};\ 100\left\{1 + \frac{12}{100}\right\}^2;\ 100\left\{1 + \frac{12}{100}\right\}^3;\ \cdot\ \cdot\ \cdot$$

This is a G.P. with common ratio $r = \left(1 + \frac{12}{100}\right)$

In general, if P is the principal and i is the compound interest rate per anum, then the amount at the end of first year becomes $p(1 + i)$. Also the amount at the end of successive years forms a G.P.

$$P\left(1 + \frac{i}{100}\right);\ P\left(1 + \frac{i}{100}\right)^2;\ \cdot\ \cdot\ \cdot$$

with $r = \left(1 + \frac{i}{100}\right)$

The n^{th} term of G.P.

The n^{th} term of G.P. is also called the general term of the standard G.P. it is given by

$$T_n = a r^{n-1}, \; n = 1, 2, 3, \cdots$$

It may be noted here that the power of r is oneless than the index of T_n, which denotes the rank of this term in the progression.

Sum of the first in terms in G.P.:

Consider the first n terms of the standard form of G.P. $a, ar, ar^2, \cdots, a r^{n-1}$.

The sum, S_n of these terms is given by

$$S_n = a + ar + ar^2 + \cdots + ar^{n-2} + ar^{n-1} \qquad (2.4)$$

Multiplying both sides by r, we get

$$r\, S_n = ar + ar^2 + ar^3 + \cdots + ar^{n-1} + ar^n \qquad (2.5)$$

Subtracting (2.5) from (2.4), we have

$$S_n - r s_n = a - a r^n$$

$$S_n (1 - r) = a \left(1 - r^n\right)$$

or $S_n = \frac{a\left(1 - r^n\right)}{1 - r}$; $r \neq 1$ and < 1

Changing the signs of the numerator and denominator,

We have $S_n = \frac{a\left(r^n - 1\right)}{r - 1}$, $r \neq 1$ and > 1

(a) If $r = 1$, G.P. becomes a, a, a, ... so that S_n in this case is $S_n = n \cdot a$

(b) If number of terms in a G.P. are infinte, then

$$S_n = \frac{a}{1-r}\ ,\ r < 1$$

$$= \frac{a}{r-1}\ ,\ r > 1$$

Example 8:

A car is purchased for Rs. 80,000. Depreciation is calculated at 5% per annum for the first 3 years and 10% per annum for the next 3 years. Find the money value of the car after a period of 6 years.

Solution:

(i) Depreciation for the first year $= 80,000 \times \frac{5}{100}$. Thus the depreciated value of the car at the end of first year is

$$= \left(80,000 - 80,000 \times \frac{5}{100}\right)$$

$$= 80,000\left(1 - \frac{5}{100}\right)$$

(ii) Depreciation for the second year

= (depreciated value at the end of first year) x Rate of depreciation for second year

$$= 80,000\left(1 - \frac{5}{100}\right)\left(\frac{5}{100}\right)$$

Thus the depreciated value at the end of the second year is = (Depreciated value after first year) - (Depreciation for second year)

$$= 80,000\left(1 - \frac{5}{100}\right) - 80,000\left(1 - \frac{5}{100}\right)\left(\frac{5}{100}\right)$$

$$= 80,000\left(1 - \frac{5}{100}\right)\left(1 - \frac{5}{100}\right)$$

$$= 80,000\left(1 - \frac{5}{100}\right)^2$$

Calculating in the same way, the depreciated value at the end of three years is

(iii) Depreciation for fourth year

$$= 80,000\left(1-\frac{5}{100}\right)^3\left(\frac{10}{100}\right)$$

Thus the depreciated value at the end of the

fourth year is = (Depreciated value after three year) x depreciated value for fourth year

$$= 80,000\left(1-\frac{5}{100}\right)^3 - 80,000\left(1-\frac{5}{100}\right)^3\left(\frac{10}{100}\right)$$

$$= 80,000\left(1-\frac{5}{100}\right)^3\left(1-\frac{10}{100}\right)$$

Calculating in the same way, the depriciated value at the end of six years becomes

$$= 80,000\left(1-\frac{5}{100}\right)^3\left(1-\frac{10}{100}\right)^3$$

$$= \text{Rs. } 49,980\cdot 24$$

2.9 Solved Problems:

1. The fixed cost for a firm is Rs. 30,000 and the unit variable cost is Rs. 0.50 If the product can be sold at Rs. 2 per unit, construct the cost and revenue functions for the firm. What will be the new break even point if selling price is reduced to Rs. 1.50 due to market fluctuations?

Solution:

Fixed cost a = 30,000 variable / unit

b = 0.50

p = Rs. 2

Revenue Function = Px = 2 x

Cost Function = a + b x

= 30000 + 0.5 x

At break even point revenue = cost

2 x = 30,000 + 0.5 x

1.5 x = 30,000

$$x = \frac{30000}{1 \cdot 5}$$

$$= \frac{300000}{15}$$

$$x = 20,000$$

Revenue Function = $p\,x = 1 \cdot 5\,x$

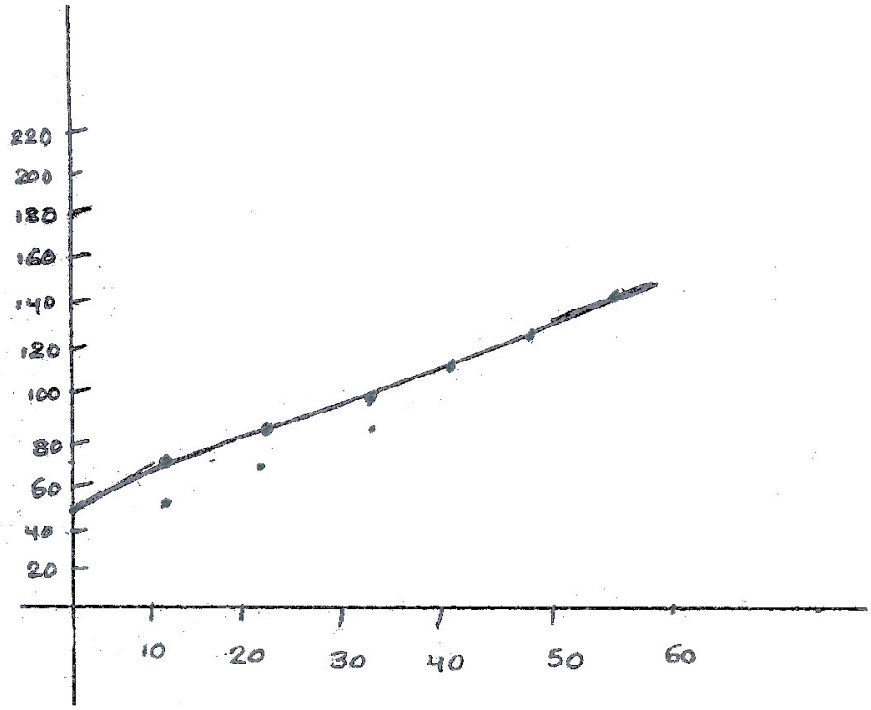
$$1 \cdot 5\,x - 0 \cdot 5\,x = 30,000$$

$$x = 30,000$$

2. Let there be two cost - functions C_1 and C_2 given by $C_1 = 0 \cdot 2\,x + 200$ and $C_2 = 0 \cdot 6\,x + 50$ where x is the level of output. Draw the graph of these functions and find out which cost function increases more rapidly with the increase in output.

Solution:

x	0	10	20	30	40	50	60
C_1	200	202	204	206	208	210	212
C_2	50	56	62	68	74	80	86



3. The firm wants to take a decision for introducing a new product. The cost of introducing the product (initial advertising promotion and fixed cost for one year of production) is estimated at Rs. 30,000. The variable cost per unit is Rs. 30 and the proposed selling price is Rs. 50

   a, find out the break - even level of production

   b, Determine the profit on sale of 2,500 units represent the cost, Revenue and profit functions graphically.

Solution:

Fixed cost a = 30,000

b = 30

P = 50

Revenue Function $p\,x = 50\,x$

Cost Function $a + b\,x = 30{,}000 + 30\,x$

At break even point revenue = cost function

$50\,x = 30{,}000 + 30\,x$

$20\,x = 30{,}000$

$x = 1500$

P = R - C

$= 50 \times 2500 - 30000$

$= -30000 + (50 - 30)\,2500$

$= -30000 + 50000$

P = 20,000

4. Vijay wants to predict sales for his company for his coming year on the basis of historical data, he concludes that a linear function passes through the observed data points for year 1 and year 6. The sales for the year 1 is Rs. 24,000 and for the year 6 is Rs. 32,000 what will be the sales for the seventh year?

Solution:

Linear function of sales

$$A = (a, f(a))$$

$$B = (b, f(b))$$

Then the output linear function

$$y = \frac{f(b) - f(a)}{b - a}\,x + \frac{-a\,f(b) + b\,f(a)}{b - a}$$

$$x = 7, \qquad y = \frac{8000}{5} \times 7 + \frac{6 \times 24000 - 1 \times 32000}{5}$$

$$= 8000 \times \frac{7}{5} + \frac{8000\,(18 - 4)}{5}$$

$$= \frac{8000}{5}\,(7 + 18 - 4)$$

$$= 1600 \times 21$$

$$= 33,600$$

5. In inventory problems, cost is a function of order size. The following data is available for a product.

Annual requirement (R) = 12,000

Ordering Cost per order (S) = 150

Cost per unit (C) = 4

Carrying cost per unit (I) = 0.20

State the ordering cost, inventory carrying cost and total cost functions.

Solution:

Ordering cost $= \frac{12,000}{Q} \times 150$

Carrying Cost $= \frac{Q}{2} \times 0 \cdot 20 \times 4$

Total Cost = ordering cost + carrying cost

$$= \frac{12,000}{Q} \times 150 + \frac{Q}{2} \times 0 \cdot 20 \times 4$$

6. The demand function for a product is given below:

$$p = 3 \cdot 50 - 0 \cdot 50\, q$$

Derive the total revenue function taking quantity demanded as independent variable.

Solution:

Total Revenue = Price x Quantity demanded

$$= p \times q$$

$$= (3 \cdot 50 - 0 \cdot 50\ q)\ q$$

$$= 3 \cdot 50\ q - 0 \cdot 50\ q^2$$

7. A market supply function for a government supported farm commodity is q = 5 p where q denotes the quantity supplied and p dentoes the market price. Each unit produced costs Rs. 2. The government feels that if the farmers as a group are to receive a reasonable price, total profit should be Rs. 300 what price would farmers have to receive in order to realise this profit ?

Solution:

Total profit = Total revenue - total cost

= Price x Quantity supplied - cost x quantity supplied

$= p \times q - c \times q$

$= q\,(p - c)$

$= (100 + 5p)\,(p - 2)$

If profit = 300, then

$$(100 + 5p)\,(p - 2) = 300$$

$$100p + 5p^2 - 200 - 10p = 300$$

$$5p^2 + 90p - 500 = 0$$

$$p = \frac{-90 \pm \sqrt{8100 - 4 \times 5\,(-500)}}{2 \times 5}$$

$$= \frac{-90 \pm \sqrt{8100 + 10000}}{10}$$

$$p = \frac{-90 \pm 134 \cdot 6}{\cdot 0} = -22 \cdot 46 \text{ and } 4 \cdot 46$$

Since negative prices have no economic meaning, the required price is Rs. 4.46 per unit.

8. A company is considering as to how much maximum reserve it should transfer (capitalise) to equity for the purpose of declaring bonus issue given the restriction that the residual reserve after capitalisation should be atleast 1/3 rd of the increased paid up capital. The

existing paid - up capital of the company is Rs. 50 lakhs and the existing level of free reserve is Rs. 75 lakhs.

Solution:

Suppose the company transfers X amount to the paid up capital. Then the increased paid up capital will be C + X, where C is the existing paid - up capital and residual reserves will be FR - X. Where F R is the existing free reserves. Hence, the maximum amount that can be capitalised by the company will be given by the equation.

$$\text{F R} - \text{X} \geq \frac{1}{3}(\text{C} + \text{X})$$

Since C = 50 and F R = 75

Substituting these values in the above equation,

We get

$$75 - \text{X} \geq \frac{1}{3}(50 + \text{X}) \text{ or } 225 - 3\text{X} \geq 50 + \text{X} \text{ or } 175 \geq 4\text{X} \text{ or } \text{X} \leq \frac{175}{4} \text{ or } \text{X} \leq 43 \cdot 75$$

Thus the maximum amount that can be capitalised is Rs. 43.75 lakh.

9. The test for capitalising is that 30% of the average profits before tax for the previous three years should yield a rate of divided on the expanded capital base of the company at 9 percent.

If the average profit earned by the company for the last three years is Rs. 22.5 lakhs and the existing capital base is Rs. 50 lakhs, how much amount can be transferred from the free reserves to the paid up capital?

Solution:

Let X be the amount to be transffered to the paid up captial. The new capital base will be $(\text{C} + \text{X})$. If A P represents the average profit for last three years, then the test for capitalisation can be represented by the following equation.

$$\frac{30}{100} \text{A} \cdot \text{P} \cdot = \frac{9}{100}(\text{C} + \text{X})$$

$$\frac{10}{3} \text{A P} = \text{C} + \text{X} \text{ or } \text{X} = \frac{10}{3} \text{ A P} - \text{C}$$

Substituting $\text{A P} = 22 \cdot 5$ and $\text{C} = 50$ we obtain

$$\text{X} = \frac{10}{3} \times 22 \cdot 5 - 50$$

$$\text{X} = 25$$

Thus, Rs. 25 lakhs can be transffered to paid up capital.

2.10 Summary:

The objective of this unit is to provide you exposure to functional relationship among decision variables. We started with the mathematical concept of function and defined terms such as constant, parameter, independent and dependent variable. Various examples of functional relationshipsare mentioned to see the concept in broad perspective. VArious types of functions which are normally used in managerial decision - making are enumerated along with suitable examples, their graphs and solution procedure. Finally, the applications of functional relationships are demonstrated through several examples.

Attention is then directed to defining the Arithmetic and Geometric Progressions and subsequently to their applications.

2.11 Technical Terms:

Arithmetic Progression (A.P.) : An A.P. is a sequence whose terms increases or decreses by a constant number.

Algebraic and Transcendental Function : When only finite number of terms are involved in a functional relationship and variables are affected only by the mathematical operations, then functions are called algebraic function, otherwise transcendental function.

Constant : A qunatity that remains fixed in the context of a given problem or situation.

Exponential Function : If the independent variable in any functional relationship appears an exponent (or power), then such functional relationship is called exponential function.

Function : It is the rule of correspondence between dependent variable and independent variable(s) so that for every assigned value to the independent variable, the corresponding unique value for the dependent variable is determined.

Geometric Progression (G.P.) : A G.P. is a sequence whose terms increases or decreases by a constant ratio.

Linear Function : A function whose graph is a straight line is called a linear function.

Logarithmic Function : The inverse of exponential function is called a logarithmic function.

Parameter : A quantity that retains the same value throughout any particular problem but may assume different values in different problem.

Polynomial Function : A function of degree n is called a polynomial function of degree n.

Series : A series is formed by connecting the terms of a sequence with plus or minus sign.

Sequence : If for any positive integer n, there corresponds a number a_n such that a_n is related to n by some rule, then the terms $a_1, a_2, \cdots, a_n$ are said to form a sequence.

Step Function : If for values of an independent variable, the dependent variable takes a constant value in different intervals then the function is called step function.

Variable : A quantity that can assume various values.

2.12 Exercise:

1. Find the domain and range of the following functions:

   a) $y = \frac{1}{x-1}$

   b) $y = \sqrt{x}$, $y \leq 0$

   c) $y = \sqrt{4-x}$, $y \geq 0$

2. Let $4p + 6q = 60$ be an equation containing variables p (price) and q (quantity). Identify the meaningful domain and range for the given function when price is considered as independent variable.

3. Draw the graph of the following functions:

   a) $y = 3x - 5$

   b) $y = x^2$

   c) $y = \log_2 x$

4. Given that $f(x) = (x-4)(x+3)$ then find,

   a) $f(4), f(-1), f(-3)$

   b) Roots of the function

5. A company sells x units of an item each day at the rate of Rs. 50 per unit. The cost of manufacturing and selling these units is Rs. 35 per unit plus a fixed daily overhead cost of Rs. 100. Determine the profit function. How would you interpret over the situation if the company manufactures and sells 400 units of the items a day.

6. Let the market supply function of an item be $g = 160 + {}^{8}\!/_{p}$, where of denote the quantity supplied and p denotes the market price. The unit cost of production is Rs. 4 It is felt that the total profit should be Rs. 500 What market has to be fixed for the item so as to achieve this profit?

7. Consider the quadratic equation $2x^2 - 8x + c = 0$. For what value of C , the equation has

   i) Real Roots ii) Equal roots and iii) Imaginary roots.

8. A news boy buys papers for p_1 paise per paper and sells then at a price of p_2 paise per paper $(p_2 > p_1)$. The unsold papers at the end of the day are bought by a waste paper dealer for p_3 paise per paper $(p_3 < p_1)$.

   i) construct the profit function of the newsboy.

   ii) Construct the opportunity loss function of the newsboy.

9. Find the 15^{th} year term of an A.P. whose first term is 12 and common difference is 2.

10. A firm produces 1500 TV sets during its first year. The total production of the firm at the end of the 15^{th} is 8300 TV sets, then

    a) Estimate by how many units, production has increased each year.

    b) Based on estimate of the annual increment in production, forecast the amount production for the 10^{th} year

11. Determine the common ratio of the G.P.

    $49,\ 7,\ {}^{1}\!/_{7},\ {}^{1}\!/_{49},\ \ldots$

    a) Find the sum of first 20 terms of G.P.

    b) Find the sum to infinite terms of G.P.

12. The population of a country in 1985 was 50 crore. Calculate the population in the year 2000 if the compounded annual rate of increase is a) 1 % b) 2%.

2.13 Reference Books:

1. Childress R.L. 1974, Mathematics for Managerial Decision, Prentice Hall Inc; Englewood - Cliffs.

2. Dean, B.V. Sassieni M.W. and Gupta S.K. 1978 Mathematics for Modern Management Wiley Eastern ; New Delhi.

3. Draper, J.E. ; and J.S.Klingman, 1972 Mathematical Analysis : Business and Economic Applications, Harpera and Row Publishes ; Newyork.

4. Raghavachari M. 1985 Mathematics for Management An introduction Tata Mc Graw - Hill Publishing Company. Ltd., New Delhi.

5. Srivatsava U.K. Shenoy G.V. and Sharma, S.C. 2008 Quantitative Techniques for Managerial decisions, New Age international Publishers, New Delhi.

Lesson Writer

Dr. S.V.S. GIRIJA

Lesson - 3

BASIC CALCULUS AND APPLICATIONS

Objectives:

After studying this lesson you should be able to understand the:

- Meaning of the term "calculus" and its branches
- Concept of limit and slope which are fundamentals to an understanding of calculates.
- Meaning of differential calculus
- The type of decision problems which can be solved with the help of differential calculus.

Structure:

3.1 Introduction:

In the past, the term "calculus" as a branch of mathematics was familiar only to scientists. The managers and students of business management were little concerned about its usefulness. But, with the increasing need of quantitative techniques, particularly the classical optimization techniques in the solution of business problems, there is a growing tendency to use quantitative techniques based on calculus in the solution of business problems. Calculus based techniques are extensively used in economics, operations management, marketing, financial management etc.

Calculus is particularly useful in the those situations where we are interested in estimating the rate at which things change. For example, it has a role to play when we are interested in knowing how the sales volume or sales is affected when the prices change or how the total cost, price, etc... are affected when the volume of output changes.

There are two branches of calculus: differential calculus and integeral calculus. These two are reverse of each other, as are addition, and subtraction and multiplication and division. Differential calculus is concerned with determining the rate of change of a given function due to a unit change in one on the independent variables while, integral calculus is concerned with the inverse problem of finding a funtion when its rate of change is given. This cannot be illustrated with real examples because integral calculus in beyond the scope of this unit. In this unit we will be concerned only with differential calculus.

Analysis in business and economics is frequently concerned with change, therefore differential calculus should find wide applications in business. Marginal analysis in economics is perhaps the most direct application of differential calculus in business. Also business problems concerned with such things as maximisation of profits and minimisation of costs under various assumptions can be solved using differential calculus.

The objective of this unit is to give you an idea about the rate of change of a function. The applications of this concept to marginal analysis and to various problems of maximisation and minimisation are discussed in this unit.

3.2 Limit and Continuity:

A) Limit: Sometimes, we wish to determine the behaviour of a function $y = f(x)$ as the independent variable x approaches some particular value, say 'a'. For example, it may be interesting to know limiting saturation level of sales as advertising efforts are increased. The formal definition of limit may look little abstract, therefore the notion of limit of a function is easier to understand in an intuitive sense. consider a function $f(x)$ defined as:

$$f(x) = x - 1$$

Now as we give values to x which are nearer and nearer to 1, the value of the function $f(x)$ become smaller and smaller and become closer and closer to zero.

This phenomenon of x approaches a value 'a' termed as 'x tends to a' and it is symbolically written as $x \to a$. The corresponding value of $f(x)$, say 'L' as $x \to a$ is called the limit of the function, and it is symbolically written as:

$$\underset{x \to a}{\text{Limit}}\ f(x) = L \quad \text{or} \quad \underset{x \to a}{\text{Lt}}\ f(x) = L$$

$$\text{or } f(x) \to L \text{ as } x \to a$$

Example 1:

If $f(x) = 2x + 5$, then $\text{Lt} \cdot f(x) = 5$. It can be illustrated as shown below:

x	$y = f(x) = 2x + 5$
2	9
1	7
1/2	6
1/5	27/5
1/10	26/5
1/100	251/50
1/1000	2501/500

Alternative symbolical notations of the limit of the given function when we allow x to take different values are as follows:

	Notations	**Example**
i)	$\underset{x \to \infty}{\text{Lt}} f(x) = L$	$\underset{x \to \infty}{\text{Lt}} \left(1 + \frac{1}{x}\right) = 1$
	or $f(x) \to L$ as $x \to \infty$	
ii)	$\underset{x \to -\infty}{\text{Lt}} f(x) = L$	$\underset{x \to -\infty}{\text{Lt}} \left(1 - \frac{1}{x}\right) = 1$
	or $f(x) \to L$ as $x \to -\infty$	
iii)	$\underset{x \to \infty}{\text{Lt}} f(x) = \infty$, and	$\underset{x \to \infty}{\text{Lt}} \left(x^3 - 8\right) = \infty$ and also
	$\underset{x \to -\infty}{\text{Lt}} f(x) = \infty$	$\underset{x \to -\infty}{\text{Lt}} \left(x^3 - 8\right) = \infty$
iv)	$\underset{x \to a}{\text{Lt}} f(x) = \infty$	$\underset{x \to 2}{\text{Lt}} \left\{ \frac{1}{(x-2)^2} \right\} = \infty$
	or $f(x) \to \infty$ as $x \to a$	

Then

$$\lim_{n \to \infty} f(x) = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

$$= e \; (= 2 \cdot 71828)$$

Also, for every real number x, we have

$$e^x = \lim_{x \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Example 3:

Let a sum of Rs. P be initially lent at the rate of r per rupee per annum to be compounded annually. Then the compound value of money at the end of n years is given by

$$A = P\,(1 + x)^n$$

But if the interest be compounded more than once a year, then we have

$$A = P\left(1 + \frac{r}{m}\right)^{mn}$$

$$= P\left[\left(1 + \frac{x}{m}\right)^{\frac{m}{n}}\right]^m$$

where m is the number of times per year compounding occurs. That is, the interest be compounded at intervals of $\frac{1}{m}$ years.

If $m \to \infty$, that is, interest is compounded at very very small intervals, then we have

$$\frac{m}{r} \to \infty, \; \frac{r}{m} \to 0 \text{ and } \lim_{m \to \infty} \left(1 + \frac{r}{m}\right)^{m/r} = e$$

and also, $A = P \cdot e^m$

Hence, a sum of Rs. P invested initially at the rate of r per rupee per annum to be compounded continuously, becomes $A = P \cdot e^m$ at the end of n years.

Continuity:

A function $y = f(x)$ is said to be continuous at a point $x = a$ if

i) $f(a)$ exists (or defined)

ii) $\underset{x \to a}{Lt}\ f(x)$ exists

iii) $\underset{x \to a}{Lt}\ f(x) = f(a)$

Condition (iii) implies that both right hand limit and left hand limit should exist and be equal to the value of the function at $x = a$. That is, limit of $f(x)$ in the neighbourhood (i.e. close to) of $x = a$ (or at $x = a + h$ and $x = a - h$, where $h \to 0$) should exist.

The limit is said to exist if its value is finite. For example, if $Lt\ .f(x) = \infty$ as $x \to a$, then this means $f(x)$ becomes arbitrarily large as x approaches a. It should be remembered that ∞ is not a number.

A function $f(x)$ is said to be continuous in (or on) an open interval (b, c) or closed interval $[b, c]$ if it is continuous at each and every point of the interval. Otherwise it is said to be discontinuous.

From this function of continuity, it follows that the graph of a function that is continuous in (or on) an interval consists of unbroken curve (i.e. a curve that can be drawn without raising the pen from the paper) over that interval as shown in Figure I (a) and I (b)

Fig I (a)

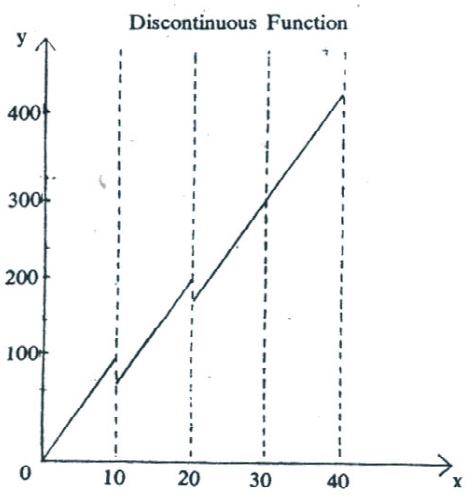
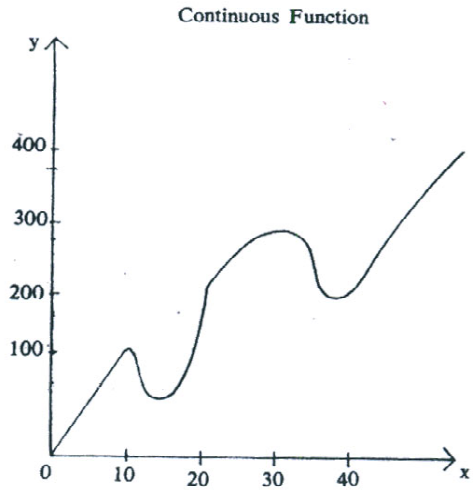


Fig I (b)



Example 4:

Discuss the nature of the following functions

(a) $f(x) = \frac{1}{x-2}$ at $x = 2$

(b) $f(x) = x^2$ at $x = 2$

Solution:

(a) The function $y = \frac{1}{x-2}$ is discontinuous at x = 2 because

$$f(2) = \frac{1}{0} = \infty$$

i.e. the function is not defined for x = 2 because it does not have finite value

(b) $f(2) = (2)^2 = 4$ (finite value)

Also $R \cdot H \cdot L \cdot = \underset{h \to 0}{Lt} (2+h)^2 = \underset{h \to 0}{Lt} (4 + h^2 + 2h) = 4\,(\text{finite})$

$$L \cdot H \cdot L \cdot = \underset{h \to 0}{Lt} (2-h)^2 = \underset{h \to 0}{Lt} (4 + h^2 - 2h) = 4\,(\text{finite})$$

Since all the conditions of continuity are satisfied, therefore function is continuous.

3.3 Concept of Slope and Rate of Change:

The term slope is used to measure the degree of steepness or rate of change of a function. In general, it is defined as the change in the dependent variable caused by one unit of change in one of the independent valuables. The slope is denoted by 'm' or '$\tan\theta$' (θ is the angle of inclination of the given line with x - axis).

Slope of a Straight Line:

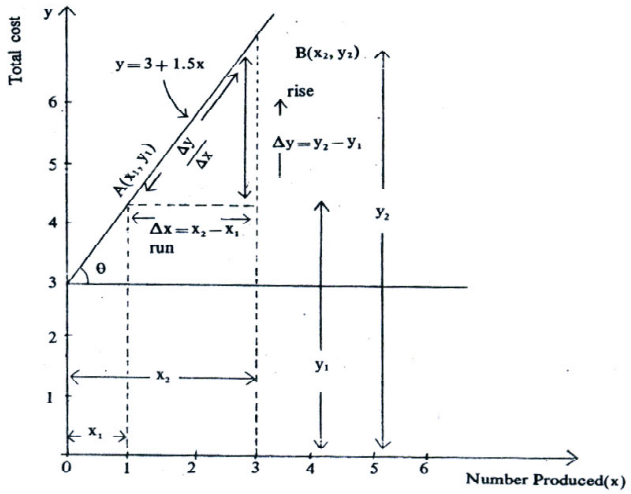
Consider the case of total cost of producing an item. Usually total cost of production is a function of the fixed (set - up) cost plus a constant additional cost for each item produced. If fixed cost is Rs. 3 and additional cost is Rs. 1.5, then total cost, y is represented by

$$y = 3 + 1 \cdot 5\,x$$

where x is the number of items produced. Clearly x is the independent variable and y is the dependent variable.

This equation has been graphed in fig II. It represents a straight line.

Fig II



Consider two points A and B on the line whose coordinates are (x_1, y_1) and (x_2, y_2) respectively. Suppose, we employ the symbol Δ (delta) to indicate a very small change in the value of a variable or quantity. This change can be positive or negative change. If Δx represents the change (or increment) in the value of x and Δy represents the change in the value of y due to change in x, then the ratio $(\Delta y/\Delta x)$ of the change in dependent variable y due to one unit change in dependent variable, is called the slope and is defined as

$$m = \tan\theta = \frac{\text{rise}}{\text{run}} \text{ or } \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{7\cdot5 - 4\cdot5}{3-1} = 1\cdot5 \text{ (Coefficient of x)}$$

Thus, in the case of straight line relationship which we are currently considering, the slope is simply given by the coefficient of the independent variable. In this case the slope is $+1\cdot5$ (the plus sign indicates that y increases when x increases and vice - versa).

Further considering the equation of the line $y = 3$ or $3 + 0\cdot x$ (i.e. cost of production is independent of the number of items produced). It is obvious that terms involving x has a coefficinet of zero. That is the slope of this line is zero and hence it is a horizontal line as shown in Figure 2. It should be noted that the slope (rate of change) of a line remains constant at all points on the line, i.e. rate of change of y as x changes is constant throughout the length of the line. However, the slope of a curve (i.e. a non linear function) changes from point to point and thus the slope must be determined for each particular point of interest.

Positive and Negative Slope:

The slope $+1\cdot5$ in the case just discussed is the example of positive slope which indicates that dependent variable y increases (or decreases) as independent variable x increases (or

decreases). But if the value of dependent variable y decreases as independent variable x increases and vice - versa, then slope is always negative. For example, let the sales of an item be the function of the price charged, and the exact relationship between these two is given by

$$y = 100 - 5x$$

In this case the slope is - 5 (negative) which indicates that as sales, y decreases with increasing values of price x and vice - versa.

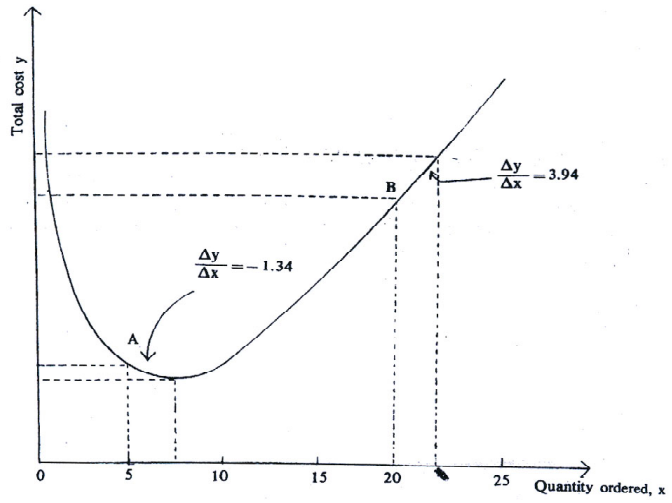
Slope of a curve (at a point):

For non - linear functions, the slope changes from point to point. Thus, it is necessary to specify the point at which the slope is to be determined. The procedure for computing the slope in this case is also same as in the case of the straight line. This means, that we must compute the ratio $\frac{\Delta y}{\Delta x}$ at a specified point. Suppose total cost 'y' of the stock of an item as a function of order quantity, 'x' is represented as :

$$y = 4x + \frac{200}{x}$$

This equation has been graphed in Figure 3. It represents a curve

Figure III



For $x = 5,\ y = 20 + \frac{200}{5} = 60$

$$x = 7 \cdot 5,\ y = 30 + \frac{200}{7 \cdot 5} = 56 \cdot 66$$

Between $x = 5$ and $x = 7 \cdot 50$, we have

$$\frac{\Delta y}{\Delta x} = \frac{56 \cdot 66 - 60}{7 \cdot 5 - 5} = -1 \cdot 34$$

For $x = 20,\ y = 80 + \frac{200}{20} = 90$

$$x = 22 \cdot 5,\ y = 90 + \frac{200}{22 \cdot 5} = 98 \cdot 88$$

Between $x = 20$ and $x = 22 \cdot 5$ we have

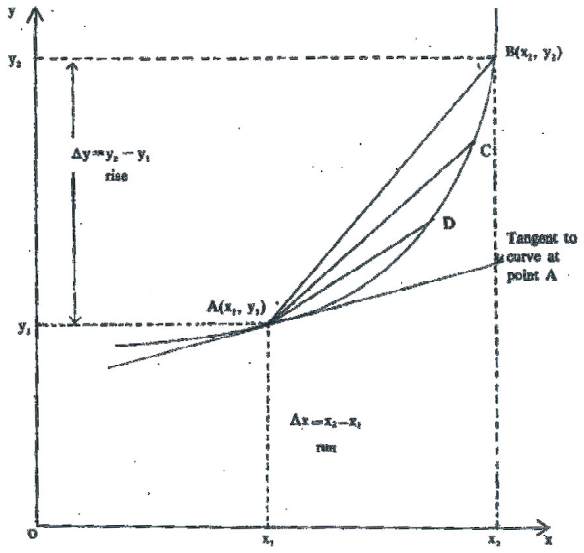
$$\frac{\Delta y}{\Delta x} = \frac{98 \cdot 88 - 90}{22 \cdot 5 - 20} = +\ 3 \cdot 94$$

From these two values, it is clear that the slope of a curve is different at different points, and the absolute value of the ratio $\frac{\Delta y}{\Delta x}$ in the first case is smaller as compared to the absolute value of the ratio $\frac{\Delta y}{\Delta x}$ in second case. This shows that the value of y is much more sensitive to changes in the lower range of x. The negative slope between $x = 5$ and $7 \cdot 5$ indicate that the total stock holding cost decreases as size of order increse on this part of the curve. Whereas between $x = 20$ and $x = 22 \cdot 5$, stock holding cost increses as size of order increases on this part of the curve.

3.4 Concept of Derivative:

The term derivative is a generalised expression for measuring the rate of change or slope of a function. Supposing A and B are two points on the curve (figure iv) whose coordinates are $(x_1,\ y_1)$ and $(x_2,\ y_2)$ respectively.

Figure IV



In figure IV, the average slope of the curve between two points A and B is measured by the slope of the line joining the points A and B. That is

$$\text{Slope of the line } AB = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} \qquad (3.1)$$

Assuming that the mathematical equation of the curve in the figure is represented by $y = f(x)$. Then

$$y_1 = \text{the value of } f(x) \text{ at } x = x_1$$

$$= f(x_1)$$

Similarly $y_2 = f(x_2)$

Substituting for y_1 and y_2 in equation (3.1), we have

$$\frac{\Delta y}{\Delta x} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \qquad (3.2)$$

As $x_2 > x_1$ then let $x_2 = x_1 + \Delta x_1$ where Δx_1 represents small change in x_1. Therefore,

$$x_2 = x_1 + \Delta x_1 \text{ and } f(x_2) = f(x_1 + \Delta x_1)$$

Substituting for x_2 and $f(x_2)$ in equation $(3 \cdot 2)$ we have

$$\frac{\Delta y}{\Delta x} = \frac{f\left(x_1 + \Delta x_1\right) - f\left(x_1\right)}{\left(x_1 + \Delta x_1\right) - x_1}$$

$$= \frac{f\left(x_1 + \Delta x_1\right) - f\left(x_1\right)}{\Delta x_1} \qquad (3.3)$$

Equation (3.3), represents the slope of the straight line A B, rather than of the curve AB.

If we keep on making Δx_1 smaller, we approach a point such as A, and obtain a line that touches the curve only at the point A. This line is the tangent to the curve at the point A (tangent at a point is defined as the line that touches the curve only at that point and does not cross the curve at that point). Now when Δx_1 is very very small, and point B will be extremely close to A. In mathematics, this is known as taking the limit of the ratio $\frac{\Delta y}{\Delta x}$ as $\Delta x_1 \to 0$. Hence from equation $(3 \cdot 3)$, we have

$$\text{Slope of the curve at point } A = \underset{\Delta x_1 \to 0}{\text{Lt}} \left[\frac{f\left(x_1 + \Delta x_1\right) - f\left(x_1\right)}{\Delta x_1} \right]$$

In general the slope of the curve at any point. $A\left(x, y\right)$ is defined as:

$$\frac{dy}{dx} = \underset{\Delta x \to 0}{\text{Lt}} \left[\frac{\Delta y}{\Delta x} \right] = \underset{\Delta x \to 0}{\text{Lt}} \left[\frac{f\left(x + \Delta x\right) - f\left(x\right)}{\Delta x} \right]$$

Hence, we can say that the the derivative of a function is the generalised expression for the slope of a function. Further, if we can calculate the derivative at any point on a curve, this means we know the value of the slope at that point. Another interpretation of the derivative $\frac{dy}{dx}$ is that it measures the rate of change of the varibale y with respect to the variable x.

At any point where the limit of $(3 \cdot 3)$ does exist, the function $y = f\left(x\right)$ is said to have a derivative or to be differentiable and $\frac{dy}{dx}$ is said to be the first derivative or the derivative of $y = f\left(x\right)$. The process of obtaining the first derivative of a function is referred to as differentiation. Various types of notations, in addition to $\frac{dy}{dx}$ are used to denote the first derivative of $y = f\left(x\right)$ with respect to x. The most common of these are

$f'(x)\ ;\ y'\ ;\ \frac{d}{dx}(y)\ ;\ D_x(y)$

3.5 Rules of Differentiation:

Some of the most commonly used rules of differentiation are as follows:

Polynomial Functions:

a) Derivative of a constant:

Let $y = K$, where K is a constant, then

$$\frac{dy}{dx} = \frac{d}{dx}(k) = 0$$

Example 5

i) If $y = 10$, then $\frac{dy}{dx} = 0$

ii) If $y = 0$, then $\frac{dy}{dx} = 0$

That is the derivative of a constant is always zero.

b) Derivative of a power function:

Let $y = kx^n$, where the coefficient K and exponent n are constant, then

$$\frac{dy}{dx} = \frac{d}{dx}\left(kx^n\right) = k\frac{d}{dx}\left(x^n\right)$$

$$= k \cdot n x^{n-1}$$

That is (i) the derivative of the product of a constant and a differentiable function is the product of the constant and the derivative of the function, and (ii) the derivative of the power function x^n equals the product of the exponent n and the variable x raised power one less than exponent, i.e., $(n-1)$.

Example 6

i) If $y = 10x$ then $\frac{dy}{dx} = 10 \cdot 1 \cdot x^{1-1} = 10x^0 = 10$

ii) If $y = 10x^4$ then $\frac{dy}{dx} = 10 \cdot 4 \cdot x^{4-1} = 40x^3$

c) Derivative of a sum (or a difference) of two (or more) functions:

Let $y = u \pm v$

where $u = f(x)$ and $v = g(x)$ are differential function of x, then

$$\frac{dy}{dx} = \frac{du}{dx} \pm \frac{dv}{dx}$$

That is derivative of the sum (or difference) of a finite number of differentiable functions equals the sum (or difference) of derivatives of the individual functions.

Example 7

i) If $u = 10x^4$ and $v = -5x^2$, $y = u + v$.

then

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx} = \frac{d}{dx}\left(10x^4\right) + \frac{d}{dx}\left(-5x^2\right)$$

$$= 40x^3 - 10x$$

ii) If $u = \frac{5}{x^2}$ and $v = 4x + 9x^2$, $y = u - v$

then

$$\frac{dy}{dx} = \frac{du}{dx} - \frac{dv}{dx}$$

$$= \frac{d}{dx}\left(5x^{-2}\right) - \frac{d}{dx}\left(4x + 9x^2\right)$$

$$= -10x^{-3} - \left(4 + 18x\right)$$

$$= \frac{-10}{x^3} - 4 - 18x$$

Algebraic Functions:

a) Derivative of a product of two functions:

Let $y = uv$

where $u = f(x)$ and $v = g(x)$ are differentiable functions of x, then

$$\frac{dy}{dx} = u \cdot \frac{dv}{dx} + v \cdot \frac{du}{dx}$$

That is the derivative of the product of two functions equas this sum; (first function) x (derivative of second function) + (second function) x (derivative of first function).

Example 8

If $u = x^3 + 5$ and $v = x^2$, $y = u \cdot v$

then

$$\frac{dy}{dx} = \left(x^3 + 5\right)\frac{d}{dx}\left(x^2\right) + x^2\frac{d}{dx}\left(x^3 + 5\right)$$

$$= \left(x^3 + 5\right)(2x) + x^2\left(3x^2 + 0\right)$$

$$= 2x^4 + 10x + 3x^4$$

$$= 5x^4 + 10x$$

b) Derivative of a quotient of two functions:

Let $y = \frac{u}{v}$, $v \neq 0$

where $u = f(x)$ and $v = g(x)$ are differentiable functions of x, then

$$\frac{dy}{dx} = \frac{v\frac{du}{dx} - u \cdot \frac{dv}{dx}}{v^2}$$

That is the derivative of the quotient of two functions equals; (the denominator) x (derivative of the numerator) - (the numerator) x (derivative of the denominator) then this difference divided by the square of the denominator.

Example 9

Let $u = x^3 + 5x + 1$ and $v = x + 2$

$$y = \frac{u}{u}$$

then

$$\frac{dy}{dx} = \frac{(x+2)\frac{d}{dx}\left(x^2 + 5x + 1\right) - \left(x^2 + 5x + 1\right)\frac{d}{dx}(x+2)}{(x+2)^2}$$

$$= \frac{(x+2)(2x + 5 + 0) - \left(x^2 + 5x + 1\right)(1+0)}{(x+2)^2}$$

$$= \frac{x^2 + 4x + 9}{(x+2)^2}$$

c) Derivative of the nth power of a function:

Let $y = u^n$

where $u = f(x)$ is a differentiable function of x and n is any number (positive or negative integer or non - integer) then

$$\frac{dy}{dx} = n \cdot u^{n-1} \frac{d}{dx}(u)$$

That is the derivative of such function equals the product of power n, the $(n-1)^{th}$ power of the function and derivative of the function.

Special Case:

If $u = f(x) = x$, then $u^n = x^n$ and $\frac{dy}{dx} = n \cdot x^{n-1}$ for any real number n.

Example 10

Let $u = \left(x^2 + 2x\right)^{-1/3}$; $y = u = \left(x^2 + 2x\right)^{-1/3}$

then

$$\frac{dy}{dx} = -\frac{1}{3}\left(x^2 + 2x\right)^{-1/3-1} \frac{d}{dx}\left(x^2 + 2x\right)$$

$$= \frac{1}{3}\left(x^2 + 2x\right)^{-4/3} \left(2x + 2\right)$$

$$= -\frac{\left(2x + 2\right)}{3\left(x^2 + 2x\right)^{4/3}}$$

d) Derivative of a function of a function:

Let $y = f(u)$ and $u = g(x)$

then $\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$

Example 11

Let $y = u^2$ and $u = 2x + 3$

Therefore

$$\frac{dy}{du} = 2u = 2\left(2x + 3\right) = 4x + 6$$

$$\frac{du}{dx} = 2 + 0 = 2$$

Hence

$$\frac{dy}{dx} = \frac{dy}{dx} \cdot \frac{du}{dx}$$

$$= \left(4x + 6\right) \cdot 2$$

$$= 8x + 12$$

Logarithmic Functions:

Let $y = \log_a u$

where $u = f(x)$ is a differentiable function. Then

$$\frac{dy}{dx} = \frac{1}{u}\log_a^e \cdot \frac{d}{dx}u$$

Exaple 12

i) Let $y = \left(\log x^2\right)^2$

then

$$\frac{dy}{dx} = 2\cdot\left(\log x^2\right)\frac{d}{dx}\left(\log x^2\right)$$

$$= 2\left(\log x^2\right)\left(\frac{1}{x^2}\cdot 2x\right)$$

$$= \frac{2}{x}\log x^2$$

ii) Let $y = \log\left(\frac{x}{x+1}\right)$

then

$$\frac{dy}{dx} = \frac{1}{\left(\frac{x}{x+1}\right)}\frac{d}{dx}\left(\frac{x}{x+1}\right)$$

$$= \frac{1}{\left(\frac{x}{x+1}\right)}\left[\frac{(x+1)1 - x\cdot 1}{(x+1)^2}\right]$$

$$= \left(\frac{x+1}{x}\right)\left[\frac{1}{(x+1)^2}\right] = \frac{1}{x(x+1)}$$

Special Case:

If a = e, then $y = \log_e^u$ and

$$\frac{dy}{dx} = \frac{1}{u}\frac{du}{dx}$$

Exponential Functions:

a) Let $y = a^u$

where $u = f(x)$ is a differentiable function of x, then

$$\frac{dy}{dx} = a^u \log a \frac{d}{dx}(u)$$

Special Case:

i) Let $y = 2^{x^2+1}$

then

$$\frac{dy}{dx} = 2^{x^2+1} \cdot \log_e^2 \cdot \frac{d}{dx}\left(x^2+1\right)$$

$$= 2^{x^2+1} \cdot \log_e^2 \left(2x + 0\right)$$

$$= 2x \cdot 2^{x^2+1} \log_e^2$$

ii) $y = \frac{e^x}{x}$

then

$$\frac{dy}{dx} = \frac{x \cdot \frac{d}{dx}\left(e^x\right) - e^x \frac{d}{dx}(x)}{x^2}$$

$$= \frac{x \cdot e^x - e^x \cdot 1}{x^2} = \frac{e^x (x-1)}{x^2}$$

b) Let $y = u^v$ or $\log_e y = v \log_e u$

where $u = f(x)$ and $v = g(x)$ are differentiable functions of x, then

$$\frac{d}{dx}\left(\log_e y\right) = \frac{d}{dx}\left(v \cdot \log_e u\right) = \frac{1}{y} \cdot \frac{dy}{dx} = v \cdot \frac{d}{dx}\left(\log_e u\right) + \log_e u \frac{d}{dx}(v)$$

$$= v\left(\frac{1}{u}\cdot\frac{du}{dx}\right) + \log_e u\left(\frac{dv}{dx}\right)$$

or $$\frac{dy}{dx} = y\left[\frac{v}{u}\cdot\frac{du}{dx} + \log_e u\,\frac{dv}{dx}\right]$$

$$= u^v\left[\frac{v}{u}\,\frac{du}{dx} + \log_e u\,\frac{dv}{dx}\right]$$

$$= v\cdot u^{v-1}\,\frac{du}{dx} + u^v\,\log_e u\,\frac{dv}{dx}$$

If may be noted that rule (a) is a special case of rule (b) where $u = a$.

Example 14

Let $y = x^{x^2+1}$, then taking log on both sides, we have

$$\log_e y\,(x^2+1)\log_e x$$

Differentiating both sides with respect to x, we have

$$\frac{1}{y}\,\frac{dy}{dx} = (2x+0)\log_e x + (x^2+1)\left(\frac{1}{x}\right)$$

$$= 2x\,\log_e x + x + \frac{1}{x}$$

$$= x(\,2\log_e x + 1) + \frac{1}{x}$$

3.6 Applications of the Derivative:

Basics, variation of one quantity y with respect to another quantity x usually and in terms of two concepts.

Average Concept and Marginal Concept

The average concept expresses the variation of y over a whole range of values of x. It is usually measured from zero to a certain selected value, say from 5 to 10. Whereas marginal concept concerns with the instantaneous rate of change in the dependent variable y for every small variation of x from a given value of x. Therefore a marginal concept is precise only when variation

in x are made smaller and smaller i.e. considering limiting value only. Hence $\underset{\Delta x \to 0}{Lt}\left(\frac{\Delta y}{\Delta x}\right)$ is interpreted as the marginal value of y.

Few applications of the derivative are discussed below:

1. Average and Marginal Cost:

Suppose the total cost y of producing and marketing x units of an item is represented by the function $u = f(x)$. Then the average cost which represents the cost per unit is given by

$$\text{Average cost } (AC) = \frac{y}{x} \text{ or } \frac{f(x)}{x}$$

Now, if the output is increased from x to $x + \Delta x$, and corresponding total cost becomes $y + \Delta y$, then the average increase in cost per unit output is given by the ratio $\frac{\Delta y}{\Delta x}$ and the marginal cost is defined as:

$$\text{Marginal Cost } (MC) = \underset{\Delta x \to 0}{Lt}\left(\frac{\Delta y}{\Delta x}\right) = \frac{dy}{dx}$$

That is, marginal cost is the first derivative of the total cost y with respect to output x and is the rate of increase in total cost with increase in output.

Example 5

The total cost $C(x)$ associated with producing and marketing x units of an item is given by $C(x) = 0 \cdot 005\, x^3 - 0 \cdot 02 x^2 - 30x + 3000$

find i) Total cost when output is 4 units

ii) Average Cost of output of 10 units

iii) Marginal Cost when output is 3 units

Solution:

i) Given that

$$C(x) = 0 \cdot 005 x^3 - 0 \cdot 02 x^2 - 30x + 3000$$

For x = units, the total cost $C(x)$ becomes

$$C(x) = 0\cdot005(4)^3 - 0\cdot02(4)^2 - 30(4) + 3000$$

$$= 0\cdot32 - 0\cdot32 - 120 + 3000$$

$$= \text{Rs. } 2880$$

ii) Average Cost $(AC) = \dfrac{C(x)}{x}$

$$= \frac{0\cdot005x^3 - 0\cdot002x^2 - 30x + 3000}{x}$$

$$= 0\cdot005x^2 - 0\cdot02(x) - 30 + \frac{3000}{x}$$

Average cost at x = 10 units becomes

$$AC = 0\cdot005\,(10)^2 - 0\cdot02(10) - 30 + \frac{3000}{10}$$

$$= 0\cdot5 - 0\cdot2 - 30 + 300$$

$$= \text{Rs. } 2703$$

iii) Marginal cost at x is given by $\dfrac{dc}{dx}$

Therefore differentiating both sides of $C(x)$ with respect to x, we have

$$\frac{d}{dx}\left(C(x)\right) = 0\cdot005\times3x^2 - 0\cdot02\times2x - 30$$

Marginal cost at x = 3 becomes

$$\frac{dC}{dx_{x=3}} = 0\cdot015(3)^2 - 0\cdot04(3) - 30$$

$$= 0\cdot135 - 0\cdot120 - 30$$

$$= \text{Rs. } 30\cdot015$$

2. Total revenue, Marginal revenue and Average revenue:

Let p be the price per unit and q is the number of units of an item sold. Then the total revenue (R) is given by

$$R = p \cdot q$$

The demand function is $p = f(q)$, therefore R becomes

$$R = q \cdot f(q)$$

Now average revenue (A R) or revenue per unit which represents the price per unit is given by

$$AR = \frac{R}{q} = \frac{p \cdot q}{q} = p\,(\text{price})$$

This shows that the average revenue and price are identical.

Since total revenue is given by $R = p \cdot q$, therefore marginal revenue (M R) is defined as:

$$MR = \frac{dR}{dq} = p + q \cdot \frac{dp}{dq}$$

$$= p\left(1 + \frac{q}{p} \cdot \frac{dp}{dq}\right)$$

Example 16

The demand for a certain product is represented by the equation:

$$p = 20 + 5q - q^2$$

where q is the number of units demanded and p is the price per unit.

Find marginal revenue function. What is the marginal revenue at $q = 2$?

Solution:

The total revenue is given by

Revenue, $R = (\text{demand})\ (\text{price})$

$$= q\left(20 + 5q - q^2\right)$$

$$= 20\,q + 5q^2 - q^3$$

Marginal revenue $(MR) = \frac{d}{dq}\left(20q + 5q^2 - q^3\right)$

$$= 20 + 10q - 3q^2$$

The marginal revenue (MR) at $q = 2$ is given by

$$MR = \frac{dR}{dq} = 20 + 10q - 3q^2$$

$$= 20 + 10(2) - 3(2)^2$$

$$= 20 + 20 - 12$$

$$= 28$$

Hence, the marginal revenue when two units are demand is Rs. 28.

3. Elastisity:

The elasticity of a function $y = f(x)$ at a point x is defined as the ratio of the rate of proportional change in y per unit proportional change in x. That is,

$$\frac{Ey}{Ex} = \frac{dy/y}{dx/x} = \frac{x}{y} \cdot \frac{dy}{dx}$$

The elasticity of a function is independent of the units in which the variables are measured because its definition is in terms of proportional changes. Notations usually used to denote elasticity are: e_y , or η_y or ε_{yx}

The above definition can also be expressed as:

$$e_y = \frac{dy/y}{dx/x} = \frac{dy/dx}{y/x} = \frac{\text{Marginal Function}}{\text{Average Function}}$$

The crucial value of $e_y = 1$. However the sign of e_y depends upon the sign of $\frac{dy}{dx}$. It may be positive, negative or zero. A part from the sign, we are also concerned about the absolute value $|e_y|$ of e_y.

a) Price Elasticity of Supply:

Let g be the supply and p be the price and the function is expressed as

$$q = f(p)$$

Then the formula for elasticity of supply is same as that of e_y that is

$$e_s = \frac{p}{q} \cdot \frac{dq}{dp}$$

The sign of e_s will also be positive because slope of supply curve is positive.

b) Price elasticity of demand:

The price elasticity of demand at price 'b' is defined as:

$$e_d = -\frac{p}{q} \underset{\Delta p \to 0}{Lt} \left\{ \frac{\Delta q}{\Delta p} \right\}$$

$$= -\frac{p}{q} \cdot \frac{dq}{dp} = -\frac{p}{q} \frac{1}{dp/dq}$$

The sign of e_d is negative, because in general the slope of demand $\frac{dq}{dp}$ is negative.

c) Marginal revenue and elasticity of demand:

You know that the total revenue (R) is given by

$$R = p \cdot q$$

where p is the price and q is the quantity sold.

Also the average revenue (A R) and marginal revenue (M R) are defined as:

Average revenue $(AR) = \frac{R}{q} = \frac{p \cdot q}{q} = p$

Marginal revenue $(MR) = \frac{dR}{dq} = \frac{d}{dq}(p \cdot q)$

$$= p \cdot 1 + q \cdot \frac{dp}{dq}$$

$$= p\left(1 + \frac{q}{p} \cdot \frac{dp}{dq}\right)$$

$$= p\left(1 + \frac{q}{p} \cdot \frac{dp}{dq}\right)$$

$$= p\left(1 + \frac{1}{|e_d|}\right)$$

$$= AR\left(1 + \frac{1}{|e_d|}\right)$$

since $|e_d| = \frac{p}{q} \cdot \frac{dq}{dp}$

From this definition of M R, it follows that

i) If $|e_d| = 1$ then $AR = 0$ and hence $MR = 0$, i.e. total revenue remains constant with a fall in price.

ii) If $|e_d| > 1$ then, AR = 0 and hence $MR > 0$ i.e., total revenue increases with an increase in demand or with a fall in price.

iii) If $|e_d| > 1$, then AR > 0 and hence MR < 0 i.e. total revenue decreases with an increse in demand or with a fall in price.

Example 17

Suppose the price p and quantity q of a commodity are related by the equation

$$q = 30 - 4p - p^2$$

Find i) Elasticity of demand e_q defined as $= -\frac{dq/q}{dp/p}$ at $p = 2$ and

ii) Marginal revenue (M R) defined as $= \frac{dR}{dq}$, where $R = p \cdot q$

Solution:

i) Elasticity of demand e_q is defined as:

$$e_q = -\frac{dq/q}{dp/p} = -\frac{p}{q}\cdot\frac{dq}{dp}$$

$$= -\frac{p}{q}\cdot\frac{d}{dp}\left(30 - 4p - p^2\right)$$

$$= -\frac{p}{\left(30 - 4p - p^2\right)}\cdot(-4 - 2p)$$

$$= \frac{4p + 2p^2}{30 - 4p - p^2}$$

Thus $e_q / p = 2 = \frac{4\times 2 + 2(2)^2}{30 - 4\times 2 - (2)^2} = \frac{16}{18} = \frac{8}{9}$

ii) Marginal revenue (M R) is defined as:

$$MR = \frac{dR}{dq} = \frac{dR}{dp}\cdot\frac{dp}{dq}$$

$$= \frac{dR}{dp}\cdot\frac{1}{dq/dp}$$

$$= \frac{d(p\cdot q)}{dp}\cdot\frac{1}{dq/dp}$$

$$= \frac{d}{dp}\left[p\cdot\left(30 - 4p - p^2\right)\right]\cdot\frac{1}{\frac{d}{dp}\left(30 - 4p - p^2\right)}$$

$$= \left[30 - 8p - 3p^2\right]\cdot\frac{1}{(0 - 4 - 2p)}$$

$$= \frac{30 - 8p - 3p^2}{-4 - 2p}$$

3.7 Concept of Maxima and Minima with Managerial Applications:

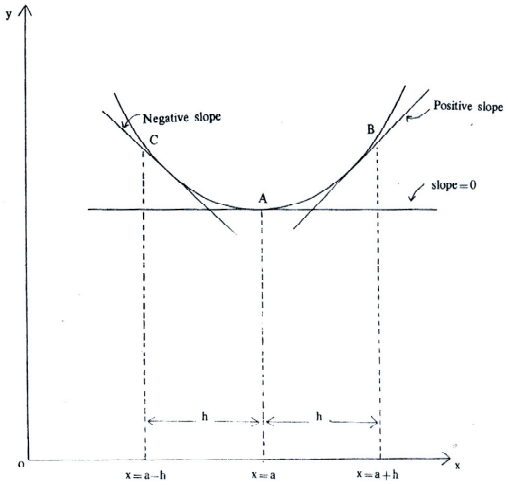
The objective of studying differential calculus is to be able to solve optimisation problem in which the decision - maker seeks either to maximise or minimise the given objective function (or goal) under certain limitations (or constraints) on available resources. In this unit unconstrained optimisation problems involving single independent variable are presented.

Conditions for maxima and minima:

The necessary condition:

Consider the function $y = f(x)$ given in figure V(a). At the point A which is the lowest point of the curve, the tangent is neither inclined to the right nor to the left. But the tangent is parallel to the x - axis and its slope is zero, i.e. $m = \tan\theta = 0$ because the slope of a horizontal line is equal to zero. The slope is measured by the first derivative, therefore the derivative at point A must be equal to zero.

Figure V(a)



From figure V(a), it is clear that the value of the function $y = f(x)$ decreases as x increases upto A, i.e. increases from $x = a - h$ to $x = a$ and then increases as x increases upto B, i.e., increases from $x = a$ to $x = a + h$. Thus $\frac{dy}{dx}$ will be negative upto A, becomes zero at A and will be positive after crossing A. This shows that if the function $f(x)$ is minimum at point A, then its first derivative at point A is equal to zero, but the converse is not true. That is,

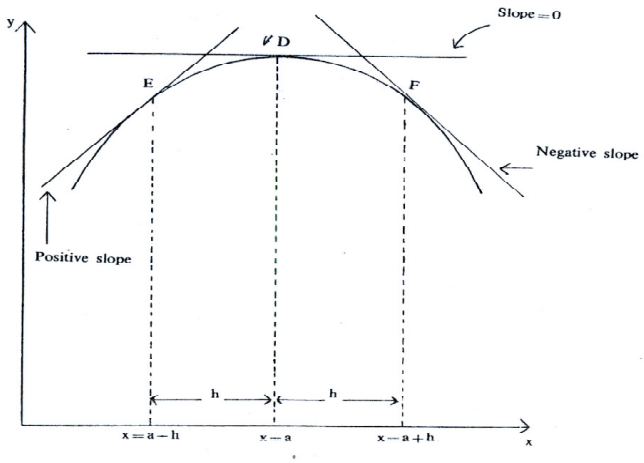
$$\frac{dy}{dx} = 0 \text{ at point A}$$

This minimum value of the function $y = f(x)$ at x = a is called local (or relative) minimum value because the value $y = f(a)$ is lessthan any other value of $f(x)$ for x in an interval around a. The word local (or relative) has been used to define this minimum value of $f(x)$ because it has been obtained with reference to a small interval containing the point.

From figure $V(b)$, it is clear that the function $f(x)$ reaches a maximum at the point D. It can also be verified that function $f(x)$ increases as x increases upto D, and the decreases after crossing D. Thus $\frac{dy}{dx}$ will be positive up to D become zero at D and will be negative after crossing D. This also shows that if the function $f(x)$ is maximum at point D, then its first derivative at that point is zero but converse is not true. That is,

$\frac{dy}{dx} = 0$ at the point D.

Figure V(b)



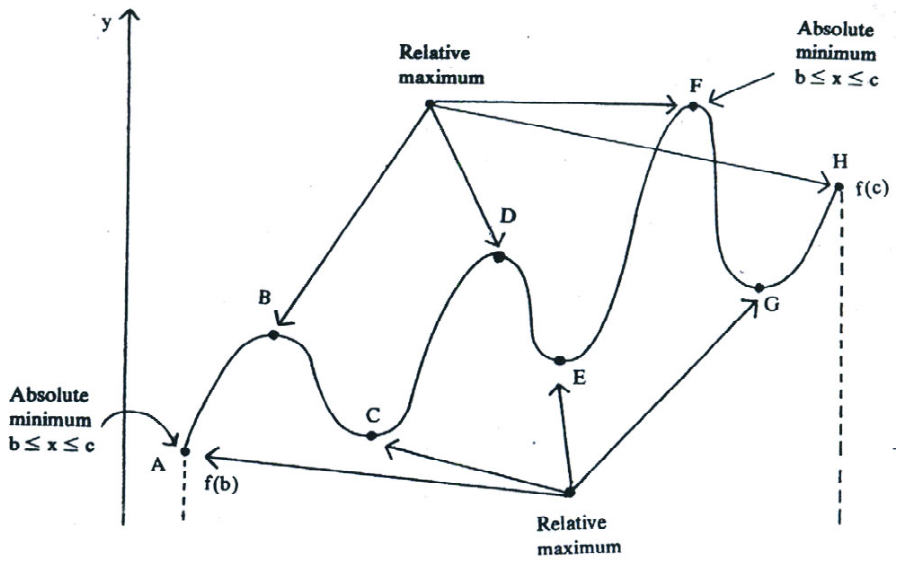
This maximum value of the function $f(x)$ at $x = a$ is called a local (or relative) maximum because $y = f(a)$ is greater than any value of $f(x)$ for x in an interval around a.

Hence, the condition that the first derivative is equal to zero at the maxima (plural or maximum) or minima (plural of minimum) is a necessary condition but not a sufficient one because it does not help us to locate absolute (or global) maximum or minimum. By absolute maximum (or minimum) we mean maximum (or minimum) value of $f(x)$ amongst all given maximum (or minimum) values in the specified interval for x.

The sufficient condition:

The function $y = f(x)$ whose graph is given in figure $V(c)$ has four maxima and four minima in the entire range from $x = b$ to $x = c$.

Figure V(c)



The slope of the curve at the points A to H is zero. Such points for which $\frac{dy}{dx} = 0$ are called the stationary points or extreme points or critical points of the function $y = f(x)$. The function has maxima at the points B, D, F, H and minima at the points A, C, E, G. The absoulte (or global) maximum occurs at the point F and absolute (or global) minimum occur at the point A. However these values of a function in an interval may occur at an end point of the interval rather than at a relative minimum or maximum value.

Let us now, examine the sign of $\frac{dy}{dx}$ in the neighbourhood the points of maxima and minima.

i) The sign of $\frac{dy}{dx}$ changes from positive to negative as x passes through the points of maxima. If you consider $\frac{dy}{dx}$ as a function of x, then you will find that it is a decreasing

function as it passes through the points of maxima, i.e. rate of change of $\frac{dy}{dx}$ is negative. In other words

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) < 0 \qquad \text{or} \qquad \frac{d^2y}{dx^2} < 0$$

at a point where $f(x)$ is a maximum.

(ii) The sign of $\frac{dy}{dx}$ changes from negative to positive as x passes through the points of minima, and hence $\frac{dy}{dx}$ is an increasing function, i.e. rate of change of $\frac{dy}{dx}$ is positive, In other words

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) > 0 \qquad \text{or} \qquad \frac{d^2y}{dx^2} > 0$$

at the point where $f(x)$ is a minimum.

However, at certain points, you may find $\frac{d^2y}{dx^2} = 0$.

Such points are called point of inflexion. In such cases, the points are neither maximum nor minimum.

Summary of the results:

	Maximum	**Minimum**
Necessary condition	$\frac{dy}{dx} = 0$	$\frac{dy}{dx} = 0$
Sufficient Condition	$\frac{dy}{dx} = 0 ; \frac{d^2y}{dx^2} < 0 \cdot$	$\frac{dy}{dx} = 0 \ ; \frac{d^2y}{dx^2} > 0$.

Summary of the procedure:

1. Take the first derivative of the given function.
2. Set the derivative equal to zero and solve the value of the independent variable at which the function is either maximum or minimum.
3. Take the second derivative of the function.
4. Evaluate the second derivative at the points obtained in step 2.
5. If second derivative is positive, then $f(x)$ is minimum at the given point. Otherwise maximum.

Example 18

Suppose a manufacturer can sell x items per week at a price, $P = 20 - 0 \cdot 001x$ rupees each when it costs, $y = 5x + 2000$ rupees to produce x items. Determine the number of items he should produce per week for maximum profit.

Solution:

The cost of producting x items $= 5x + 2000$

The price of one item $= 20 - 0 \cdot 0001 . x$

Threrefore selling price of x items $= x(20 - 0 \cdot 001x)$

Let Z be the profit function. Then it is given by

Z = Revenue - Cost

$$= \left(20x - 0 \cdot 0011x^2\right) - (5x + 2000)$$

$$= -0 \cdot 001\, x^2 + 15x - 2000 \quad \text{and} \quad \frac{dz}{dx} = -0 \cdot 002x + 15$$

For maximum profit,

$$\frac{dZ}{dx} = -0 \cdot 002x + 15 = 0 \quad \text{or} \quad 0 \cdot 002x = 15 \quad \text{or} \quad x = \frac{15}{0 \cdot 002} = 7500$$

Now $\dfrac{d^2 Z}{dx^2} = \dfrac{d}{dx}\left(\dfrac{dZ}{dx}\right)$

$$= \frac{d}{dx}(-0 \cdot 002x + 15) = -0 \cdot 002\,(-ve)$$

So profit is maximum when 7500 items are produced and sold.

3.8 Solved Problems:

1. If $f(x) = 3x^2 + 2x - 4$, find $\lim_{x \to 0} f(x)$

Given $f(x) = 3x^2 + 2x - 4$

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \left(3x^2 + 2x - 4\right)$$

$$= 3(0)^2 + 2(0) - 4$$

$$\lim_{x \to 0} f(x) = -4$$

2. If $f(x) = 2 - \frac{1}{x^2} + \frac{4}{x}$ find $\lim_{x \to \infty} f(x)$

Given $f(x) = 2 - \frac{1}{x^2} + \frac{4}{x}$

$$\lim_{x \to \infty} f(x) = \lim_{x \to \infty} \left(2 - \frac{1}{x^2} + \frac{4}{x}\right)$$

$$= 2 - \frac{1}{\infty} + \frac{4}{\infty}$$

$$= 2$$

3. Find $\lim_{x \to 2} \left(x^3 - 4x + 8\right)$

Given $f(x) = x^3 - 4x + 8$

$$\lim_{x \to 2} f(x) = \lim_{x \to 2} \left(x^3 - 4x + 8\right)$$

$$= 2^3 - 4(2) + 8$$

$$= 8$$

4. Find $\lim_{x \to 2} \frac{x^2 + x + 2}{x^2 + 1}$

Given $f(x) = \frac{x^2 + x + 2}{x^2 + 1}$

$$\lim_{x \to 2} f(x) = \lim_{x \to 2} \frac{x^2 + x + 2}{x^2 + 1}$$

$$= \frac{4 + 2 + 2}{4 + 1}$$

$$\lim_{x \to 2} f(x) = 8/5$$

5. If $f(x) = \frac{x^4 - 16}{x^2 - 4}$, find $\lim_{x \to 2} f(x)$

Given $f(x) = \frac{x^4 - 16}{x^2 - 4}$

$$\lim_{x \to 2} f(x) = x \xrightarrow{\text{Lim}} 2 \frac{(x^4 - 16)}{x^2 - 4}$$

$$= \lim_{x \to 2} \frac{(x^2 + 4)\cancel{(x^2 - 4)}}{\cancel{(x^2 - 4)}}$$

$$= \lim_{x \to 2} x^2 + 4$$

$$= 4 + 4 = 8$$

6. Determine the derivative of following functions:

1. $f(x) = x^5$

Given $y = x^5$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^5\right)$$

$$= 5x^4$$

2. $f(x) = 3x^2$

Given $y = 3x^2$

$$\frac{dy}{dx} = \frac{d}{dx}\left(3x^2\right)$$

$$= 3\frac{d}{dx}\left(x^2\right)$$

$$= 3 \times 2 \times x$$

$$= 6x$$

3. $f(x) = x^{-3}$

Given $y = x^{-3}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^{-3}\right)$$

$$= -3x^{-2}$$

4. $f(x) = x^{-1}$

Given $y = x^{-1}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^{-1}\right)$$

$$= -1$$

5. $f(x) = x^{-1/4}$

Given $y = x^{-1/4}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^{-1/4}\right)$$

$$= -1/4 \cdot x^{-1/4 + 1}$$

$$= -1/4 \; x^{3/4}$$

6. $f(x) = x^{1/2}$

Given $y = x^{1/2}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^{1/2}\right)$$

$$= 1/2 \; x^{1/2 - 1}$$

$$= 1/2 \; x^{-1/2}$$

7. Calculate the derivatives of following functions:

1. $f(x) = 4x^2 + 2x^{-1/2} + 3$

Given $y = 4x^2 + 2x^{-1/2} + 3$

$$\frac{dy}{dx} = \frac{d}{dx}\left(4x^2 + 2x^{-1/2} + 3\right)$$

$$= \frac{d}{dx}\left(4x^2\right) + \frac{d}{dx}\left(2x^{-1/2}\right) + \frac{d}{dx}(3)$$

$$= 4 \times 2x + \cancel{2} \times \left(\frac{-1}{\cancel{2}}\right) x^{-1/2 + 1} + 0$$

$$= 8x - x^{1/2}$$

8. $f(x) = \frac{x^5}{5} + \frac{x^{-2}}{3}$

Given $y = \frac{x^5}{5} + \frac{x^{-2}}{3}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(\frac{x^5}{5} + \frac{x^{-2}}{3}\right)$$

$$= \frac{1}{\cancel{5}} \times \cancel{5}\, x^4 + \frac{1}{3}(-2)\, x^{-3}$$

$$= x^4 - \frac{2}{3}\, x^{-3}$$

9. $f(x) = x^{-4} + x^{-3} + x^{-2} + x^{-1}$

$$y = x^{-4} + x^{-3} + x^{-2} + x^{-1}$$

$$\frac{dy}{dx} = \frac{d}{dx}\left(x^{-4} + x^{-3} + x^{-2} + x^{-1}\right)$$

$$= -4\, x^{-5} - 3x^{-4} - 2x^{-3} - x^{-2}$$

10. $f(x) = 20x^5 - 10^4 + 5x^6 + 15\, x^{-17}$

$$y = 20x^5 - 10x^4 + 5x^6 + 15x^{-17}$$

$$\frac{dy}{dx} = \frac{d}{dx}\left(20x^5 - 10x^4 + 5x^6 + 15x^{-17}\right)$$

$$= 20 \times 5x^4 - 10 \times 4x^3 + 5 \times 6x^5 - 15 \times 17\, x^{-18}$$

$$= 100\, x^4 - 40\, x^3 + 30\, x^5 - 225\, x^{-18}$$

11. $f(x) = a\sqrt{x}$

Given $y = a\, x^{1/2}$

$$\frac{dy}{dx} = \frac{d}{dx}\left(a\, x^{1/2}\right)$$

$$= a \cdot \frac{1}{2}\, x^{1/2 - 1}$$

$$= \frac{a}{2}\, x^{-1/2}$$

12. Find the derivatives of the following functions:

$$f(x) = x^4\left(x^3 + 9\right)$$

Solution:

Given $y = x^4\left(x^3 + 9\right)$

$$\frac{dy}{dx} = x^4 \frac{d}{dx}\left(x^3 + 9\right) + \left(x^3 + 9\right)\frac{d}{dx}\left(x^4\right)$$

$$= x^4 \times 3x^2 + \left(x^3 + 9\right)4x^3$$

13. Given the demand function $D = 15 - 0 \cdot 5\, p,$ where D is the quantity demanded and P is the price, determine the price elasticity at $P = 12$.

Solution:

The price elasticity is defined as follows:

$$e = -\frac{\%\ \text{change in demand}}{\%\ \text{change in price}}$$

$$= -\frac{\frac{dD}{D} \times 100}{\frac{dp}{P} \times 100} = -\frac{P\, d\, D}{D\, d\, P}$$

Since $D = 15 - 0 \cdot 5\, P \qquad \frac{dD}{dP} = -0 \cdot 5$

Hence $e = -\frac{P}{D}(-0\cdot5) = -\frac{P}{15-0\cdot5P}(-0\cdot5)$

Therefore, the elasticity at $P = 12$ is

$$e = \frac{-12}{15-6}(-0\cdot5) = \frac{6}{9} = 0\cdot66$$

14. A store has a square display space that is estimated to effect sales in proportion to area. The effect for a constant variation of K is given by $S = K\,x^2$ where S is the amount of sales and x is the side of the square. Calculate the rate of sales and x is the side of the square. Calculate the rate of change of sales with respect to dimension of the square.

Solution:

Rate of change of sales with respect to dimension of the square is given by $\frac{dS}{dx}$ Now

$$S = K\,x^2$$

$$\Rightarrow \frac{dS}{dx} = \frac{d}{dx}\left(K\,x^2\right) = 2k\,x$$

15. Let A be the amount of advertisement and S be the sales. The relationship between S and A for a firm is given by

$$S = \left(3A^2 - 2A\right)^{1/3}$$

Find the effect of increase in advertisement on sales.

Solution: The rate of change of sales with respect to advertisement is given by

$$\frac{dS}{dA} = \frac{d}{dA}\left[\left(3A^2 - 2A\right)^{1/3}\right]$$

$$= \frac{1}{3}\left(3A^2 - 2A\right)^{1/3-1}\frac{d}{dA}\left(3A^2 - 2A\right)$$

$$= \frac{1}{3}\left(3A^2 - 2A\right)^{-2/3}\left(6A - 2\right)$$

16. A company has examined its cost structure and revenue structure and has determined that C the total cost, R total revenue and x the number of units produced are related as

$C = 100 + 0 \cdot 015\ x^2$ and R = 3x

Find the production rate x that will maximize profits of the company. Find that profit. Find also the profit when x = 120.

Solution:

Let P denote the profit of the company, then

P = Revenue - Cost = R - C

$$= 3x - \left(100 + 0 \cdot 015\ x^2\right) = 3x - 100 - \frac{15}{1000}\ x^2$$

$$\frac{dP}{dx} = 3 - \frac{30x}{1000}$$

For maximum or minimum values $\frac{dP}{dx} = 0$

$$3 - \frac{30x}{1000} = 0 \Rightarrow \frac{3x}{100} = 3 \Rightarrow x = 100 \text{ units}$$

also $\frac{d^2P}{dx^2} = -\frac{3}{100} < 0$ when x = 100

Profits is maximum when x = 100

Maximum profit $= 3 \times 100 - 0 \cdot 015\ (100)^2 - 100$

$$= 300 - 150 - 100 = 50 \text{ rupees}$$

Profit when x = 120 is

$$P = 3 \times 120 - 100 - 0 \cdot 015\ (120)^2$$

$$= 360 - 100 - 216 = 44 \text{ rupees}$$

17. The demand function for a particular commodity is $y = 15\ e^{-x/3}$ for $0 \le x \le 8$, where y is the price per unit and x is the number of units demanded. Determine the price and the quantity for which the revenue is maximum. (Hint : Revenue ; R = y.x)

Solution:

Demand $y = 15 \cdot e^{-x/3}$ for $0 \le x \le 8$

Revenue $R = x\,y = 15\,x\,e^{-x/3}$

For maximisation of revenue, we have

$$\frac{dR}{dx} = 15\,e^{-x/3} + \left(-\frac{15}{3}\right) x\,e^{-x/3}$$

$$= 15\,e^{-x/3} - 5\,x\,e^{-x/3}$$

$$\frac{dR}{dx} = 0 \Rightarrow 3e^{-x/3} - x\,e^{-x/3} = 0$$

Either $x = 3$ or $e^{-x/3} = 0 \Rightarrow x \to \infty$ (absurd)

Hence x = 3.

Also $\frac{d^2R}{dx^2} = -\frac{3}{3}e^{-x/3} - e^{-x/3} + \frac{x}{3}e^{-x/3}$

$$= -2e^{-x/3} + \frac{x}{3}e^{-x/3}$$

$$= -2\,e^{-1} + e^{-1} = -e^{-1} = \frac{-1}{e} < 0$$

Hence the maximum profit is yielded by substituting x = 3 in the revenue equation.

$$R = 15\,x\,e^{-x/3} = 45\,e^{-1} = \frac{45}{2 \cdot 72} = 16 \cdot 54$$

18. If the demand law is $p = \frac{10}{(x+1)^2}$, find the elasticity of demand in terms of x.

Solution:

The elasticity of demand is defined as

$$\eta_d = -\frac{p}{x} \cdot \frac{dx}{dp}$$

Given $p = \frac{10}{(x+1)^2} = 10(x+1)^{-2}$

$$\frac{dP}{dx} = 10(-2)(x+1)^{-3} = \frac{-20}{(x+1)^3}$$

Substituting the values we get

$$\eta_d = -\frac{10}{(x+1)^2} \times \frac{1}{x} \times \left\{ -\frac{(x+1)^3}{20} \right\} = \frac{x+1}{2x}$$

19. Find the elasticity of demand for the demand function $x = \frac{27}{p^3}$, where x is the demand of a good at a price p.

Solution:

Marginal quantity demanded is

$$\frac{dx}{dP} = -\frac{81}{p^4}$$

Average quantity demanded is

$$\frac{x}{p} = \frac{27}{p^3} \cdot \frac{1}{p} = \frac{27}{p^4}$$

Hence elasticity of demand $\eta_d = \left| \frac{dx/dp}{x/p} \right|$

$$\eta_d = \left| \frac{-81}{p^4} \times \frac{p^4}{27} \right| = 3$$

20. The average cost function (A C) for a commodity is given by

$$AC = x + 5 + \frac{36}{x}$$

in terms of the output x. Find the outputs for which A C is increasing and the outputs for which AC is decreasing, with increasing output.

Also, find the total cost C and the marginal cost (M C) as function of x.

Solution: Slope of $\mathrm{AC} = \frac{d}{dx}\left(x + 5 + \frac{36}{x}\right) = 1 - \frac{36}{x^2}$

AC is increasing if $1 - \frac{36}{x^2} > 0$, i.e., if $x^2 > 36$ or $x > 6$ and

decreasing if $1 - \frac{36}{x^2} < 0$, i.e., if $x < 6$.

Now $\mathrm{AC} = x + 5 + \frac{36}{x} = \frac{x^2 + 5x + 36}{x}$

Total Cost $(C) = x$, $\mathrm{AC} = x^2 + 5x + 36$

Marginal Cost $(MC) = \frac{dC}{dx} = 2x + 5$

21. The total cost function of a firm is given by

$$C = 0 \cdot 04\, q^3 - 0 \cdot 9\, q^2 + 10q + 10$$

Find (a) average cost (AC)

(b) Marginal Cost (M C)

(c) Slope of AC

(d) Slope of MC

(e) Value of q at which average variable cost is minimum.

Solution:

(a) $\mathrm{AC} = \frac{C}{q} = 0 \cdot 04\ q^2 - 0 \cdot 9\, q + 10 + \frac{10}{q}$

(b) Marginal Cost $(MC) = \frac{dC}{dq} = 0 \cdot 12\, q^2 - 1 \cdot 8q + 10$

(c) Slope of $AC = \frac{d}{dq}\left(\frac{C}{q}\right)$

$$= \frac{d}{dq}\left(0 \cdot 04\, q^2 - 0 \cdot 9\, q + 10 + \frac{10}{q}\right)$$

$$= \left(0 \cdot 08\, q - 0 \cdot 9 - \frac{10}{q^2}\right)$$

$$= \frac{1}{q}\left(0 \cdot 08\, q^2 - 0 \cdot 9q - \frac{10}{q}\right)$$

$$= \frac{1}{q}\left[\left(0 \cdot 12\, q^2 - 1 \cdot 8\, q + 10\right) - \left(0 \cdot 04q^2 - 0 \cdot 9q + 10 + \frac{10}{q}\right)\right]$$

$$= \frac{1}{q}\left[MC - AC\right]$$

(d) Slope of $MC = \frac{d}{dq}\left(\frac{dMC}{dq}\right) = 0 \cdot 24\, q - 1 \cdot 8$

(e) when AVC is minimum, the slope of AVC curve is zero, i.e.,

$$\frac{d}{dq}\left(AVC\right) = 0 \quad \text{or} \quad \frac{d}{dq}\left(0 \cdot 04q^2 - 0 \cdot 9q + 10\right) = 0$$

$$\Rightarrow \quad 0 \cdot 08\, q - 0 \cdot 9 = 0 \quad \text{or} \quad q = \frac{0 \cdot 9}{0 \cdot 08} = 11 \cdot 25$$

22. Let the cost function of a firm be given by the following equation:

$$C = 300x - 10x^2 + \frac{1}{3}x^3, \quad \text{where C stands for cost and x for output.}$$

Calculate (i) Output at which marginal cost is minimum.

(ii) Output, at which average cost is minimum

(iii) Output, at which average cost is equal to marginal cost.

Solution:

(i) $C = 300x - 10x^2 + \frac{1}{3}x^3$

$$MC = \frac{dC}{dx} = 300 - 10(2x) + \frac{1}{3} \cdot 3x^2$$

$$= 300 - 20x + x^2$$

Differentiating w.r.t. x and equating to zero, we have

$$\frac{d(MC)}{dx} = -20 + 2x = 0$$

or x = 10 is the necessary condition for marginal cost minimisation.

To get the sufficient condition, we have

$\frac{d^2(MC)}{dx^2} = 2$, a positive quantity which means that marginal cost is minimum at x = 10.

(ii) Average Cost $(AC) = \frac{C}{x} = \frac{300x - 10x^2 + \frac{1}{3}x^3}{x}$

$$= 300 - 10x + \frac{1}{3}x^2$$

Now to find output at which average cost is minimum, we have to differentiate the AC and equating it to zero.

$$\therefore \frac{d(AC)}{dx} = 0 - 10 + \frac{1}{3} \cdot 2x = 0 \text{ or } x = 15$$

Also $\frac{d^2(AC)}{dx^2} = \frac{d}{dx}\left(-10 + \frac{2}{3}x\right) = \frac{2}{3}$, a positive quantity.

$\therefore$ Second condition is also satisfied. Hence the output at which AC is minimum is given by x = 15.

(iii) Now $A C = M C$

$$\Rightarrow 300 - 10x + \frac{1}{3}x^2 = 300 - 20x + x^2$$

$$\Rightarrow \frac{3x^2}{3} = 10x \quad \text{or} \quad x = 15$$

Hence for $x = 15$ average cost is equal to marginal cost.

23. The total variable cost of a monthly output x tons by a firm producting a variable metal is Rs. $\frac{1}{10}x^3 - 3x^2 + 5x$ and the fixed cost is Rs. 300 per month. Draw the average cost curve when cost includes (i) Variable cost only, (ii) all costs. Find the output for minimum average cost in each case.

Solution: We have

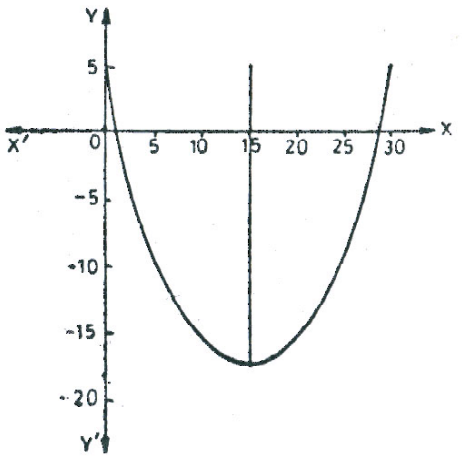
$$T C = \text{Total Cost} = \frac{1}{10}x^3 - 3x^2 + 5x + 300$$

and $$TVC = \text{Total Variable Cost} = \frac{1}{10}x^3 - 3x^2 + 5x$$

(i) When cost includes variable cost only:

$$A V C = \text{Average Variable Cost} = \frac{TVC}{x} = \frac{1}{10}x^2 - 3x + 5$$

It is a parabola with vertex at (15, - 17.5) and the axis of the parabola is x = 15. The graph of the curve is shown in the figure below:



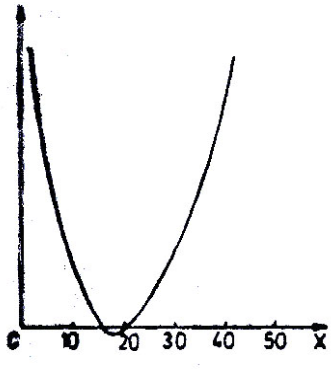
$$\frac{d(AVC)}{dx} = \frac{1}{5}x - 3$$

and $\frac{d^2(AVC)}{dx^2} = \frac{1}{5} > 0$

$\frac{d(AVC)}{dx} = 0 \Rightarrow \frac{1}{5}x - 3 = 0$ or x = 15.

Hence average cost is minimum when the output is 15 tons.

(ii) When cost includes all costs:



A C = Average Cost $= \frac{TC}{x}$

$$= \frac{1}{10}x^2 - 3x + \frac{300}{x}$$

The graph of the curve is shown in the adjoining figure.

$\frac{d(AC)}{dx} = \frac{1}{5}x - 3 - \frac{300}{x^2}$ and $\frac{d^2(AC)}{dx^2} = \frac{1}{5} + \frac{600}{x^3} > 0$

$\frac{d(AC)}{dx} = 0 \Rightarrow \frac{1}{5}x - 3 - \frac{300}{x^2} = 0$

Which gives $x = 19 \cdot 1$

Hence average cost is minimum when the output is 19.1 tons.

24. Find the profit maximising output given the following revenue and cost functions.

$$R(Q) = 1000\,Q - 2Q^2,\quad C(Q) = Q^3 - 59Q^2 + 1315Q + 2000$$

Solution:

We have $P = \text{Profit} = R(Q) - C(Q)$

$$= \left(1000\,Q - 2Q^2\right) - \left(Q^3 - 59Q^2 + 1351Q + 2000\right)$$

$$= -Q^3 + 57Q^2 - 315Q - 2000$$

First order condition:

$$\frac{dP}{dQ} = 0$$

$$\therefore \quad \frac{dP}{dQ} = -3Q^2 + 114Q - 315$$

$$\frac{dP}{dQ} = 0 \quad \Rightarrow -3Q^2 + 114Q - 315 = 0$$

$$\text{or} \quad Q^2 - 38Q + 105 = 0 \quad \text{or} \quad (Q-3)(Q-35) = 0$$

$$\therefore \quad Q = 3 \quad \text{or} \quad Q = 35$$

Second order condition:

$$\frac{d^2P}{dQ^2} < 0$$

$$\frac{d^2P}{dQ^2} = -6Q + 114$$

$$\left.\frac{d^2P}{dQ^2}\right|_{Q=3} = -18 + 114 = 96 > 0$$

$$\left.\frac{d^2P}{dQ^2}\right|_{Q=35} = -210 + 114 = -96 < 0$$

Hence the profit maximising output is given by Q = 35.

25. A radio manufacturer produces x sets per week at a total cost of Rs. $\left(x^2 + 78x + 2500\right)$. He is a monopolist and the demand function for his product is $x = \frac{600 - P}{8}$ when the price is Rs. p per set. Show that maximum net revenue (i.e., Profit) is obtained when 29 sets are produced per week. What is the monopoly price?

Solution: Total Cost $(C) = x^2 + 78x + 2500$

Marginal $(MC) = \frac{dC}{dx} = 2x + 78$

Demand function is $x = \frac{600 - p}{8}$

$\Rightarrow 8x = 600 - p$

$\Rightarrow p = 600 - 8x$

Now total revenue for x sets is

$R = p \times x = (600 - 8x)\ x = 600x - 8x^2$

Marginal revenue $(MR) = \frac{dR}{dx} = \frac{d}{dx}\left(600x - 8x^2\right) = 600 - 16x$ .....(**)

Net revenue will be maximum at the level of output, where $MR = MC$

$\therefore 2x + 78 = 600 - 16x$

$\Rightarrow 18x = 522$

$\Rightarrow x = \frac{522}{18} = 29$

Hence in order to maximise his profit, the manufacturer should manufacture 29 sets per week. Also the monopoly price is given by

$p = 600 - 8x = 600 - 8 \times 29 = \text{Rs. } 368$

Aliter, We know: Net revenue = Total revenue - Total cost

$\pi = px - C = x(600 - 8x) - \left(x^2 + 78x + 2500\right)$

For maxima and minima $\frac{dP}{dx} = 0 \quad \Rightarrow \; 600 - 16x - 2x - 78 = 0$ or x = 29.

(Remark: Also examine whether second order condition is satisfied at output level).

26. The total revenue function of a firm is given as $R = 21q - q^2$ and its total cost function as $C = \frac{1}{3}q^3 - 3q^2 - 7q + 16$, where q is the output. Find

(i) The output at which the total revenue is maximum, and

(ii) The output at which the total cost is minimum

Solution:

(i) $R = 21\ q - q^2$

Differentiating w.r.t. q and equating to zero, we have

$$\frac{dR}{dq} = 21 - 2q = 0$$

or $q = \frac{21}{2} = 10 \cdot 5$ is the necessary condition for revenue maximisation.

To get the sufficient condition, we have

$\frac{d^2R}{dq^2} = -2$, a negative quantity, which means the revenue is maximum at $p = 10 \cdot 5$

(ii) $C = \frac{1}{3}q^3 - 3q^2 - 7q + 16$

Differentiating w.r.t. q and equating to zero, we have

$$\frac{dC}{dq} = \frac{1}{3} \cdot 3q^2 - 3 \times 2q - 7 = 0$$

$$\Rightarrow \; q^2 - 6q - 7 = 0$$

$$\Rightarrow (q - 7)(q + 1) = 0$$

$\Rightarrow q = 7$ or $q = -1$ is the necessary condition for cost maximisation or minimisation. $q = -1$ is not admissible as output cannot be negative.

To get the sufficient condition, we have

$$\frac{d^2C}{dq^2} = 2q - 6$$

$\left[\frac{d^2C}{dq^2}\right]_{q=7} = 2 \times 7 - 6 = 8$ a positive quantity which means that cost is minimum at $q = 7$.

27. The unit demand function is $x = \frac{1}{3}(25 - 2p)$ where x is the number of units and p is the price. Let the average cost per unit be Rs. 40 Find

(a) The revenue function R in terms of price p,

(b) The cost function C,

(c) The profit function P,

(d) The price per unit that maximizes the profit function and

(e) the maximum profit.

Solution:

(a) $R(x) = xp = \frac{1}{3}(25 - 2p)\ p = \frac{1}{3}\left(25p - 2p^2\right)$

(b) $C(x) = 40x = 40 \cdot \frac{1}{3}(25 - 2p) = \frac{40}{3}(25 - 2p)$

(c) $P(x) = R(x) - C(x)$

$$= \frac{1}{3}\left(25p - 2p^2\right) - \frac{40}{3}(25 - 2p)$$

$$= \frac{25p}{3} - \frac{2p^2}{3} - \frac{1000}{3} + \frac{80p}{3}$$

$$= \frac{1}{3}\left[-2p^2 + 105p - 1000\right]$$

(d) The derivative of $P(x)$ is

$$P'(x) = \frac{1}{3}(-4p + 105)$$

Solving the equation $P'(x) = 0$ we find that

$$P = \frac{105}{4} = 26 \cdot 25$$

Using second derivative test, we have

$$P''(x) = -\frac{4}{3} < 0$$

$\therefore$ Maximum profit is found when $p = 26 \cdot 25$

(e) Maximum profit is

$$P(x) = \frac{1}{3}\left[-2\left(\frac{105}{4}\right)^2 + 105\left(\frac{105}{4}\right) - 1000\right] = 126 \cdot 04$$

28. The demand function faced by a firm is $p = 500 - 0 \cdot 2x$ and its cost function is $C = 25x + 10000$ (p = price, x = output and C = cost). Find the output at which the profits of the firm are maximum. Also find the price it will charge.

Solution: Revenue, $R(x) = p \cdot x = x(500 - 0 \cdot 2x) = 500x - 0 \cdot 2x^2$

Profit = Revenue - Cost

$$\Rightarrow P(x) = R(x) - C(x) = 500x - 0 \cdot 2x^2 - (25x + 10000)$$

$$= 475x - 10000 - 0 \cdot 2x^2$$

For maximum or minimum:

$$\frac{dP}{dx} = 475 - 0 \cdot 2 \times 2x = 475 - 0 \cdot 4x = 0$$

$$\Rightarrow x = \frac{475}{0 \cdot 4} = 1187 \cdot 50$$

Also $\dfrac{d^2P}{dx^2} = -0 \cdot 4 < 0$

Hence the profit is maximum, when the output $(x) = 1187 \cdot 50$. At this level, the price is given by

$$p = 500 - 0 \cdot 2\, x$$

$$= 500 - 0 \cdot 2\,(1187 \cdot 50) = 262 \cdot 50$$

3.9 Summary:

The objective of this unit was to provide you with some exposure to differential calculus. Differential calculus is useful to solve optimisation problems, problems in which the aim is either to maximise or minimise a given objective function. Because of this reason it has found wide applications in this field. Applications of the derivative in both micro economics theory (cost, revenue, elasticity) and macro economic theory (income, consumption, savings) are good examples of its application in business.

This unit begins with a discussion on the concept of limit and continuity and then attention is directed to defining the slope of a linear function and proceeds with a discussion that extends this to include the slope of non - linear function. This is followed by the definition of the term derivative and rules for obtaining the derivatives of the more commonly encountered functional forms. The term derivative is a generalised expression for measuring the rate of change or slope of a function. Through several examples, the concepts of average cost, marginal cost, total revenue, marginal revenue, average revenue and elasticity are demonstarted by using the derivative first.

The procedures for determining local maxima and minima for the given function are demonstrated through an example and graph. A step by step procedure for finding maximum and minimum of a function is outlined. Each section in this unit is followed by an unsolved exercise for practice to the reader.

3.10 Technical Terms:

Classic Optimisation : Locating the maximum and / or minimum value(s) of a function through the application of differential calculus.

Continuity : A function is said to be continuous at a point x = a if (i) f (a) exists (ii) $\underset{x \to a}{Lt}\ f(x)$ exists and (iii) $\underset{x \to a}{Lt}\ f(x) = f(a)$

Critical Point : Any point that satisfies the necesary condition $\frac{dy}{dx} = 0$. These points may be maxima, minima or points of inflection.

Derivative : A function that expresses the slope of another function at every point.

Differential Calculus : It is concerned with determining the rate of change of a given function due to an unit change in one of the independent variables.

Integral Calculus : It is concerned with the inverse problem of find a function when its rate of change is given.

Limit : The method of knowing the behaviour of a function $y = f(x)$ as the independent variabel x approaches some particualar value.

Local maximum : A point on a curve that is highest than the points on both sides of itself. A point where $\frac{dy}{dx} = 0$ and $\frac{d^2y}{dx^2} > 0$

Local Minimum : A point on a curve that is lower than the points on both sides of itself. A point where $\frac{dy}{dx} = 0$ and $\frac{d^2y}{dx^2} > 0$

Point of inflection : A point on a curve at which the $\frac{dy}{dx}$ may or not be zero; $\frac{d^2y}{dx^2} = 0$

Slope : The rate of change in the dependent variable (y) for a unit change in the independent variable (x).

Tangent : A staright line that touches a non - linear function at only one point, not cutting through the curve at the point. The slope of the tangent is used as a measure of the slope of the curve at that point.

3.11 Exercise:

1. Evaluate

(a) $\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)$

(b) $\lim_{n \to \infty} \left(\frac{n-2}{n+1}\right)$

(c) $\lim_{x \to \infty} e^{1/x + 8}$

(d) $\lim_{x \to 0} \frac{2 - e^x}{e^{1/x}}$

(e) $\lim_{n \to \infty} \frac{1 + 2 + 3 + \cdots + n}{n^2}$

(f) $\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^{4x}$

2. (a) If $f(x) = \frac{4x^3 + 2x - 3}{x^2 + x + 5}$, find $\lim_{x \to \infty} f(x)$

(b) A wholesaler follows the following sales policy. If the quantity demanded is 50 units or less, the price per unit is Rs. 300. If the quantity demanded is more than 50 and not more than 100, the unit price is Rs. 250. If the quantity demanded is more than 100, the price per unit is Rs. 200. Construct the function and discuss the continuity.

3. Determine the derivative of following functions:

(a) $f(x) = \frac{x^5}{5} + \frac{x^{-2}}{3}$

(b) $f(x) = a\sqrt{x}$

(c) $f(x) = \left(x^2 + 3\right)\left(x^2 + 9\right)$

(d) $f(x) = \left(x^2 + 2x + 3\right)^{-6}$

(e) $f(x) = \frac{x^2 + 3x + 2}{x^2 - 3}$

(f) $f(x) = \frac{x^{-2} + x^{-1}}{x^2 - 3}$

(g) $y = 5x$

(h) $y = 5x^2 + 3x + 2$

4. The total cost $C(x)$ of purchasing x units of an item within each interval is as follows:

$$C(x) = \begin{cases} 15x & ; \quad 0 \le x \le 200 \\ 13x & ; \quad 200 < x \le 400 \\ 10x & ; \quad x > 400 \end{cases}$$

Find the points of discontinuity.

5. Suppose a salesman is paid a fixed sum of Rs. 500 per month together with a bonus of Rs. 2 for all items sold. Derive functional relationship for his salary and determine the slope of the line.

6. Suppose, total cost, y of the stock of an item as a function of order size, x is represented by equation

$$y = 4x + \frac{200}{x}$$

Comapare the slope between x = 8 and 9 with between 20 and 21. Also interpret your result.

7. The sales S for a product with price x is given by

find (i) Total sales revenue, $S = 2000\, e^{-0.5\, x}$, $R = x \cdot 5$ at x = 2

(ii) Marginal revenue $\frac{dR}{dx}$, at x = 2

8. The cost of fuel for running a train is proportional to the square of the speed generated in Kilometres per hour, and costs Rs. 75 per hour, at 17 kilometers per hour. What is the most economical speed, if the fixed charges are Rs. 400 per hour?

9. The sales S (in Rs. 1000's) of a product as a function of advertising expenditure x is given by $S = 2000 + 4000\left(1 - e^{-(0.01)x}\right)$ find the limit of S as $x \to \infty$ and interpret your result.

10. The demand for a certain product is represented by the equation p = 300 - 6q where p is the price per unit and q is the number of units demanded. Find the revenue function. What is the slope of the revenue function ? At what price is margianl revenue zero?

11. The demand of (in Kg) for a commodity when its price p (in Rs.) is given by $p = 108 - \left(\frac{3}{5q}\right)$

find the elasticity of demand when the price is Rs. 12.

12. The cost of fuel for running a train is proportional to the square of the speed generated in kilometers per hour, and costs Rs. 75 per hour at 17 kilometres per hour. What is the most economical speed, if the fixed charges are Rs. 400 per hour.

3.12 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics for Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. S.V.S. GIRIJA

Lesson - 4

MATRIX ALGEBRA AND APPLICATIONS

Objectives:

After studying this lesson you should be able to know

- Basic concepts of a matrix
- Methods of representing large quantities of data in matrix form
- Various operations concerning matrices
- The solution methods of simultaneous linear equations
- Applications of matrix algebra in various decision models

Structure:

4.1 Introduction:

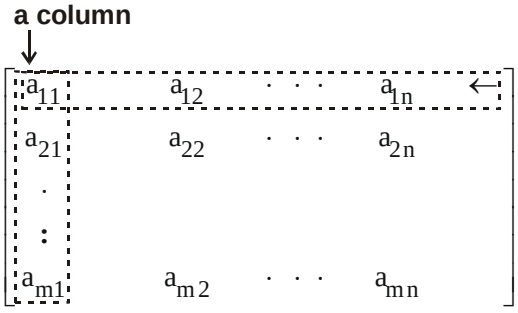
Matrices have proved their usefulness in quantitative analysis of managerial decisions in several disciplines like marketing, finance, production, personnel, economics, etc. Many quantitative methods such as linear programming, game theory, Markov models, input - output models and some statistical models have matrix algebra as their underlying theoretical base. All these models are built by establishing a system of linear equations which represent the problem to be solved . The simultaneous linear equations involving more than three variables cannot be solved by using "ordinary algebra". Real - world business problems may involve more than three variables, then in such cases matrices are used to represent a complex system of equations and large quantities of data in a compact form. Once the system of equations is represented in matrix form, they can be solved easily and quickly by using a computer. The limitation of matrix algebra is that it is applicable only in those cases where assumption of linearity can be made.

The main objective of this unit is to provide (i) some basic theoretical matrix operations - addition, subtraction and multiplication (ii) A procedure for solving a system of linear simultaneous equations, and (iii) a few applications of matrix algebra.

4.2 Matrix : Definition and Notations:

A matrix is a rectangular array of ordered numbers. The term ordered implies that the position of each number is significant and must be determined carefully to represent the information contained in the problem. These numbers (also called elements of the matrix) are arranged in rows and columns of the rectangular array and enclosed by either square brackets, [ ]; or parantheses ( ), or by paire of double vertical line || ||.

A matrix consisting of m rows and n columns is written in the following form:



where $a_{11}, a_{12}, \cdots$ denote the numbers (or elements) of the matrix. The dimension (or order) of the matrix is determined by the number of rows and columns. Here, in the given matrix, there are m rows and n columns. Therefore, it is of the dimension $m \times n$ (read as m by n). In the dimension of the given matrix the number of rows is always specified first and then the number of columns.

Boldface capital letters such as A, B, C ... are used to denote entire matrix. The matrix is also sometimes represented as $A = \left[a_{ij}\right]_{m \times n}$ where a_{ij} denotes the i^{th} row and the j^{th} element of A. Some examples of the matrices are

$$A = \begin{bmatrix} -1 & 1 \\ 2 & 3 \end{bmatrix}_{2\times 2} ; B = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 4 & -3 \end{bmatrix}_{2\times 3} ; C = \begin{bmatrix} 5 & 5 & 10 \\ 6 & 2 & 10 \\ 2 & 1 & 2 \end{bmatrix}_{3\times 3}$$

The matrix A is a 2 x 2 matrix because it has 2 rows and 2 columns. Similarly, the matrix B is a 2 x 3 matrix while matrix C is a 3 x 3 matrix.

4.3 Some Special Matrices:

(a) Square Matrix:

A matrix in which the number of rows equals the number of columns is called a square matrix. For example

$$\begin{bmatrix} 2 & 3 & 7 \\ 3 & 5 & 2 \\ 4 & 3 & 1 \end{bmatrix}_{3\times 3}$$

in a square matrix of dimension 3. The elements 2, 5 and 1 in this matrix are called the **diagonal elements** and the diagonal is called the **principal diagonal.**

(b) Diagonal Matrix:

A square matrix, in which all non - diagonal elements are zero whereas diagonal elements are non - zero is called a diagonal matrix. For example

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

is a diagonal matrix of dimension 3.

(c) Scalar Matrix:

A diagonal matrix in which all diagonal elements are equal is called a scalar matrix. For example

$$\begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix}_{3\times 3}$$

is a scalar matrix, where k is a real (or complex) number.

(d) Identify (or unit) matrix:

A scalar matrix in which all diagonal elements are equal to one, is called an identity (or unit) matrix and is denoted by I. Following are two different identity matrices

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}_{2\times 2} ; I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

An identity matrix of dimension n is denoted by I_n. It has n elements in its diagonal each equal to 1 and other elemetns are zero.

(e) The zero (or null) matrix:

A matrix is said to be the zero matrix if every element of it is zero. It is denoted as 0. Following are three different zero matrices.

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}_{2\times 2} ; \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{2\times 3} ; \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}_{3\times 2}$$

4.4 Matrix Representation of Data:

Before discussing the operations on matrices, it is necessary for you to know a few situations in which data can be represented in matrix form.

1. Transportation Problem:

The unit cost of transportation of an item from each of the two factories to each of the three warehouses can be represented in a matrix as shown below:

		Warehouses		
		W_1	W_2	W_3
Factory	F_1	20	15	30
	F_2	25	20	15

Similarly, we can also construct a time matrix $\left[t_{ij}\right]$, where t_{ij} = time of transportation of an item from factory i to warehouse j. Note that the time of transportation is independent of the amount shipped.

2. Distance Matrix:

The distance (in kms.) between given number of cities can be represented as matrix as shown below:

City

City	A	B	C	D
A	–	1,470	2,158	1,732
B	1,470	–	1,853	2,385
C	2,158	1,853	–	1,635
D	1,732	2,365	1,635	–

3. Diet Matrix:

The vitamin content of two types of foods and two types of vitamins can be represented in a matrix as shown below:

Food	Vitamins A	Vitamins B
F_1	150	120
F_2	170	100

(4) Assignment Matrix:

The time required to perform three jobs by three workers can be represented in a matrix as shown below:

Worker	J_1	J_2	J_3
W_1	5	3	2
W_2	4	5	3
W_3	2	4	6

(5) Pay - Off Matrix:

Suppose two players A and B play a coin tossing game. If outcome (H, H) or (T, T) occurs, then player B loses RS. 10 to player A, otherwise gains as shown in the matrix.

Player A	Player B: H	Player B: T
H	10	–10
T	–10	10

The minus sign with the pay off means that player A pays to B.

(6) Brand Switching Matrix:

The proportion of users in the population surveyed switching to brand j of an item in a period, given that they were using brand i can be represented as a matrix.

$$\text{From}\begin{matrix}\\ \text{Brand 1}\\ \text{Brand 2}\\ \text{Brand 3}\end{matrix}\begin{matrix}\text{To}\\ \text{Brand 1}\\ \text{Brand 2}\\ \text{Brand 3}\end{matrix}\begin{bmatrix}\text{Brand 1} & \text{Brand 2} & \text{Brand 3}\\ 0\cdot3 & 0\cdot6 & 0\cdot1\\ 0\cdot6 & 0\cdot3 & 0\cdot1\\ 0\cdot2 & 0\cdot5 & 0\cdot3\end{bmatrix}$$

Here the sum of the elements of each row is 1 because these are proportions.

4.5 Operations on Matrices:

1. Addition (or Subtraction) of Matrices:

The addition (or subtraction) of two or more matrices is possible only if these matrices have the same dimensions, i.e., matrices must have the same number of rows and same number of columns.

The sum (or difference) of matrices is obtained by adding (or subtracting) the corresponding elements of the given matrices. For example, if

$$A=\begin{bmatrix}1 & 3\\ 2 & 4\end{bmatrix}_{2\times2} \quad \text{and} \quad B=\begin{bmatrix}-1 & 7\\ 0 & 8\end{bmatrix}_{2\times2}$$

then $$A+B=\begin{bmatrix}1-1 & 3+7\\ 2+0 & 4+8\end{bmatrix}_{2\times2}=\begin{bmatrix}0 & 10\\ 2 & 12\end{bmatrix}_{2\times2}$$

$$A-B=\begin{bmatrix}1-(-1) & 3-7\\ 2-0 & 4-8\end{bmatrix}=\begin{bmatrix}2 & -4\\ 2 & -4\end{bmatrix}$$

Note that $A-B\neq B-A$

Example 1:

A company produces three types of products A, B and C. The total annual sales (in 000's of units) of these products for the years 1985 and 1986 in the four regions is given below.

For the year 1985:

Product \ Region	Eastern	Western	Southern	Northern
A	15	8	6	12
B	5	24	7	8
C	8	4	31	6

For the year 1986:

Product \ Region	Eastern	Western	Southern	Northern
A	17	10	5	7
B	5	22	11	4
C	13	6	39	5

Find the total sales of three products for two years.

Solution:

The total sales of three products for two years can be obtained by adding the sales of two years as shown below:

Product \ Region	Eastern	Western	Southern	Northern
A	15 + 17 = 32	8 + 10 = 18	5 + 5 = 10	12 + 7 = 19
B	5 + 5 = 10	24 + 22 = 46	7 + 11 = 18	8 + 4 = 12
C	8 + 13 = 21	4 + 6 = 10	31 + 39 = 70	5 + 6 = 11

Properties of matrix addition:

If A, B and C are any three matrices of same dimension, then

i) Matrix addition is commutative, i.e.,

$$A + B = B + A$$

ii) Matrix addition is associative, i.e.,

$$(A + B) + C = A + (B + C)$$

iii) For any matrix A of dimension $m \times n$, there is a zero matrix of the same dimension such that

$$A + 0 = 0 + A = 0$$

This shows that zero matrix is the additive identity.

(iv) If for any matrix A of dimension $m \times n$, there exists another matrix B of the same dimension such that

$$A + B = B + A = 0$$

then B is called the additive inverse (or negative) of A and is dentoed by - A.

2. Scalar Multiplication:

If $A\left[a_{ij}\right]$ is any matrix of dimension $m \times n$ and k is any scalar (real number); then the multiplication KA is obtained by simply by simply multiplying each element of A by the scalar K. That is

$$AK = KA = \left[k\, a_{ij}\right]$$

Example 2:

The sales figures in Example 1 are given in thousands of units. If we want to express sales figures in actual units, then we have to multiply the given matrices by 1000. For illustration, let us consider the data matrix of 1985. That is, if

$$A = \begin{bmatrix} 15 & 8 & 5 & 12 \\ 5 & 24 & 7 & 8 \\ 8 & 4 & 31 & 6 \end{bmatrix}$$

then

$$1000\ A = \begin{bmatrix} 15\times1000 & 8\times1000 & 5\times1000 & 12\times1000 \\ 5\times1000 & 24\times1000 & 7\times1000 & 8\times1000 \\ 8\times1000 & 4\times1000 & 31\times1000 & 6\times1000 \end{bmatrix}$$

$$= \begin{bmatrix} 15,000 & 8,000 & 5,000 & 12,000 \\ 5,000 & 24,000 & 7,000 & 8,000 \\ 8,000 & 4,000 & 31,000 & 6,000 \end{bmatrix}$$

Properties of scalar multiplication:

i) $K(A+B) = KA + KB$

where A and B are two matrices of same dimension and K is a scalar number.

ii) $(K_1 + K_2)A = K_1A + K_2A$

where A is a matrix and K_1 and K_2 are two distinct scalar numbers.

3. Multiplication of Matrices:

The matrix multiplication consists of the following steps:

a) Check on compatibility: The following dimensional arrangement must hold for compatibility in matrix multiplication.

dimensions : lead matrix x lag matrix = product

$$(m \times p) \times (p \times n) = m \times n$$

In other words, the number of columns in the first matrix must be equal to the number of rows in the second matrix. If this condition does not exist, then the matrices are said to be incompatible and their multiplication is not defined.

b) The operation of multiplication: For multiplication of two matrices the following procedure should be adopted:

i) The element of a row of the lead matrix A should be multiplied by the corresponding elements of a column of the lag matrix B.

ii) The product is then summed and the location of this resulting element in the new matrix C determines which row from A has to be multiplied with which column from B.

To illustrate this, let is take two matrices A and B as defined below:

$$A = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 7 \end{bmatrix}_{2\times 3}, \quad B = \begin{bmatrix} 2 & 3 \\ 3 & 5 \\ 5 & 7 \end{bmatrix}_{3\times 2}$$

then

$$\underset{(2\times 3)}{A} \times \underset{(3\times 2)}{B} = \underset{(2\times 2)}{C}$$

$$\begin{matrix} R_1 \\ R_2 \end{matrix}\begin{bmatrix} 2 & 3 & 5 \\ 3 & 5 & 7 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 3 & 5 \\ 5 & 7 \end{bmatrix} = \begin{bmatrix} R_1 \times C_1 & R_1 \times C_2 \\ R_2 \times C_1 & R_2 \times C_2 \end{bmatrix}_{2\times 2}$$

$$\begin{matrix} & C_1 & C_2 \end{matrix}$$

$$= \begin{bmatrix} 38 & 56 \\ 56 & 83 \end{bmatrix}_{2\times 2}$$

Example 3:

There are two families A and B. There are 2 men, 3 women and 1 child in family A and 1 man, 1 woman and 2 children in family B. The recommended daily allowance for calories is, man 2400; Women 1900, child 1800 and for proteins man 55 gm, woman 45 gm, and child 33 gm.

Represent the above information by matrices. Using matrix multiplication, calculate the total requirement of calories and proteins for each of the two families.

Solution:

Let C = Family

$$\begin{matrix} & \text{man} & \text{women} & \text{child} \\ A & 2 & 3 & 1 \\ B & 1 & 1 & 2 \end{bmatrix}_{(2\times 2)}$$

and

$$D = \begin{array}{c} \\ \text{Man} \\ \text{Women} \\ \text{Child} \end{array} \begin{array}{c} \begin{array}{cc} \text{Calory} & \text{Protein} \end{array} \\ \begin{bmatrix} 2400 & 55 \\ 1900 & 45 \\ 1800 & 33 \end{bmatrix}_{(3\times 2)} \end{array}$$

If you look at the dimensions of two matrices C and D, then you will find that the condition for multiplication is satisfied. Therefore, the total requirement of calories and proteins for each of the two families is determined by multiplying C and D, as shown below:

$$C \times D \quad \begin{array}{c} R_1 \rightarrow \\ R_2 \rightarrow \end{array} \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix} \times \begin{bmatrix} 2400 & 55 \\ 1900 & 45 \\ 1800 & 33 \end{bmatrix} = \begin{bmatrix} R_1C_1 & R_1C_2 \\ R_2C_1 & R_2C_2 \end{bmatrix}_{(2\times 2)}$$

C_1 (Calory) C_2 (Protein)

Calory

$$= \text{Family} \begin{array}{c} A \\ B \end{array} \begin{bmatrix} 1,03,000 & 2,78,000 \\ 61,000 & 1,33,000 \end{bmatrix}_{(2\times 2)}$$

2. If A and B are two non - zero compatible matrices with respect to multiplication, then their product.

   i) is always zero matrix ii) is never a zero matrix

   iii) may be a zero matrix iv) None of these

3. A factroy employes 50 skilled workers and 20 unskilled workers. The daily wages to skilled and unskilled workers are Rs. 30 and Rs. 17 respectively. Using matrix notation find

   a) the number of workers matrix

   b) the total daily payment made to the workers

Properties of matrix multiplication:

i) Matrix multiplication in general, is not commutative i.e.,

$$AB \neq BA$$

ii) Matrix multiplication is associative i.e.,

$$A(BC) = (AB)C$$

where A, B, C are any three matrices of dimension $m \times n,\ n \times p\ , p \times q$ respectively.

iii) Matrix multiplication is distributive

$$A\,(B + C) = A\,B + A\,C$$

where A, B, C are any three $m \times n,\ n \times p$ and $n \times p$ matrices respectively.

4. Transpose of Matrix:

Let A be any matrix. The matrix obtained by interchanging rows and columns of A is called the transpose of A and is denoted by A' or A^t. Thus if $A = [a_{ij}]$ is an $m \times n$ matrix, then $A^t = [a_{ij}]$ will be $n \times m$ matrix. For example, the transpose of the matrix is

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & 0 \end{bmatrix}_{2 \times 3}$$

$$A^t = \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 4 & 0 \end{bmatrix}_{3 \times 2}$$

Properties of transpose of matrices:

i) Transpose of a sum (or difference) of two matrices is the sum (or difference) of the transposes, i.e.,

$$(A \pm B)^t = A^t \pm B^t$$

ii) Transpose of a transpose is the original matrix

$$(A^t)^t = A$$

iii) Transpose of a product of two matrices is the product of their transposes taken in reverse order

$$(A\,B)^t = B^t\,A^t$$

4.6 Determinant of a Square Matrix:

The determinant of a square matrix is a scalar (i.e., a number). Determinants are possible only for square matrices. For more clarity, we shall be defining it in stages, starting with square

matrix of order 1, then for matrix of order 2 etc. The determinant of a square matrix A is denoted either by $|A|$ or det. A.

i) Determinant of order 1: Let $A = [a_{11}]$ be a matrix of order 1. Then det $A = a_{11}$

ii) Determinant of order 2: Let

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

be a squre matrix of order 2, then det A is defined as

$$\det\cdot\ A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}\, a_{22} - a_{21}\, a_{12}$$

for example

$$\det\cdot\ A = \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix} = 3 \times 2 - 1 \times 4 = 2$$

To write the expansion of a determinant to matrices of order 3, 4, ... let us first define two important terms.

a) Minor: Let A be a square matrix of order m. Then minor of an element a_{ij} is the determinant of the residual matrix (or submatrix) obtained from A by deleting row i and column j containing the element a_{ij}.

In the $|A|$, the minor of the element a_{ij} is denoted by M_{ij}. Thus, in the determinant of order 3.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

the minor of the element a_{11} is obtained by deleting first row and first column containing element a_{11} and is written as

$$M_{11} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$$

similarly, minor of a_{12} is

$$M_{12} = \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix}$$

b) Cofactor: The cofactor c_{ij} of an element a_{ij} is defined as

$$c_{ij} = (-1)^{i+j} M_{ij}$$

where M_{ij} is the minor of an element a_{ij}

Now using the concept of minor and cofactor, you can write the expansion of a determinant of order 3 as shown below:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13}$$

$$a_{11}(-1)^{1+1} M_{11} + a_{12}(-1)^{1+2} M_{12} - a_{13}(-1) M_{13}$$

$$= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}(a_{22}\, a_{33} - a_{32}\, a_{23}) - a_{12}(a_{21}\, a_{33} - a_{31}\, a_{23}) + a_{13}(a_{21}\, a_{32} - a_{31}\, a_{22})$$

The expansion of the given determinant can also be done by choosing elements in any row and column. In the above example expansion was done by using the elements of the first row.

Example 4:

Find the value of the determinant

$$\det \cdot A = \begin{vmatrix} 1 & 18 & 72 \\ 2 & 40 & 96 \\ 2 & 45 & 75 \end{vmatrix}$$

Solution:

If you expand the determinant by using the elements of the first column, then you will get

$$\begin{vmatrix} 1 & 18 & 72 \\ 2 & 40 & 96 \\ 2 & 45 & 75 \end{vmatrix} = 1\begin{vmatrix} 40 & 96 \\ 45 & 75 \end{vmatrix} - 2\begin{vmatrix} 18 & 72 \\ 45 & 75 \end{vmatrix} + 2\begin{vmatrix} 18 & 72 \\ 40 & 96 \end{vmatrix}$$

$$= 1(3000 - 4320) - 2(1350 - 3240) + 2(1728 - 2880)$$

$$= 1\text{x}\,(-1320) - 2\,\text{x}\,(-1890) + 2\,(-1152)$$

$$= -1320 + 3780 - 2304$$

$$= -3624 + 3780 = 156$$

Properties of determinants:

Following are the useful properties of determinants of any order. These properties are very useful in expanding the determinants.

1. The value of a determinant remains unchanged. If rows are changed into columns and columns into rows, i.e.,

$$|A| = |A^t|$$

2. If two rows (or columns) of a determinant are interchanged, then the value of the determinant so obtained is the negative of the original determinant.

3. If each element in any row or column of a dterminant is multiplied by a constant number say K, then the determinant so obtained is K times the original determinant.

4. The value of a determinant in which two rows (or columns) are equal is zero.

5. If any row (or column) of a determinant is replaced by the sum of the row and a linear combination of other rows (or columns), then the value of the determinant so obtained is equal to the value of the original determinant.

6. The rows (or columns) of a determinant are said to be linearly dependent if $|A| = 0$ otherwise independent.

Example 5:

Verify the following result

$$\begin{vmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{vmatrix} = (a - b)(b - c)(c - a)$$

Applying row operations (Property 5)

$$\begin{vmatrix} R_2 \to R_2 + (-1)\, R_1 \\ R_3 \to R_3 + (-1)\, R_1 \end{vmatrix}$$

On the given determinant, the new determinant so obtained

$$\begin{vmatrix} 1 & a & a^2 \\ 0 & b-a & b^2-a^2 \\ 0 & c-a & c^2-a^2 \end{vmatrix}$$

Expanding the new determinant by the elements of first column, you will get

$$\begin{vmatrix} b-a & b^2-a^2 \\ c-a & c^2-a^2 \end{vmatrix} = \begin{vmatrix} b-a & (b-a)(b+a) \\ c-a & (c-a)(c+a) \end{vmatrix}$$

Again performing row operations,

$$R_2 \to \frac{1}{(b-a)} R_2$$

$$R_3 \to \frac{1}{(c-a)} R_3$$

You will have

$$(b-a)(c-a)\begin{vmatrix} 1 & b+a \\ 1 & c+a \end{vmatrix}$$

$$= (b-a)(c-a)\{(c+a)-(b+a)\}$$

$$= (b-a)\ (c-a)\ (c-b)$$

$$= (a-b)(b-c)(c-a)$$

4.7 Inverse of a Matrix:

If for a given square matrix A, another square matrix B of the same order is obtained such that

$$A\,B = B\,A = I$$

then matrix B is called the inverse of A and is denoted by $B = A^{-1}$

Before start discussing the procedure of finding the inverse of a matrix, it is important to know the following results:

1. The matrix $B = A^{-1}$ is said to be the inverse of matrix A if and only if $AA^{-1} = A^{-1} A = I$

2. That is if the inverse of a square matrix multiplied by the original matrix, then result is an identity matrix. The inverse A^{-1} does not mean $1/A$ or I/A. This is simply a notation to denote the inverse of A.

3. Every square matrix may not have an inverse. For example, zero matrix has no inverse. Because, inverse of square matrix exists only if the value of its determinant is non - zero, i.e., A^{-1} exists if and only if $|A| \neq 0$

   For example let B be the inverse of the matrix A, then

   $$A B = B A = I \quad \text{or} \quad |AB| = I$$

   or $$|A||B| = 1 \quad (|I| = 1)$$

   Hence $|A| \neq 0$

4. If A square matrix A has an inverse, then it is unique. It can also be proved by letting two inverse B and C of A.

   We then have

   $$A B = BA = I \cdot \cdot \cdot (i) \text{ and } A C = C A = I \cdot \cdot \cdot (ii)$$

   Pre multiplying (i) by C, we get C A B = C I

   $$\text{or} \quad I B = C I \qquad \text{or} \qquad B = C(C A = I)$$

   This implies that the inverse of a square matrix is unique.

Singular Matrix:

A matrix is said to be singular if its determinnanat is equal to zero; Otherwise non-singular.

Properties of the inverse:

i) The inverse of the inverse is the original matrix, i.e., $(A^{-1})^{-1} = A$

ii) The inverse of the transpose of a matrix is the transpose of its inverse, i.e.

$$(A^t)^{-1} = (A^{-1})^t$$

iii) The identity matrix is its own inverse, i.e., $I^{-1} = I$

iv) The inverse of the product of two non - singular matrices is equal to the product of two inverses in the reverse order, i.e., $(A\,B)^{-1} = B^{-1}, A^{-1}$

Method of finding inverse of a matrix:

The procedure of finding inverse of a square matrix $A = [a_{ij}]$ of order n can be summarised in the following steps:

1. Construct the matrix of co-factors of each element a_{ij} in $|A|$ as follows:

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mn} \end{bmatrix}$$

In this case cofactors are the elements of the matrix

2. Take the transpose of the matrix of cofactors constructed in step 1. It is called adjoint of A and is denoted by Adj. A.

3. Find the value of $|A|$

4. Apply the following formula to calculate the inverse of A

$$A^{-1} = \frac{\text{Adj } A}{|A|}, |A| \neq 0$$

Example 6:

Find the inverse of the matrix

$$A = \begin{bmatrix} 1 & 3 & 0 \\ -2 & 3 & 3 \\ 1 & 1 & 4 \end{bmatrix}$$

Solution:

The determination of matrix A is expanded with respect to the elements of first row:

$$|A| = \begin{vmatrix} 1 & 3 & 0 \\ -2 & 3 & 3 \\ 1 & 1 & 4 \end{vmatrix} = 1\begin{vmatrix} 3 & 3 \\ 1 & 4 \end{vmatrix} - 3\begin{vmatrix} -2 & 3 \\ 1 & 4 \end{vmatrix} + 0\begin{vmatrix} -2 & 3 \\ 1 & 1 \end{vmatrix}$$

$$= 9 - 3(-11) = 42$$

Since $|A| \neq 0$ therefore the inverse of A exists. The matrix of cofactor of elements A is

$$C_{11} = (-1)^{1+1} M_{11} = \begin{vmatrix} 3 & 3 \\ 1 & 4 \end{vmatrix} = 9$$

$$C_{12} = (-1)^{1+2} M_{12} = \begin{vmatrix} -2 & 3 \\ 1 & 4 \end{vmatrix} = 11$$

$$C_{13} = (-1)^{1+3} M_{13} = \begin{vmatrix} -2 & 3 \\ 1 & 1 \end{vmatrix} = -5$$

$$C_{21} = (-1)^{2+1} M_{21} = \begin{vmatrix} -3 & 0 \\ 1 & 4 \end{vmatrix} = -12$$

$$C_{22} = (-1)^{2+2} M_{22} = \begin{vmatrix} 1 & 0 \\ 1 & 4 \end{vmatrix} = 4$$

$$C_{23} = (-1)^{2+3} M_{23} = \begin{vmatrix} -1 & 3 \\ 1 & 1 \end{vmatrix} = 2$$

$$C_{31} = (-1)^{3+1} M_{31} = \begin{vmatrix} 3 & 0 \\ 3 & 3 \end{vmatrix} = 9$$

$$C_{32} = (-1)^{3+1} M_{32} = \begin{vmatrix} -1 & 0 \\ -2 & 3 \end{vmatrix} = -3$$

$$C_{33} = (-1)^{3+3} M_{33} = \begin{vmatrix} 1 & 3 \\ -2 & 3 \end{vmatrix} = 9$$

The matrix of cofactors of elements of matrix A is

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} = - \begin{bmatrix} 9 & 11 & -5 \\ 12 & 4 & 2 \\ 9 & -3 & 9 \end{bmatrix}$$

The adj. A is now constructed by taking transpose of the cofactor matrix:

$$\text{Adj. } A = (\text{Co} - \text{Factor } A)^t \begin{bmatrix} 9 & -12 & 9 \\ 11 & 4 & -3 \\ -5 & 2 & 9 \end{bmatrix}$$

Hence $$A^{-1} = \frac{\text{Adj } A}{|A|}$$

$$= \frac{1}{42} \begin{bmatrix} 9 & -12 & 9 \\ 11 & 4 & -3 \\ -5 & 2 & 9 \end{bmatrix}$$

4.8 Solution of Linear Simultaneous Equations:

As mentioned earlier in this unit, matrix algebra is useful in solving a set of linear simultaneous equations involving more than two variables. Now the procedure for getting the solution will be demonstrated.

Consider the set of linear simultaneous equations

$$x - y + z = 4$$

$$2x + 5y - 2z = 3$$

These equations can also be solved by using ordinary algebra. However, to demonstrate the use of matrix algebra, the first step is to write the given system of equations, in matrix form as follows:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 5 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \text{ or } AX = B$$

where $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 5 & -2 \end{bmatrix}$

is known as the coefficient matrix in which coefficients of x are written in first column, coefficients of y in second column and the coefficients of z in the third column.

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

is the matrix of unknown variables x, y and z and $B = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$

is the matrix formed with the right hand terms in equations which do not involve unknowns x, y and Z.

Generalising the situation, let us consider m linear equations in n - unknowns $x_1, x_2, \cdots, x_n$;

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ &a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{aligned}$$

Writing this system of equations in matrix form,

where $AX = B$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \text{----------} & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}_{n \times 1}$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_m \end{bmatrix}_{m \times 1}$$

Classification of linear equations:

If matrix B is zero matrix, i.e. B = 0 then the system A X = 0 is said to be homogeneous system. Otherwise the system is said to be non - homogeneous.

Homogeneous Equations:

When the system is homogeneous, i.e., $b_1 = b_2 = \cdot \cdot \cdot = b_m = 0$ the only possible solution is $X = 0$ or $x_1 = x_2 = \cdot \cdot \cdot x_n = 0$. It is called a trivial solution. Any other solution if it exists is called non trivial solution of the homogeneous linear equations.

In order to solve the equation $\Delta X = 0$, we perform such an elementary operations or transformations on the given coefficient matrix A which does not change the order of the matrix. An elementrary operation is of any one of the following three types.

i) The interchange of any two rows (or columns)

ii) The multiplication (or division) of the elements of any row (or column) by any non - zero number, e.g. the R_i (row i) can be replaced by $K R_i \ (K \neq 0)$.

iii) The addition of the elements of any row (or column) to the corresponding elements of any other row (or column) multiplied by any number, e.g. $R_i \ (\text{row i})$ can be replaced by $R_i + K R_j$ where R_j is the row j and $K \neq 0$.

The elementary operation is called row operation if it applies to rows, and column operation if it applies to columns.

For the purpose of applying these elementary operations, we form another matrix called augmented matrix as shown below:

$$[A : B] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & \cdot b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & \cdot b_2 \\ \text{----------} & & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} & \cdot b_m \end{bmatrix}$$

Solution Method:

We shall apply Gauss - Jordon Method (also called Triangular form Reduction Method) to solve homogeneous linear equations. In this method the given system of linear equations is reduced to an equivalent simpler system (i.e. system having the same solution as the given one). The new system looks like:

$$x_1 + b_1 x_2 + C_1 x_3 = d_1$$

$$x_2 + C_2 x_3 = d_2$$

$$x_3 = d_3$$

This method helps, not only to find solution to homogeneous equations but also to non - homogeneous system of equations having any number of unknowns.

Example 7: Solve the following system of equations using Gauses Jordon Method.

$$x_1 + 3x_2 - x_3 = 0$$
$$2x_1 - x_2 + 4x_3 = 0$$
$$x_1 - 11x_2 + 14x_3 = 0$$

Solution: The given system of equations in matrix form is:

$$\begin{bmatrix} 1 & 3 & -2 \\ 2 & -1 & 4 \\ 1 & -11 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{ or } \Delta X = 0$$

The augmented matrix becomes

$$[A : 01] \begin{bmatrix} 1 & 3 & -2 & :0 \\ 2 & -1 & 4 & :0 \\ 1 & -11 & 14 & :0 \end{bmatrix}$$

Applying elementary row operations

$$R_2 \to R_2 - 2R_1$$

$$R_3 \to R_3 - R_1$$

The new equivalent matrix is:

$$\begin{bmatrix} 1 & 3 & -2 & :0 \\ 0 & 7 & 8 & :0 \\ 0 & -14 & 16 & :0 \end{bmatrix}$$

Again applying $R_3 \to R_3 - 2R_2$. The new equivalent matrix is:

$$\begin{bmatrix} 1 & 3 & -2 & :0 \\ 0 & -7 & 8 & :0 \\ 0 & 0 & 0 & :0 \end{bmatrix}$$

The equations equivalent to the given system of equations obtained by elementary row operations are:

$$x_1 + 3x_2 - 2x_3 = 0$$

$$-7x_2 + 8x_3 = 0 \quad \text{or} \quad x_2 - \left(8/7\right) x_3 = 0$$

$$0 = 0$$

The last equation, through true, is redundant and the system is equivalent to

$$x_1 + 3x_2 - 2x_3 = 0$$

$$x_2 - (8/7)\, x_3 = 0$$

This is not in triangular form because the number of equations being less than the number of unknowns.

This system can be solved in terms of x_3 by assigning an arbitary constant value, k to it. The general solution to the given system is given by

$$x_3 = k$$

$$x_2 = (8/7)\, k$$

$$x_1 + 3x_2 = 2\, k_3 \quad \text{or} \quad x_1 = -3(8/7)k + 2k = (-10/7)\, k$$

Exercise 8:

Solve the following system of equations using Gauss - Jordon Method

i) $4x_1 + x_2 = 0$ ii) $x_1 - 2x_2 + 3x_3 = 0$

$-8x_2 + 2x_2 = 0$ $2x_1 + 5x_2 + 6x_3 = 0$

Non - Homogeneous Linear Equations:

The non - homogeneous linear equations can be solved by any of the following three methods.

1. Matrix Inverse Method
2. Cramer's Method
3. Gauss - Jordon Method

Again, for the purpose of demonstrating above solution methods, we shall consider three equations with three unknowns.

1. Matrix Inverse Method:

Let $AX = B$

be the given system of linear equations, and also A^{-1} be the inverse of A.

Pre multiplying both sides of the equation by A^{-1}

$$A^{-1}(AX) = A^{-1}B$$

$$(A^{-1}A)X = A^{-1}B$$

$$IX = A^{-1}B$$

$$X = A^{-1}B$$

Where I is the identity matrix.

The value of X gives the general solution to the given set of simultaneous equations. This solution is thus obtained by (i) first finding A^{-1}, and (ii) Post multiplying A^{-1} by B.

When the system has a solution, it is said to be consistent, otherwise inconsistent. A consistent system has either just one solution or infinitely many solutions.

Example 8:

The daily cost, C of operating a hospital is a linear function of the number of in patients I and out - patients. P, plus a fixed cost a, i.e.,

$$C = a + b\ P + d\ I$$

Given the following data for three days, find the values of a, b and d by setting up a linear system of equations and using the matrix inverse.

Day	Cost (in Rs.)	No.of in patients, I	No.of out-patients, P
1	6,950	40	10
2	6,725	35	9
3	7,100	40	12

Solution:

Based on the given daily cost equation, the system of equations for three days cost can be written as:

$$a + 10b + 40d = 6,950$$

$$a + 9b + 35d = 6,725$$

$$a + 12b + 40\ d = 7,100$$

The system can be written in the matrix form as follows:

$$\begin{bmatrix} 1 & 10 & 40 \\ 1 & 9 & 35 \\ 1 & 12 & 40 \end{bmatrix} \begin{bmatrix} a \\ b \\ d \end{bmatrix} = \begin{bmatrix} 6,950 \\ 6,725 \\ 7,100 \end{bmatrix}$$

Which is of the form $AX = B$, where

$$A = \begin{bmatrix} 1 & 10 & 40 \\ 1 & 9 & 35 \\ 1 & 12 & 40 \end{bmatrix};\ X = \begin{bmatrix} a \\ b \\ d \end{bmatrix}, \text{and } B = \begin{bmatrix} 6 & 950 \\ 6 & 725 \\ 7 & 100 \end{bmatrix}$$

The inverse of a matrix A is obtained as follows:

$$\text{Adj } A = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}^t = \begin{bmatrix} 60 & 5 & -3 \\ -80 & 0 & 2 \\ 10 & -5 & 1 \end{bmatrix}^t = \begin{bmatrix} 60 & -80 & 10 \\ 5 & 0 & -5 \\ -3 & 2 & 1 \end{bmatrix}$$

$$|A| = \begin{vmatrix} 1 & 10 & 40 \\ 1 & 9 & 35 \\ 1 & 12 & 40 \end{vmatrix} = 1\begin{bmatrix} 9 & 35 \\ 12 & 40 \end{bmatrix} - 10\begin{bmatrix} 1 & 35 \\ 1 & 40 \end{bmatrix} + 40\begin{bmatrix} 1 & 9 \\ 1 & 12 \end{bmatrix}$$

$$= (360 - 420) - 10(40 - 35) + 40(12 - 9)$$

$$= -10 \neq 0$$

Since $|A| \neq 0$ therefore inverse of matrix A exists and is computed as

$$A^{-1} = \frac{\text{Adj} \cdot A}{|A|}$$

$$= -\frac{1}{10}\begin{bmatrix} 60 & -80 & 10 \\ 5 & 0 & -5 \\ 3 & 2 & 1 \end{bmatrix}$$

$$\therefore X = A^{-1} B$$

or $$\begin{bmatrix} a \\ b \\ d \end{bmatrix} = -\frac{1}{10}\begin{bmatrix} 60 & -80 & 10 \\ 5 & 0 & -5 \\ -3 & 2 & 1 \end{bmatrix}\begin{bmatrix} 6,950 \\ 6,725 \\ 7,100 \end{bmatrix}$$

$$= -\frac{1}{10}\begin{bmatrix} 60 \times 6,950 - 80 \times 6,725 + 10 \times 7,100 \\ 5 \times 6,950 + 0 \times 6,725 - 5 \times 7,100 \\ -3 \times 6,950 + 2 \times 6,725 + 1 \times 7,100 \end{bmatrix}$$

$$= -\frac{1}{10}\begin{bmatrix} -50,000 \\ -750 \\ -300 \end{bmatrix} = \begin{bmatrix} 5000 \\ 75 \\ 30 \end{bmatrix}$$

or $a = 5000,\ b = 75$ and $d = 30$.

2 Cramer's Method:

When the number of equations is equal to the number of unknowns and the determinant of the coefficeints has non - zero value, then the system has a unique solution which can be found by using Cramer's formula.

$$x_j = \frac{D_j}{D}, \quad j = 1, 2, \cdots, n$$

where $D = |a_{ij}|$ and determinant D_j is obtained from D by replacing column j by the column of constant terms (i.e., matrix B).

Example 9:

An automobile company uses three types of steel, S_1, S_2 and S_3 for producing three different types of cars C_1, C_2 and C_3. Steel requiremetns (in tons) for each type of car and total available steel of all the three types is summarised in the following table:

Types of steel	Type of Car			Total Steel Available
	C_1	C_2	C_3	
S_1	2	3	4	29
S_2	1	1	2	13
S_3	3	2	1	16

Determine the number of cars of each type which can be produced.

Solution:

Let x_1, x_2 and x_3 be the number of cars of the type C_1, C_2 and C_3 respectively which can be produced. Then system of three linear equations is:

$$2x_1 + 3x_2 + 4x_3 = 29$$

$$x_1 + x_2 + 2x_3 = 13$$

$$3x_1 + 2x_2 + x_3 = 16$$

These equations can also be represented in matrix form as shown below:

$$\begin{bmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 29 \\ 13 \\ 16 \end{bmatrix}$$

The determinant of the coefficients matrix is

$$D = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix} = 2\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - 3\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} + 4\begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}$$

$$= 2(1-4) - 3(1-6) + 4(2-3)$$

$$= 5 (\neq 0)$$

Applying Cramer's Method

$$X_1 = \frac{D_1}{D} = \frac{1}{5}\begin{bmatrix} 29 & 3 & 4 \\ 13 & 1 & 2 \\ 16 & 2 & 1 \end{bmatrix} = 2$$

$$X_2 = \frac{D_2}{D} = \frac{1}{5}\begin{bmatrix} 2 & 29 & 4 \\ 1 & 13 & 2 \\ 3 & 16 & 1 \end{bmatrix} = 3$$

$$X_3 = \frac{D_3}{D} = \frac{1}{5}\begin{bmatrix} 2 & 3 & 29 \\ 1 & 1 & 13 \\ 3 & 2 & 16 \end{bmatrix} = 4$$

Hence, the number of cars of type C_1, C_2 and C_3 which can be produced are 2, 3 and 4 respectively.

4.9 Applications of Matrices:

1. Markov Models:

A particular mathematical model which is concerned with the brand - switching behavior of consumers who are essentially repeat - buyers of the product, is known as Markov brand - switching model. These models help in predicting the market share of a product at time period t, if the market share at the time period $(t-1)$ is known. Markov models have also been used in the study of (i) equipment maintenance and failure probability, (ii) stock market price movements, etc...

The general expression for forecasting the buying levels at time $t = n + 1$ is given by

$$R_{n+1} = R_n\, p$$

where,

$$p = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \text{----------} & & & \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}$$

is the matrix of transition probabilities. Each element of it represents the probability that a customer will change his liking from one brand to another in his next purchase. This is the reason for calling them transition probabilities and $\sum_{j}^{n} p_{ij} = 1$

R = matrix of order $(1 \times n)$

representing the buying levels (or state probabilities) at a particular time period. If we know the buying levels at time t = 0, then we can find them at any time by solving the above equation by the relation.

$R_1 = R_0 P\,, \quad n = 0$

$R_2 = R_1 P = (R_0\, P)\, P = R_0 P^2 \quad ; \; n = 1$

.....

Now $R_n = R_0 p^n \quad ; \; n = n - 1$

Now as the time passes, i.e. $n \to \infty$ the purchasing levels (or market shares) tends to settle down to an equilibrium (or steady state). That is, once an equilibrium state is reached there will be no change in the future market shares. Thus

$$\lim_{n \to \infty} R_{n+1} = \lim_{n \to \infty} R_n \cdot P \qquad \text{or} \qquad R = R\,P$$

This relationship can be used to determine market shares in the long run.

Example 10:

Consider the following matrix of transition probabilities of a product available in the market in two brands:

	Brand A	Brand B
Brand A	0 . 9	0 . 1
Brand B	0 . 3	0 . 7

Determine the market shares of each of the brand in equilibrium position.

Solution:

If the row vector (matrix having only one row) represents the market share of the two brands at equilibrium, then

$$\underset{\text{i·e·}}{R} = RP$$

$$(r_1, r_2) = (r_1, r_2)\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

$$= (r_1, r_2)\begin{bmatrix} 0\cdot 9 & 0\cdot 1 \\ 0\cdot 3 & 0\cdot 7 \end{bmatrix}$$

or $r_1 = 0\cdot 9\, r_1 + 0\cdot 3 r_2$ or $-0\cdot 1 r_1 + 0\cdot 3 r_2 = 0 \quad \cdots (i)$

$r_2 = 0\cdot 1 r_1 + 0\cdot 7 r_2$ or $0\cdot 1 r_1 - 0\cdot 3 r_2 = 0 \quad \cdots (ii)$

These are two linear homogeneous simultaneous equations. But these are not independent since one can be derived from the other. Hence, in order to solve, one more equation is needed, which is

$$r_1 + r_2 = 1 \quad \cdots (iii)$$

This is because the market shares have been expressed in percentage, so the sum of market shares will be 1.

Solving equations (i) and (ii) with the help of equation (iii), to get market shares in an equilibrium condition,

$$r_1 = 0\cdot 75 \quad \text{and} \quad r_2 = 0\cdot 25$$

Hence the expected market shares in an equilibrium condition for brand A will be 0.75 and that of brand B will be 0.25

2. Input - Output Analysis:

The method of "input analysis" was first proposed by Wasily W. Leotief in the 1930s. This method is based on the concept of "economic inter - dependence", which means that every sector (or industry) of the economy is related to every other sector. That is, they are all inter - dependent and inter - related. This means, any change in one sector (such as strike) will affect all other industries to a varying degree. However, this technique does not explain or establish as to why such effects occur.

The input - output model is based on the following assumptions:

i) An economy is decomposed into n sectors (or industries) and each of these produces only one kind of product. Each of the sectors uses as input, the output of the other sectors. Let $x_j\ (j = 1, 2, \cdots, n)$ be the gross production (output) of the j^{th} sector.

ii) Let a_{ij} represents rupee value of the output from sector i which sector j must consume to produce one rupee worth of its own product. It can be calculated as follows:

$$a_{ij} = \frac{\text{Rupee value of the product of sector i required by sector j}}{\text{Rupee value of the total output of sector j}}$$

The a_n's for all i and j can be represented in matrix form as shown below:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}_{n \times n}$$

The matrix A is the technical input - output coefficient matrix. This matrix remains unchanged so long as the structure of the economy remains unchanged.

iii) There is neither shortages or surpluses of product under consideration. In other words, gross product of each sector is sufficient to meet the final demand as well as demands of other sectors. LEt $d_j \ (j = 1, 2, \cdots, n)$ be the final demand (in rupee value) for product produced by each of n sectors.

The input - output table displayed in the following table summarises information about the economy in question.

Producers Sectors \ Consumer Sectors			Intermediate use Sectors 1	2 ..... j .....n			Final use	Final demand
Sector	1	x_1	a_{11}	a_{12}	a_{1j}	a_{1n}	x_1	d_1
	2	x_2	a_{21}	a_{22}	a_{2j}	a_{2n}	x_2	d_2
	.....		.....	.....			.....	
	i	x_i	a_{i1}	a_{i2}	a_{rj}	x_{in}	x_i	d_i
	.....		.....	.....			.....	
	n	x_n	a_{n1}	a_{n2}	a_{nj}	a_{nn}	x_n	d_n

If the economy is assumed to be in a state of dynamic equilibrium (i.e. neither shortages nor surpluses) so that the total output is just sufficient to meet the input needs of each sector as well as the needs of the final demand of all sectors themselves, then

Output = Input

= Need of each sector + Final demand

$$x_i = \sum_{j=1}^{n} a_{ij} x_j + d_i \quad ; \text{ for sector } i = 1, 2, ..., n$$

In matrix notation, we have

$$x = A X + D$$

where $\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$; $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$

and $D = \begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ . \\ d_n \end{bmatrix}$

The above equation can also be rewritten as:

$$X = A X + D$$

$$I X = A X + D$$

$$I X - A X = D$$

$$(I - A) X = D$$

$$(I - A) X = D$$

$$X (I - A)^{-1} D \text{ provided } |I \ A'| \neq 0$$

Where I is the identity matrix. The value of X gives how much each sector must produce which is just sufficient to meet the final demand as well as the demand of all sectors themselves.

Example 11:

Given the following input - output table, calculate the gross output so as to meet the final demand of 200 units of Agriculture and 800 units of Industry.

Producer Sector	Consumer Sector		Final	Total
	Agriculture	Industry	Demand	Output
Agriculture	300	600	100	1000
Industry	400	1200	400	2000

Solution:

Using the notations as discussed above

a_{11} = Rupee value of the product of sector agriculture used by agriculture rupee value of the total output of sector agriculture

$$= \frac{300}{1000} = 0 \cdot 3$$

Similarly $$a_{12} = \frac{600}{2000} = 0 \cdot 6$$

$$a_{21} = \frac{400}{1000} = 0 \cdot 4$$

$$a_{22} = \frac{1200}{2000} = 0 \cdot 6$$

Thus the technological matrix A and final demand matrix D, becomes

$$A = \begin{bmatrix} 0 \cdot 3 & 0 \cdot 3 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix} \quad ; \quad D = \begin{bmatrix} 200 \\ 800 \end{bmatrix}$$

None $$I - A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 \cdot 3 & 0 \cdot 3 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix} = \begin{bmatrix} 1 - 0 \cdot 3 & 0 - 0 \cdot 3 \\ 0 - 0 \cdot 4 & 1 - 0 \cdot 6 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \cdot 7 & -0 \cdot 3 \\ -0 \cdot 4 & 0 \cdot 4 \end{bmatrix}$$

$$|I - A| = \begin{bmatrix} 0\cdot7 & -0\cdot3 \\ -0\cdot4 & 0\cdot4 \end{bmatrix} = 0\cdot28 - 0\cdot12$$

$$= 0\cdot16 \ (\neq 0)$$

$$(I - A)^{-1} = \frac{\text{Adj}\cdot(I-A)}{|I-A|} = \frac{1}{0\cdot16}\begin{bmatrix} 0\cdot4 & 0\cdot3 \\ 0\cdot4 & 0\cdot7 \end{bmatrix}$$

$$X = (I - A)^{-1} \ D = \frac{1}{0\cdot16}\begin{bmatrix} 0\cdot4 & 0\cdot3 \\ 0\cdot4 & 0\cdot7 \end{bmatrix}\begin{bmatrix} 200 \\ 800 \end{bmatrix}$$

$$= \frac{1}{0\cdot16}\begin{bmatrix} 320 \\ 640 \end{bmatrix} = \begin{bmatrix} 2000 \\ 4000 \end{bmatrix}$$

Hence the gross output of agriculture and industry must be 2000 units and 4000 units respectively.

4.10 Solved Problems:

1. (a) Prove that

$$\begin{vmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{vmatrix} = (a-b)(b-c)(c-a)$$

Solution:

$$\begin{vmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 1 \\ a-b & b-c & c \\ a^2-b^2 & b^2-c^2 & c^2 \end{vmatrix} \text{ apply } \begin{matrix} c_1 - c_2 \\ c_2 - c_3 \end{matrix}$$

$$= (a-b)(b-c)\begin{vmatrix} 0 & 0 & 1 \\ 1 & 1 & c \\ a+b & b+c & c^2 \end{vmatrix}$$

$$= (a-b)(b-c)\left\{\begin{vmatrix} 1 & 1 \\ a+b & b+c \end{vmatrix}\right\}$$

$$= (a-b)(b-c)(b+c-a-b)$$

$$= (a-b)(b-c)(c-a)$$

(b) Prove that

$$\begin{vmatrix} a+b+2c & a & b \\ c & b+c+2a & b \\ c & a & c+a+2b \end{vmatrix} = 2(a+b+c)^3$$

Solution:

$$\begin{vmatrix} a+b+2c & a & b \\ c & b+c+2a & b \\ c & a & c+a+2b \end{vmatrix} = \begin{vmatrix} 2a+2b+2c & a & b \\ 2a+2b+2c & b+c+2a & b \\ 2a+2b+2c & a & c+a+2 \end{vmatrix}$$

apply $c_1+c_2+c_3$

$$= 2(a+b+c)\begin{vmatrix} 1 & a & b \\ 1 & b+c+2a & b \\ 1 & a & c+a+2b \end{vmatrix}$$

$$= 2(a+b+c)\begin{vmatrix} 1 & a & b \\ 0 & b+c+a & 0 \\ 0 & 0 & c+a+b \end{vmatrix} \text{ apply } \begin{matrix} R_2-R_1 \\ R_3-R_1 \end{matrix}$$

$$= 2(a+b+c)\left\{\begin{vmatrix} b+c+a & 0 \\ 0 & c+a+b \end{vmatrix}\right\}$$

$$= 2(a+b+c)^3$$

2. Evaluate

$$\begin{vmatrix} 0 & ab^2 & ac^2 \\ a^2b & 0 & bc^2 \\ a^2c & b^2c & 0 \end{vmatrix}$$

Solution:

$$\begin{vmatrix} 0 & ab^2 & ac^2 \\ a^2b & 0 & bc^2 \\ a^2c & b^2c & 0 \end{vmatrix} = abc \begin{vmatrix} 0 & b^2 & c^2 \\ a^2 & 0 & c^2 \\ a^2 & b^2 & 0 \end{vmatrix}$$

$$abc\left(a^2b^2c^2\right)\begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix}$$

$$= a^3b^3c^3\left[-1(0-1)+(1-0)\right] = 2a^3b^3c^3$$

3. a) Find the adjoint of the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 2 & -1 & 3 \end{pmatrix}$$

and verify the theorem

$$A\left(\text{Adj } A\right) = \left(\text{Adj } A\right)A = |A|I_3$$

Solution:

$$\text{Adj } A = \text{transpose of} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}_{3\times 3}$$

where $A_{11} = \text{Co factor of } a_{11} \text{ in } |A| = (-1)^{1+1}\begin{vmatrix} 2 & -3 \\ -1 & 3 \end{vmatrix} = 3$

$$A_{12} = \text{co factor of } a_{12} \text{ in } |A| = (-1)^{1+2}\begin{vmatrix} 1 & -3 \\ 2 & 3 \end{vmatrix} = -9$$

$$A_{13} = \text{co factor of } a_{13} \text{ in } |A| = (-1)^{1+3}\begin{vmatrix} 1 & 2 \\ 2 & -1 \end{vmatrix} = -5$$

Similarly

$$A_{21} = (-1)^{2+1}\begin{vmatrix} 1 & 1 \\ -1 & 3 \end{vmatrix} = -4, \qquad A_{22} = (-1)^{2+2}\begin{vmatrix} 1 & 1 \\ 2 & 3 \end{vmatrix} = 1$$

$$A_{23} = (-1)^{2+3}\begin{vmatrix} 1 & 1 \\ 2 & -1 \end{vmatrix} = 3, \qquad A_{31} = (-1)^{3+1}\begin{vmatrix} 1 & 1 \\ 2 & -3 \end{vmatrix} = -5$$

$$A_{32} = (-1)^{3+2}\begin{vmatrix} 1 & 1 \\ 1 & -3 \end{vmatrix} = 4, \qquad A_{33} = (-1)^{3+3}\begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix} = 1$$

Therefore

$$\text{Adj } A = \begin{pmatrix} 3 & -9 & -5 \\ -4 & 1 & 3 \\ -5 & 4 & 1 \end{pmatrix}^t = \begin{pmatrix} 3 & -4 & -5 \\ -9 & 1 & 4 \\ -5 & 3 & 1 \end{pmatrix}$$

Also $|A| = 1\cdot 3 - 1\cdot(4) + 2\cdot(-5) = -11$

Now

$$A\,(\text{Adj } A) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 2 & -1 & 3 \end{pmatrix}\begin{pmatrix} 3 & -4 & -5 \\ -9 & 1 & 4 \\ -5 & 3 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} -11 & 0 & 0 \\ 0 & -11 & 0 \\ 0 & 0 & -11 \end{pmatrix} = -11\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= |A|\, I_3 \qquad \text{..... (1)}$$

Also $(\text{Adj A})\ A = \begin{pmatrix} 3 & -4 & -5 \\ -9 & 1 & 4 \\ -5 & 3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -3 \\ 2 & -1 & 3 \end{pmatrix}$

$$= \begin{pmatrix} -11 & 0 & 0 \\ 0 & 11 & 0 \\ 0 & 0 & 11 \end{pmatrix} = |A| I_3 \qquad \text{..... (2)}$$

From (1) and (2) we get

$$A(\text{Adj A}) = (\text{Adj } A) = |A| I_3$$

b) Find the inverse of the matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Solution:

$$A^{-1} = \frac{\text{Adj A}}{|A|}$$

$$|A| = (ad - bc)$$

Adj A = transpose of $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix}$

Now

$$A_{11} = (-1)^{1+1} d = d, \quad A_{12} = (-1)^{1+2} c = -c$$

$$A_{21} = (-1)^{2+1} b = -b, \quad A_{22} = (-1)^{2+2} a = a$$

$$\therefore \text{ Adj } A = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Hence $A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$

c) Compute the inverse of the matrix:

$$\begin{pmatrix} 1 & 0 & -4 \\ -2 & 2 & 5 \\ 3 & -1 & 2 \end{pmatrix}$$

Solution:

We know $A^{-1} = \frac{\text{Adj } A}{|A|}$

$$|A| = \begin{vmatrix} 1 & 0 & -4 \\ -2 & 2 & 5 \\ 3 & -1 & 2 \end{vmatrix}$$

$$= 1(4+5) - 0(-4-15) - 4(2-6) = 25$$

$$\text{Adj } A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}^t = \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{pmatrix}$$

The co-factors of the elements of A are

$$A_{11} = (-1)^{1+1}\begin{vmatrix} 2 & 5 \\ -1 & 2 \end{vmatrix} = 9, \qquad A_{12} = (-1)^{1+2}\begin{vmatrix} -2 & 5 \\ 3 & 2 \end{vmatrix} = 19$$

$$A_{13} = (-1)^{1+3}\begin{vmatrix} -2 & 2 \\ 3 & -1 \end{vmatrix} = -4, \qquad A_{21} = (-1)^{2+1}\begin{vmatrix} 0 & -4 \\ -1 & 2 \end{vmatrix} = 4$$

4. a) Solve the following system of equations using Gauss - Jordan Method.

$$x_1 + 2x_2 + x_3 = 6$$

$$2x_1 + 3x_2 + 4x_3 = 12$$

$$3x_1 + x_2 + 2x_3 = 7$$

Solution:

Augmented matrix is

$$A = \begin{pmatrix} 1 & 2 & 1 & 6 \\ 2 & 3 & 4 & 12 \\ 3 & 1 & 2 & 7 \end{pmatrix} \begin{matrix} R_2 \to R_2 - 2R_1 \\ R_3 \to R_3 - 3R_1 \end{matrix}$$

$$\sim \begin{pmatrix} 1 & 2 & 1 & 6 \\ 0 & -1 & 2 & 0 \\ 0 & -5 & -1 & -11 \end{pmatrix} \begin{matrix} R_3 \to -R_3 \\ R_3 \to R_3 + 5R_2 \end{matrix}$$

$$\sim \begin{pmatrix} 1 & 2 & 1 & 6 \\ 0 & -1 & 2 & 0 \\ 0 & 0 & 11 & 11 \end{pmatrix} \begin{matrix} R_1 \to R_1 + 2R_2 \\ R_3 \to \frac{1}{11} R_3 \end{matrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 5 & 6 \\ 0 & -1 & 2 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} R_2 \to R_2 - 2R_3 \\ R_1 \to R_1 - 5R_3 \end{matrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & -1 & 0 & -2 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad R_2 \to -R_2$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

The solution of system of equations is

$$x_1 = 1,\ x_2 = 2,\ x_3 = 1.$$

b) Solve the following equations:

$$5x - 6y + 4z = 15$$

$$7x + 4y - 3z = 19$$

$$2x + y + 6z = 46$$ by matrix inversion method.

solution:

Given system in the matrix notation is

$$\begin{pmatrix} 5 & -6 & 4 \\ 7 & 4 & -3 \\ 2 & 1 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 15 \\ 19 \\ 46 \end{pmatrix}$$

$$A\ X = B$$

$$X = A^{-1} B$$

$$A^{-1} = \frac{\text{Adj } A}{|A|} \quad \text{where} \quad |A| = \begin{vmatrix} 5 & -6 & 4 \\ 7 & 4 & -3 \\ 2 & 1 & 6 \end{vmatrix} = 419$$

Adj A = transpose of cofactor matrix = $\begin{pmatrix} 27 & 40 & 2 \\ -48 & 22 & 43 \\ -1 & -17 & 62 \end{pmatrix}$

$$A^{-1} = \frac{\text{Adj } A}{|A|} = \frac{1}{419}\begin{pmatrix} 27 & 40 & 2 \\ -48 & 22 & 43 \\ -1 & -17 & 62 \end{pmatrix}$$

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{1}{419}\begin{pmatrix} 27 & 40 & 2 \\ -48 & 22 & 43 \\ -1 & -17 & 62 \end{pmatrix}\begin{pmatrix} 15 \\ 19 \\ 46 \end{pmatrix}$$

$$= \frac{1}{419}\begin{pmatrix} 27\cdot 15 + 40\cdot 19 + 2\cdot 46 \\ -48\cdot 15 + 22\cdot 19 + 43\cdot 46 \\ -1\cdot 15 - 17\cdot 19 + 62\cdot 46 \end{pmatrix}$$

$$= \frac{1}{419}\begin{pmatrix} 1257 \\ 1676 \\ 2514 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 6 \end{pmatrix}$$

$\therefore \quad x = 3,\ y = 4 \ \text{ and } \ z = 6\cdot$

5. Solve the equations by Cramer's rule

$$3x_1 + x_2 + x_3 = 1$$

$$2x_1 + 2x_3 = 0$$

$$5x_1 + x_2 + 2x_3 = 2$$

Solution:

Given system can be written as $A\,X = B$

i.e. $$\begin{pmatrix} 3 & 1 & 1 \\ 2 & 0 & 2 \\ 5 & 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

where $A = \begin{pmatrix} 3 & 1 & 1 \\ 2 & 0 & 2 \\ 5 & 1 & 2 \end{pmatrix}$, $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, $B = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$

$$\Delta = |A| = 3(-2) - 1(-6) + 1(2) = 2$$

$$\Delta_1 = \begin{vmatrix} 1 & 1 & 1 \\ 0 & 0 & 2 \\ 2 & 1 & 2 \end{vmatrix} = 1(-2) + 2(2) = 2$$

$$\Delta_2 = \begin{vmatrix} 3 & 1 & 1 \\ 2 & 0 & 2 \\ 5 & 2 & 2 \end{vmatrix} = 1(4) - 2(1) + 2(-2) = -2$$

$$\Delta_3 = \begin{vmatrix} 3 & 1 & 1 \\ 2 & 0 & 0 \\ 5 & 1 & 2 \end{vmatrix} = 1(2) + 2(-2) = -2$$

$$x_1 = \frac{\Delta_1}{\Delta} = \frac{2}{2} = 1$$

$$x_2 = \frac{\Delta_2}{\Delta} = \frac{-2}{2} = -1$$

$$x_3 = \frac{\Delta_3}{\Delta} = \frac{-2}{2} = -1$$

Applications to matrices:

6. Mr X is a sole trader, manufacturing tables and chairs. Each table requires 5 hours of labour and 6 units of material. A chair requires 3 labour hours and 3 units of material. If Mr X plans to produce 10 tables and 15 chairs in the next week, how many hours will he need to work and how much material will he require?

Solution:

The labour requirement is $(10 \times 5) + (15 \times 3) = 95$ hours

The material requirement is $(10 \times 6) + (15 \times 3) = 105$ units

The matrix solution would be :

Tables	Chairs	Labour	Materials	Labour	Materials

$$(10 \quad 15)_{1\times 2} \times \begin{pmatrix} 5 & 6 \\ 3 & 3 \end{pmatrix}_{2\times 2} = (95 \quad 105)_{1\times 2}$$

It may be noted that

$$\begin{pmatrix} 5 & 6 \\ 3 & 3 \end{pmatrix}_{2\times 2} \times \begin{pmatrix} 10 \\ 15 \end{pmatrix}_{2\times 1} = \begin{pmatrix} 140 \\ 75 \end{pmatrix}_{2\times 1}$$

is incorrect as labour hours are being added to units of material.

7. A firm produces different pump units, each of which requires some components shown below in a tabular form:

Pump	Housing	Impeller	Bolts	Couplings	Inlets	Armoured Hose
Type A	1	1	5	4	2	8 m
Type B	1	1	7	3	2	4 m
Type C	1	1	3	5	2	3 m

The firm receives an order for 8 Type A pump units, 4 Type B units and 2 Type - C units. Using the notion of Matrix multiplication, obtain the matrix whose elements may represent the quantities of each item required to make up the order.

Solution:

The specifications of the different pump units with their components can be represented by the following matrix.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 5 & 7 & 3 \\ 4 & 3 & 5 \\ 2 & 2 & 2 \\ 8 & 4 & 3 \end{bmatrix}_{6\times 3}$$

Where each column represents the type of the pump and each row represents the different components required. The firm has received order for 8 type A, 4 type B, 2 type C units. This can be represented by the matrix,

$$\begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}_{3\times 1}$$

Therefore the matrix multiplication of these two matrices gives

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 5 & 7 & 3 \\ 4 & 3 & 5 \\ 2 & 2 & 2 \\ 8 & 4 & 3 \end{bmatrix}_{6\times 3} \times \begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}_{3\times 1} = \begin{bmatrix} 1\times 8+1\times 4+1\times 3 \\ 1\times 8+1\times 4+1\times 3 \\ 5\times 8+7\times 4+3\times 3 \\ 4\times 8+3\times 4+5\times 3 \\ 2\times 8+2\times 4+2\times 3 \\ 8\times 8+4\times 4+3\times 3 \end{bmatrix}_{6\times 1}$$

The first element of matrix (= 14) gives the number of components for housing the second (= 14) gives that of impeller and so on.

8. The following matrix gives the number of units of three products (P, Q and R) that can be processed per hour on three machines (A, B and C).

$$\begin{array}{c} \\ P \\ Q \\ R \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 10 & 12 & 15 \\ 13 & 11 & 20 \\ 16 & 18 & 14 \end{bmatrix} \end{array}$$

Determine by using matrix algebra, how many units of each product can be produced, if the hours available on machines A, B and C are 54, 56 and 48 respectively.

Solution:

$$\text{Units of products} = \begin{array}{c} \\ P \\ Q \\ R \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 10 & 12 & 15 \\ 13 & 11 & 20 \\ 16 & 18 & 14 \end{bmatrix} \end{array} \quad \begin{array}{c} \\ \begin{bmatrix} 54 \\ 46 \\ 48 \end{bmatrix} \end{array} \begin{array}{l} \\ A \\ B \\ C \end{array}$$

$$= \begin{bmatrix} 540 + 552 + 720 \\ 702 + 506 + 960 \\ 864 + 828 + 672 \end{bmatrix}$$

$$= \begin{bmatrix} 1812 \\ 2168 \\ 2364 \end{bmatrix} \begin{matrix} \text{P} \\ \text{Q} \\ \text{R} \end{matrix}$$

$\therefore$ 1812, 2168 and 2364 units of product P, Q and R are produced respectively.

9. The following matrix gives the proportionate mix of constituents for three fertilisers:

		Constituent			
		A	B	C	D
	1	0.5	0	0.5	0
Fertiliser	2	0.2	0.3	0	0.5
	3	0.2	0.2	0.1	0.5

i) If sales are 1000 tins (of one kilogram) per week, 20% being fertiliser 1, 30% being fertiliser 2, and 50% fertiliser 3, how much of each constituent is used?

ii) If the cost of each constituent is 50 paise, 60 paise, 75 paise and 100 paise per 100 grams, respectively how much does a one kilogram tins of each fertiliser cost?

iii) What is the total cost per week?

Express the calculations and answers in matrix form.

Solution:

(i) The sales of fertilisers per week can be expressed as the following matrix:

$$1000\begin{pmatrix} 0\cdot 2 & 0\cdot 3 & 0\cdot 5 \end{pmatrix} = \begin{pmatrix} 200 & 300 & 500 \end{pmatrix}$$

Thus

$$\begin{pmatrix} 200 & 300 & 500 \end{pmatrix} \begin{pmatrix} 0\cdot 5 & 0 & 0\cdot 5 & 0 \\ 0\cdot 2 & 0\cdot 3 & 0 & 0\cdot 5 \\ 0\cdot 2 & 0\cdot 2 & 0\cdot 1 & 0\cdot 5 \end{pmatrix}$$

$$= \begin{pmatrix} 260 & 190 & 150 & 400 \end{pmatrix}$$

Requirements of constituents are:

A : 260, B : 190, C : 150, D : 400.

(ii) Costs of each constituent are 50 p, 60 p, 75 p and 100 p per 100 grams, i.e., 500 p, 600p, 750p, and 1000 p per 1000 grams (one kilogram) of each constituent, respectively.

Thus

$$\begin{pmatrix} 0\cdot5 & 0 & 0\cdot5 & 0 \\ 0\cdot2 & 0\cdot3 & 0 & 0\cdot5 \\ 0\cdot2 & 0\cdot2 & 0\cdot1 & 0\cdot5 \end{pmatrix} \times \begin{vmatrix} 500 \\ 600 \\ 750 \\ 1000 \end{vmatrix} = \begin{pmatrix} 625 \\ 780 \\ 795 \end{pmatrix}$$

cost per 1 kg tin of fertilizer are:

1 : Rs. 6.25, 2 : Rs. 7.80, 3 : Rs. 7.95

(iii) The total cost of fertiliser if 1,000 one - kilogram this are needed per week may be calculated by either

$$\begin{pmatrix} 200 & 300 & 500 \end{pmatrix} \begin{pmatrix} 625 \\ 780 \\ 795 \end{pmatrix} = \begin{pmatrix} 7,56,500 \end{pmatrix}$$

or by $\begin{pmatrix} 260 & 190 & 150 & 400 \end{pmatrix} \begin{vmatrix} 500 \\ 600 \\ 750 \\ 1000 \end{vmatrix} = \begin{pmatrix} 7,56,500 \end{pmatrix}$

Hence total cost per week is Rs. 7,565.

10. The total cost of manufacturing three types of motor car is given by the following table:

	Labour (hrs)	Materials (units)	Sub - contracted work (units)
Car A	40	100	50
Car B	80	150	80
Car C	100	250	100

Labour costs Rs. 20 per hour, units of material cost Rs. 5 each and units of sub - contracted work cost Rs. 10 per unit. Find the total cost of manufacturing 3,000 ; 2,000 and 1,000 vehicles of type A, B and C respectively.

(Express the cost as a triple product of a three element row matrix, a 3 x 3 matrix and a three element column matrix and perform the multiplication according to the same rules you used for 2 x 2 matrices).

Solution: Let matrix P present labour hours, material used d and sub-contracted work for three types of cars A, B, C respectively.

$$\therefore\ P = \begin{bmatrix} 40 & 100 & 50 \\ 80 & 150 & 80 \\ 100 & 250 & 100 \end{bmatrix}$$

Further let matrix Q represent labour cost per unit, material cost and cost of sub - contracted work

$$Q = \begin{bmatrix} 20 \\ 5 \\ 10 \end{bmatrix}$$

The cost of each car A, B, C is now given by the column matrix.

$$PQ = \begin{bmatrix} 1800 \\ 3150 \\ 4250 \end{bmatrix}$$

Let the number of cars A, B, C to be manufactured in that order be represented by the row matrix.

$$R = \begin{bmatrix} 3000 & 2000 & 1000 \end{bmatrix}$$

Hence the total cost of manufacturing three cars A, B and C is given by the matrix.

$$PQR = \begin{bmatrix} 1800 \\ 3150 \\ 4250 \end{bmatrix} \times \begin{bmatrix} 3000 & 2000 & 1000 \end{bmatrix}$$

$$= \begin{bmatrix} 1,59,50,000 \end{bmatrix}$$

11. A manufacturer produces three products P, Q and R which he sells in two markets. Annual sales volumes are indicated as follows:

Markets	Products		
	P	Q	R
I	10,000	2,000	18,000
II	6,000	20,000	8,000

If unit sale prices of P, Q and R are Rs. 2.50, 1.25 and 1.50 respectively, find the total revenue in each market with the help of matrix Algebra.

If the unit costs of the above 3 commodities are Rs. 1.80, 1.20 and 0.80 respectively, find his gross profits.

Solution:

Total revenue in each market is obtianed from the matrix product:

$$\begin{bmatrix} 2\cdot 50 & 1\cdot 25 & 1\cdot 50 \end{bmatrix} \times \begin{bmatrix} 10000 & 6000 \\ 2000 & 20000 \\ 18000 & 8000 \end{bmatrix} = \begin{bmatrix} 54500 & 52000 \end{bmatrix}$$

$$\text{Total cost} = \begin{bmatrix} 1\cdot 80 & 1\cdot 20 & 0\cdot 80 \end{bmatrix} \times \begin{bmatrix} 10000 & 6000 \\ 2000 & 20000 \\ 18000 & 8000 \end{bmatrix}$$

$$= \begin{bmatrix} 34800 & 41200 \end{bmatrix}$$

Profits from market $A = 54500 - 34800 = 19700$

Profits from market $B = 52000 - 41200 = 10800$

12. In a certain city there are 25 colleges and 100 schools. Each school and college has 5 peons. 1 clerk and 1 cashier. Each college in addition has 1 accountant and 1 head - clerk. The monthly salary of each of them is as follows:

Peon - Rs. 300 ; Clerk - Rs. 500 ; Cashier - Rs. 600 ; Accountant - Rs. 700 and Head clerk - Rs. 800.

Using matrix notation find

(a) the total number of posts of each kind in schools and colleges taken together

(b) the total monthly salary bill of each school and college separately and

(c) the total monthly salary bill of all the schools and colleges taken together

Solution:

(a) Consider the row matrix of order 1 x 2

$$A = \begin{bmatrix} 25 & 100 \end{bmatrix}$$

This represents the number of colleges and schools in that order.

Let $B = \begin{bmatrix} 5 & 2 & 1 & 1 & 1 \\ 5 & 2 & 1 & 0 & 0 \end{bmatrix}$

where columns represent number of peons, clerks, cashier, accountant, head - clerk while rows represents colleges and schools in that order.

Then

$$AB = \begin{bmatrix} 25 & 100 \end{bmatrix}_{1\times 2} \times \begin{bmatrix} 5 & 2 & 1 & 1 & 1 \\ 5 & 2 & 1 & 0 & 0 \end{bmatrix}_{2\times 5}$$

$$= \begin{bmatrix} 625 & 250 & 125 & 25 & 100 \end{bmatrix}_{1\times 5}$$

where first element represents total number of peons, second, represents total number of clerks, third represents total number of cashiers, fourth represents total number of accountants and fifth represents total number of head - clerks.

(b) Let the column matrix

$$C = \begin{bmatrix} 300 \\ 500 \\ 600 \\ 700 \\ 800 \end{bmatrix}$$

represent montly salary of peon, clerk, cashier, accountant and head clerk in that order. Then

$$BC = \begin{bmatrix} 5 & 2 & 1 & 1 & 1 \\ 5 & 2 & 1 & 0 & 0 \end{bmatrix}_{2\times 5} \times \begin{bmatrix} 300 \\ 500 \\ 600 \\ 700 \\ 800 \end{bmatrix}_{5\times 1}$$

$$= \begin{bmatrix} 1500+1000+600+700+800 \\ 1500+1000+600+0+0 \end{bmatrix}_{2\times 1} = \begin{bmatrix} 4600 \\ 3100 \end{bmatrix}_{2\times 1}$$

Thus total monthly salary bill of each college is Rs. 4600 and of each school is Rs. 3100

(c) The total monthly salary bill of all schools and colleges taken together is

$$A(BC) = \begin{bmatrix} 25 & 100 \end{bmatrix}_{1\times 2} \times \begin{bmatrix} 4600 \\ 3100 \end{bmatrix}_{2\times 1}$$

$$= \begin{bmatrix} 1,5,000 + 3,10,000 \end{bmatrix}_{1\times 1}$$

$$= \begin{bmatrix} 4,25,000 \end{bmatrix}$$

13. The allocation of service department costs to production departments and other service departments is one area where matrix algebra may be used.

Consider the following data:

	Service Departments		**Production Department**	
	Maintenance	**Electricity**	**Matching**	**Assembly**
Manhours of maintenance time	-	3,000	16,000	1,000
Units of electricity Consumed	20,000	-	1,30,000	50,000
Department costs before any allocation of service departments	Rs. 50,000	Rs. 4,000	Rs. 1,40,000	Rs. 2,06,000

You are required to:

(i) Calculate the total costs to be allocated to the production departments using matrix algebra (formulate the problem and show all workings):

(ii) Show the allocation to the production departments, using matrix methods.

Solution:

(i) Let X be the total cost of the maintenance department (i.e., including an allocation of electricity costs).

Let Y be the total cost of electricity (i.e., including an allocation of maintenance costs).

Proportion of maintenance time consumed by electricity department is

$$\frac{3000}{3000 + 16000 + 1000} = \frac{3000}{20000} = 0\cdot 15$$

i.e., 15% of the maintenance dept. costs should be allocated to the electricity department.

$\therefore \quad Y = 4000 + 0 \cdot 15\ X$ ..... (*)

Likewise, the proportion of total electricity consumption used by the maintenance department is

$$\frac{20000}{20000 + 130000 + 50000} = \frac{20000}{200000} = 0 \cdot 1$$

So that 10% of the electricity cost should be allocated to the maintenance department.

$\therefore \quad X = 50000 + 0 \cdot 1\ Y$ ..... (* *)

From (*) and (* *), we get

$$-0 \cdot 15\ X + Y = 4000$$

$$X - 0 \cdot 1\ y = 50000$$

i.e., $\begin{pmatrix} -0 \cdot 15 & 1 \\ 1 & -0 \cdot 1 \end{pmatrix} \times \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 4000 \\ 50000 \end{pmatrix}$

$$\Rightarrow \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} -0 \cdot 15 & 1 \\ 1 & -0 \cdot 1 \end{pmatrix}^{-1} \times \begin{pmatrix} 4000 \\ 50,000 \end{pmatrix}$$

$$= \frac{1}{0 \cdot 985} \begin{pmatrix} 0 \cdot 1 & 1 \\ 1 & 0 \cdot 15 \end{pmatrix} \times \begin{pmatrix} 4000 \\ 50000 \end{pmatrix}$$

$$= \begin{pmatrix} 51,168 \\ 11675 \end{pmatrix}$$

Hence X = Rs. 51,168 and Y = Rs. 11,675.

(ii) The proportions of maintenance and electricity consumed by the production departments are:

	Maintenance	**Electricity**
Machine	$\frac{16,000}{20,000} = 0 \cdot 8$	$\frac{1,30,000}{2,00,000} = 0 \cdot 65$
Assembly	$\frac{1,000}{20,000} = 0 \cdot 05$	$\frac{50,000}{2,00,000} = 0 \cdot 25$

Accordingly the allocations of maintenance costs to the production department is

$$\begin{pmatrix} 0\cdot 8 & 0\cdot 65 \\ 0\cdot 05 & 0\cdot 25 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$$

$$= \begin{pmatrix} 0\cdot 8 & 0\cdot 65 \\ 0\cdot 05 & 0\cdot 25 \end{pmatrix}\begin{pmatrix} 51{,}168 \\ 11{,}675 \end{pmatrix} = \begin{pmatrix} 48{,}523 \\ 5{,}477 \end{pmatrix}$$

i.e., Rs. 48,523 machining and RS. 5,477 to assembly, a total of Rs. 54,000.

14. A, B and C has Rs. 480, Rs. 760 and Rs. 710 respectively. They utilised the amounts to purchase three types of shares of prices x, y and z respectively. A purchases 2 shares of price x , 5 of price y and 3 of price z. B purchases 4 shares of price x, 3 of price y and 6 of price z, C purchases 1 share of price x, 4 of price y and 10 of price z. Find x, y and z.

Solution:

We obtain the following set of simultaneous linear equations:

$$2x + 5y + 3z = 480$$

$$4x + 3y + 6z = 760$$

$$x + 4y + 10z = 710$$

The above system of equations in the matrix notation is

$$\begin{bmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 4 & 10 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 480 \\ 760 \\ 710 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 4 & 10 \end{bmatrix}^{-1} \times \begin{bmatrix} 480 \\ 760 \\ 710 \end{bmatrix} \quad \text{..... (*)}$$

Now $A^{-1} = \dfrac{\text{Adj } A}{|A|}$; where $|A| = \begin{vmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 4 & 10 \end{vmatrix} = -119$

and $\text{Adj } A = \begin{bmatrix} +6 & -38 & +21 \\ -34 & +17 & 0 \\ +13 & -3 & -14 \end{bmatrix}$ (Try yourself)

From (*), we get

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = -\frac{1}{119}\begin{bmatrix} +6 & -38 & +21 \\ -34 & +17 & 0 \\ +13 & -3 & -14 \end{bmatrix} \times \begin{bmatrix} 480 \\ 760 \\ 710 \end{bmatrix}$$

$$= -\frac{1}{119}\begin{bmatrix} 6\times 480 - 38\times 760 + 21\times 710 \\ -34\times 480 + 17\times 760 + 0\times 710 \\ 13\times 480 - 3\times 760 - 14\times 710 \end{bmatrix}$$

$$= -\frac{1}{119}\begin{bmatrix} -11090 \\ -3400 \\ -5980 \end{bmatrix} = \begin{bmatrix} 11090/119 \\ 3400/119 \\ 5980/119 \end{bmatrix}$$

Hence $x = \frac{11090}{119}$, $y = \frac{3400}{119}$, $z = \frac{5980}{119}$.

15. To control a certain crop disease it is necessary to use 8 units of chemical A, 14 units of chemical B and 13 units of chemical C. One barrel of spray P contains one unit of A, 2 units of B and 3 units of C. One barrel of spray Q contains 2 units of A, 3 units of B and 2 units of C. One barrel of spray R contains one unit of A, 2 units of B and 2 units of C. How many barrels of each type of spray should be used to control the disease?

Solution:

To grasp the situation easily, let us tabulate the data as follows:

		Spray			**Requirement in chemicals**
		P	Q	R	
Chemical	A	1	2	1	8
	B	2	3	2	14
	C	3	2	2	13
Quantity in each spray		x	y	z	

Let x barrels of spray P, y barrels of spray Q and z barrels of spray R be used to control disease. Then

$$x + 2y + z = 8$$

$$2x + 3y + 2z = 14$$

$$3x + 2y + 2z = 13$$

Writting the equations in the matrix form, we get

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 3 & 2 & 2 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 8 \\ 14 \\ 13 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 3 & 2 & 2 \end{bmatrix}^{-1} \times \begin{bmatrix} 8 \\ 14 \\ 13 \end{bmatrix}$$

Now $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 3 & 2 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} +2 & -2 & +1 \\ +2 & -1 & 0 \\ -5 & +4 & -1 \end{bmatrix}$ (Try yourself)

$$\therefore \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} +2 & -2 & +1 \\ +2 & -1 & 0 \\ -5 & +4 & -1 \end{bmatrix} \times \begin{bmatrix} 8 \\ 14 \\ 13 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$\Rightarrow \quad x = 1, \quad y = 2 \quad$ and $\quad z = 3 \cdot$

Hence 1 barrel of the spray P, 2 barrels of spray Q and 3 barrels of spray R should be used to control the disease.

16. The XYZ Bakery Ltd. produces three basic pastry mixes A, B and C. In the past the mix of ingredients has been as shown in the following matrix.

		Flour	Fat	Sugar
Type	A	5	1	1
	B	6.5	2.5	0.5
	C	4.5	3	2

(All quantities in kilogram weight)

Due to changes in consumer tastes it has been decided to change the mixes using the following amendment matrix:

		Flour	Fat	Sugar
Type	A	0	+ 1	0
	B	- 0.5	+ 0.5	- 0.5
	C	+ 0.5	0	0

Using matrix algebra you are required to calculate:

(i) the matrix for the new mix

(ii) the production requirements to meet an order for 50 units of type A, 30 units of type B and 20 units of type C of the new mix;

(iii) the amount of each type that must be made to totally use up 3700 kgs. of flour 1700 kgs of fat and 800 kgs of sugar that are at present in the stores.

Solution:

(i) The new mix is given by the addition of the original mix matrix and the amendment matrix.

$$\begin{pmatrix} 5 & 1 & 1 \\ 6\cdot5 & 2\cdot5 & 0\cdot5 \\ 4\cdot5 & 3 & 2 \end{pmatrix} + \begin{pmatrix} 0 & +1 & 0 \\ -0\cdot5 & +0\cdot5 & +0\cdot5 \\ +0\cdot5 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 1 \\ 6 & 3 & 1 \\ 5 & 3 & 2 \end{pmatrix}$$

Therefore, the answer to part (i) is

	Flour	Fat	Sugar
Type A	5	2	1
Type B	6	3	1
Type C	5	3	2

(ii) To determine the production requirements it is necessary to multiply the order vector by the new mix matrix,

$$\begin{pmatrix} 50 & 30 & 20 \end{pmatrix} \begin{pmatrix} 5 & 2 & 1 \\ 6 & 3 & 1 \\ 5 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 530 & 250 & 120 \end{pmatrix}$$

$\therefore$ 530 kgs of flour, 250 kgs of fat, 120 kgs of sugar.

(iii) $5X_1 + 6X_2 + 5X_3 = 3700$

$2X_1 + 3X_2 + 3X_3 = 1700$

$X_1 + X_2 + 2X_3 = 800$

$$\begin{pmatrix} 5 & 6 & 5 \\ 2 & 3 & 3 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 3700 \\ 1700 \\ 800 \end{pmatrix}$$

$\Rightarrow A X = B$

$\Rightarrow X = A^{-1} B$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 5 & 6 & 5 \\ 2 & 3 & 3 \\ 1 & 1 & 2 \end{pmatrix}^{-1} \times \begin{pmatrix} 3700 \\ 1700 \\ 800 \end{pmatrix}$$

On simplification, we get

$X_1 = 400,\ X_2 = 200$ and $X_3 = 100 \cdot$

17. A mixture is to be made of three foods A, B, C. The three foods A, B, C contain nutrients P, Q, R as shown in the tabular column. How to form a mixture which will have 8 gms of P, 5 gms of Q and 7 gms of R.

gms per kg of

Food	Nutrient P	Nutrient Q	Nutrient R
A	1	2	5
B	3	1	0
C	4	2	2

Solution:

Let x kgs of food A, y kgs of food B, and z kgs of food C be choosen to make up the mixture.

Then we have the equations,

$x + 3y + 4z = 8$

$$2x + y + 2z = 5$$

$$5x + 2z = 7$$

Expressing these equations as a single matrix equation, we have

$$\begin{pmatrix} 1 & 3 & 4 \\ 2 & 1 & 2 \\ 5 & 0 & 2 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \\ 7 \end{pmatrix}$$

or $$\begin{pmatrix} 1 & 3 & 4 \\ 0 & -5 & -6 \\ 0 & -15 & -18 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ -11 \\ -13 \end{pmatrix} \quad \begin{matrix} \text{Apply} \\ R_2 + (-2) R_1 \\ R_3 + (-5) R_1 \end{matrix}$$

or $$\begin{pmatrix} 1 & 3 & 4 \\ 0 & 5 & 6 \\ 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 11 \\ 0 \end{pmatrix} \quad \begin{matrix} \text{Apply} \\ (-1) R_2 \\ R_3 + (-3) R_2 \end{matrix}$$

Therefore, we have

$$x + 3y + 4z = 8 \qquad \text{..... (*)}$$

$$5y + 6z = 11 \qquad \text{..... (* *)}$$

Let $z = a$. From (* *), $5y + 6a = 11$, i.e., $y = \dfrac{11 - 6a}{5}$

Substituting in (*), $x + 3\dfrac{(11 - 6a)}{5} + 4a = 8$

$$\Rightarrow 5x + 3(11 - 6a) + 20a = 40$$

$$\Rightarrow 5x = 7 - 2a \quad \text{or} \quad x = \frac{7 - 2a}{5}$$

$\therefore$ The solution is $x = \dfrac{7 - 2a}{5}$, $y = \dfrac{11 - 6a}{5}$, $z = a$.

As 'a' changes we can get any number of solutions and thus there are any number of mixtures. Since x, y, z take non - negative values $z \geq 0,\ i \cdot e \cdot\ a \geq 0 \cdot$

Considering the value of x, we have

$$\frac{7-2a}{5} \geq 0,\ i \cdot e.,\ 7-2a \geq 0,\ i \cdot e.,\ 7 \geq 2a,\ i \cdot e \cdot,\ a \leq \frac{7}{2} \qquad \text{..... (I)}$$

Considering the value of y,

$$\frac{11-6a}{5} \geq 0, i \cdot e \cdot, 11-6a \geq 0,\ i \cdot e \cdot,\ 11 \geq 6a\ ,\ i \cdot e \cdot,\ a \leq \frac{11}{6} \qquad \text{..... (II)}$$

The restriction (II) covers the restriction (I)

Therefore, we have $0 \leq a \leq \frac{11}{6}$

$\therefore$ when $a = 1,\ x = 1,\ y = 1$ and $z = 1 \cdot$

18. A B C company has two service departments S_1 and S_2 and four production departments P_1, P_2, P_3 and $P_4 \cdot$

Overhead is allocated to the production departmetns for inclusion in the stock valuation. The analysis of benefits received by each department during the last quarter and the overhead expense incurred by each department were:

Service Department	Percentages to be ailocated to departments					
	S_1	S_2	P_1	P_2	P_3	P_4
S_1	0	20	30	25	15	10
S_2	30	0	10	35	20	5
Direct overhead expense Rs. '000	20	40	25	30	20	10

You are required to:

(i) Express the total overhead of the service departments in the form of simultaneous equations:

(ii) Express these equations in a matrix form;

(iii) Determine the total overhead to be allocated from each of S_1 and S_2 to the production departments.

Solution: (i) Let

$S_1 =$ total overhead of service department S_1

$S_2 =$ total overhead of service department S_2

Then $S_1 = 20,000 + 0 \cdot 3\, S_2$

$S_2 = 40,000 + 0 \cdot 2\ S_1$

Written as simultaneous equations, this becomes

$S_1 - 0 \cdot 3\ S_2 = 20,000$

$-0 \cdot 2\ S_1 + S_2 = 40,000$

(ii) In matrix form, the equations are written as

$$\underset{E}{\begin{pmatrix} 20,000 \\ 40,000 \end{pmatrix}} = \underset{A}{\begin{pmatrix} 1 & -0 \cdot 3 \\ -0 \cdot 2 & 1 \end{pmatrix}} \times \underset{S}{\begin{pmatrix} S_1 \\ S_2 \end{pmatrix}}$$

$$\Rightarrow \underset{S}{\begin{pmatrix} S_1 \\ S_2 \end{pmatrix}} = \underset{A^{-1}}{\begin{pmatrix} 1 & -0 \cdot 3 \\ -0 \cdot 2 & 1 \end{pmatrix}^{-1}} \times \underset{E}{\begin{pmatrix} 20,000 \\ 40,000 \end{pmatrix}}$$

(iii) By the normal rules for finding the inverse of a 2×2 matrix, this equals

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \frac{1}{0 \cdot 94} \begin{pmatrix} 1 & 0 \cdot 3 \\ 0 \cdot 2 & 1 \end{pmatrix} \times \begin{pmatrix} 20,000 \\ 40,000 \end{pmatrix} = \begin{pmatrix} 34,043 \\ 46,808 \end{pmatrix}$$

The allocation of overhead from S_1 and S_2 becomes:

$$\left(S_1 \right) \times \left(0 \cdot 3 \quad 0 \cdot 25 \quad 0 \cdot 15 \quad 0 \cdot 1 \right) = \begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix}$$

$$\Rightarrow (34,043) \times (0\cdot3 \;\; 0\cdot25 \;\; 0\cdot15 \;\; 0\cdot1) = \begin{pmatrix} 10,213 \\ 8,511 \\ 5,106 \\ 3,404 \end{pmatrix}$$

and $$(S_2) \times (0\cdot1 \;\; 0\cdot35 \;\; 0\cdot2 \;\; 0\cdot05) = \begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix}$$

$$\Rightarrow (46,808)\ (0\cdot1 \;\; 0\cdot35 \;\; 0\cdot2 \;\; 0\cdot25) = \begin{pmatrix} 4,681 \\ 16,383 \\ 9,362 \\ 2,340 \end{pmatrix}$$

The final allocation becomes:

Department	**Total** **Rs.**	P_1 **Rs.**	P_2 **Rs.**	P_3 **Rs.**	P_4 **Rs.**
S_1	27,234	10,213	8,511	5,106	3,404
S_2	32,766	4,681	16,383	9,362	2,340
Total	60,000	14,894	24,894	14,468	5,744

19. Given the following transaction matrix, find the input - output coefficient:

Purchasing Sector / Producing Sector	**Agriculture**	**Industry**	**Final Demand**
Agriculture	300	600	100
Industry	400	1200	400
Consumer	300	200	---

Find also total output as well as total input:

Solution:

Total output for agriculture is

$$300 + 600 + 100 = 1000 \quad \text{and}$$

For industry $400 + 1200 + 400 = 2000$

Similarly total input for agriculture is

$$300 + 400 + 300 = 1000 \quad \text{and}$$

For industry $600 + 1200 + 200 = 2000$

The above transaction can be put in the following way:

Purchasing sector output/ Producting sector input	Agriculture	Industry	Final Demand	Total Output
Agriculture	300	600	100	1000
Industry	400	1200	400	2000
Consumer	300	200	0	500
Total input	1000	2000	500	3500

Now to find out input - output coefficients:

A coefficient is obtained by industry's input by total output. It is an indication of the number of any industry's output needed to produce one unit of another industry's output.

Therefore, coefficient of input - output can be obtained as follows:

$$\frac{300}{1000} = 0 \cdot 30 \quad ; \quad \frac{600}{2000} = 0 \cdot 30$$

$$\frac{400}{1000} = 0 \cdot 40 \quad ; \quad \frac{1200}{2000} = 0 \cdot 60$$

$$\frac{300}{1000} = 0 \cdot 30 \quad ; \quad \frac{200}{2000} = 0 \cdot 10$$

which can be represented as follows:

Purchasing sector output / Producing sector input	Agriculture	Industry
Agriculture	0.30	0.30
Industry	0.40	0.60
Consumer	0.30	0.10

20. Suppose the interrelationship between the production of two industries R and S in a given year is

			Current Consumser	
	R	**S**	**Demand**	**Total output**
R	14	6	8	28
S	7	18	11	36

If the forecast demand in two years is

$$D_2 = \begin{bmatrix} 20 \\ 30 \end{bmatrix}$$

What should be total output X be?

Solution: Step I:

To obtain the input - output matrix, we determine how much of each of the two products R and S is required to produce one unit of R. For example, to obtain 28 units of R requires the use of 14 units of R and 7 units of S (the entries in column one). Forming the ratios, we find that to produce 1 unit of R requires $\frac{14}{28} = \frac{1}{2}$ of R, $\frac{7}{28} = \frac{1}{4}$ of S. If we want say X_1 units of R, we will require $\frac{1}{2} X_1$ units of R, $\frac{1}{4} X_1$ units of S.

Continuing in this way, we can construct the input - output matrix as follows:

$$A = \begin{array}{c} \\ R \\ S \end{array} \begin{array}{c} \begin{array}{cc} R & S \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{2} \end{bmatrix} \end{array}$$

It may be noted that column 1 represents the amounts R, S required for one unit of R, column 2 represents the amounts of R, S required for one units of S. For example, the entry in row 1, column 2 represents the amount of S needed to produce one unit of S.

As a result of placing the entries this way, if

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

represents the total output required to obtain a given demand, the product AX represents the amounts of R and S required for internal consumption. Here the total output is

$$X = \begin{bmatrix} 28 \\ 36 \end{bmatrix}$$

The correctness of the values in A may be verified by noting that

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \times \begin{bmatrix} 28 \\ 36 \end{bmatrix} = \begin{bmatrix} 20 \\ 25 \end{bmatrix}$$

where $\begin{bmatrix} 20 \\ 25 \end{bmatrix}$ represents the internal needs of R and S.

If the demand vector is

$$D_0 = \begin{bmatrix} 8 \\ 11 \end{bmatrix}$$

then for production to equal consumption, we must have

Internal needs + Consumer demand = Total output ..... (*)

In terms of the input - output matrix A, the total output X, and the demand vector D_0, (*) becomes

$$A \quad X \quad + D_0 \quad = \quad X$$

Again the correctness of this result may be verified since for the demand vector D_0, we know the output is

$$X = \begin{bmatrix} 28 \\ 36 \end{bmatrix}$$

To find the total output X, required to achieve a future demand

$$D_2 = \begin{bmatrix} 20 \\ 30 \end{bmatrix}$$

we need to solve for X in

$$A X + D_2 = X$$

Simplifying we have

$$(I - A)\ X = D_2$$

Solving for X we have

$$X = (I - A)^{-1}\ D_2$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{6} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}^{-1} \times \begin{bmatrix} 20 \\ 30 \end{bmatrix}$$

$$= \frac{24}{5} \begin{bmatrix} \frac{1}{2} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \times \begin{bmatrix} 20 \\ 30 \end{bmatrix}$$

$$= \frac{24}{5} \begin{bmatrix} 15 \\ 20 \end{bmatrix} = \begin{bmatrix} 72 \\ 96 \end{bmatrix}$$

Hence the total output of R and S for the forecast D_2 is

$$X_1 = 72, \quad X_2 = 96.$$

21. Given the following transaction matrix, find the gross output to meet the final demand of 200 units of Agriculture and 800 units of Industry.

Producing Sector	Purchasing Sector		Final Demand
	Agriculture	Industry	
Agriculture	300	600	100
Industry	400	1200	400

Solution:

Producing Sector	Purchasing Sector		Final	Total
	Agriculture	Industry	Demand	Output
Agriculture	300	600	100	1000
Industry	400	1200	400	2000

The input - output coefficients can be obtained as follows:

$$a_{11} = \frac{300}{1000} = \frac{3}{10} \quad , \quad a_{12} = \frac{600}{2000} = \frac{3}{10}$$

$$a_{21} = \frac{400}{1000} = \frac{2}{5} \quad , \quad a_{22} = \frac{1200}{2000} = \frac{3}{5}$$

The technology matrix is

$$A = \begin{pmatrix} \frac{3}{10} & \frac{3}{10} \\ \frac{2}{5} & \frac{3}{5} \end{pmatrix}$$

$$(I - A) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{3}{10} & \frac{3}{10} \\ \frac{2}{5} & \frac{3}{5} \end{pmatrix} = \begin{pmatrix} \frac{7}{10} & -\frac{3}{10} \\ -\frac{2}{5} & \frac{2}{5} \end{pmatrix}$$

$$|I - A| = \frac{7}{10} \times \frac{2}{5} - \left(-\frac{2}{5}\right) \times \left(-\frac{3}{10}\right) = \frac{8}{50} = \frac{4}{25}$$

$$(I - A)^{-1} = \frac{25}{4} \begin{pmatrix} \frac{2}{5} & \frac{3}{10} \\ \frac{2}{5} & \frac{7}{10} \end{pmatrix}$$

Now $X = (I - A)^{-1} D$

$$\Rightarrow \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (I - A)^{-1} D = \frac{25}{4} \begin{pmatrix} \frac{2}{5} & \frac{3}{10} \\ \frac{2}{5} & \frac{7}{10} \end{pmatrix} \begin{pmatrix} 100 \\ 400 \end{pmatrix}$$

$$= \frac{25}{4} \begin{pmatrix} 160 \\ 320 \end{pmatrix} = \begin{pmatrix} 1000 \\ 2000 \end{pmatrix}$$

which verifies the given data.

The new demand vector is $D = \begin{pmatrix} 200 \\ 800 \end{pmatrix}$

Then

$$X = (I - A)^{-1} D = \frac{25}{4} \begin{pmatrix} \frac{2}{5} & \frac{3}{10} \\ \frac{2}{5} & \frac{7}{10} \end{pmatrix} \times \begin{pmatrix} 200 \\ 800 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \frac{25}{4} \begin{pmatrix} 320 \\ 640 \end{pmatrix} = \begin{pmatrix} 2000 \\ 4000 \end{pmatrix}$$

Hence the Agriculture and Industry sector must produce 2000 and 4000 units to meet the final demand.

4.11 Summary:

Matrices play an important role in quantitative analysis of managerial decisions. They also provide very convenient and compact methods of writing a system of linear simultaneous equations and methods of solving them. These tools have also become very useful in all functional areas of management. Another distinct advantage of matrices is that once the system of equations can be set up in matrix form, they can be solved quickly using a computer.

A number of basic matrix operations (such as matrix addition, subtraction, multiplication) were discussed in this unit. This was followed by a discussion on matrix inversion and procedure for finding matrix inverse. Number of examples were given in support of the above said operations and inverse of a matrix.

Finally, two important applications of matrix algebra - predicting market shares using Markov models and predicting the effect of a change in the output (or demand) of one sector of the economy on the output of the other sectors, using input - output models were discussed.

4.12 Technical Terms:

Co - factor : The number $C_{ij} = (-1)^{i+j} M_{ij}$ is called the co factor of element a_{ij} in A.

Determinant : A unique scalar quantity associated with each square matrix, adjoint of a matrix inverse of a matrix, Transpose matrix.

4.13 Exercise:

1. Show that

$$\begin{vmatrix} 3 & -7 \\ 8 & 6 \end{vmatrix} = 74, \quad \begin{vmatrix} 1 & 2 \\ 3 & 6 \end{vmatrix} = 0, \quad \begin{vmatrix} x & y \\ -1 & 1 \end{vmatrix} = x + y, \quad \begin{vmatrix} -4 & 2 \\ -3 & -4 \end{vmatrix} = 22.$$

2. (a) Show that

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} \begin{vmatrix} 5 & 6 \\ 7 & 8 \end{vmatrix} = 4$$

(b) Show that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} b & q \\ p & c \end{vmatrix} + \begin{vmatrix} p & d \\ a & q \end{vmatrix} = 0$$

3. Show that

$$\begin{vmatrix} 1 & 0 & 2 \\ 1 & 2 & 5 \\ 6 & 8 & 0 \end{vmatrix} = -48, \quad \begin{vmatrix} 3 & 4 & 8 \\ 2 & 1 & 3 \\ 7 & -2 & 0 \end{vmatrix} = 14.$$

4. Show that

$$\begin{vmatrix} 3 & 4 & 7 \\ 2 & 1 & 3 \\ -5 & -1 & 2 \end{vmatrix} = -40$$

5. Show that

$$\begin{vmatrix} 1 & 2 & 3 \\ a & -a & b \\ -a & 0 & -b \end{vmatrix} = ab - 3a^2$$

6. Show that

$$\begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix} = abc - af^2 - bg^2 - ch^2 + 2fgh$$

7. Evaluate the following:

$$\begin{vmatrix} x & 1 & 2 \\ 2 & x & 2 \\ 3 & 1 & x \end{vmatrix}, \begin{vmatrix} 1 & a & a^2 \\ 0 & 1 & 2a \\ 1 & b^1 & b^2 \end{vmatrix}, \begin{vmatrix} 1^2 & 2^2 & 3^2 \\ 2^2 & 3^2 & 4^2 \\ 3^2 & 4^2 & 5^2 \end{vmatrix}$$

8. Show that

$$\begin{vmatrix} 2 & 45 & 55 \\ 1 & 92 & 32 \\ 3 & 68 & 87 \end{vmatrix} = 54$$

(Hint: Apply $R_1 - 2R_2$, $R_3 - 3R_2$ and expand)

9. (a) Prove that

$$\begin{vmatrix} 1 & 1 & 1 \\ a & b & c \\ bc & ca & ab \end{vmatrix} = (b-c)(c-a)(a-b)$$

(b) Show that

$$\begin{vmatrix} 1 & x & y+z \\ 1 & y & z+x \\ 1 & z & x+y \end{vmatrix} = 0$$

10. Find the value of

$$\begin{vmatrix} 1 & \omega & \omega^2 \\ \omega & \omega^2 & 1 \\ \omega^2 & 1 & \omega \end{vmatrix}$$, where ω is cube root of unity.

(Hint: $1 + \omega + \omega^2 = 0$)

11. Show that

$$\begin{vmatrix} a-b & b-c & c-a \\ b-c & c-a & a-b \\ c-a & a-b & b-c \end{vmatrix} = 0$$

12. Prove that

$$\begin{vmatrix} a & b & c \\ a-b & b-c & c-a \\ b+c & c+a & a+b \end{vmatrix} = a^3 + b^3 + c^3 - 3abc$$

(Hint: Apply $c_1 + c_2 + c_3$)

13. $$\begin{vmatrix} a-b-c & 2a & 2a \\ 2b & b-c-a & 2b \\ 2c & 2c & c-a-b \end{vmatrix} = (a+b+c)^3$$

(Hint : Apply $R_1 + R_2 + R_3$)

14. $$\begin{vmatrix} 1+a & 1 & 1 \\ 1 & 1+b & 1 \\ 1 & 1 & 1+c \end{vmatrix} = abc\left(1+\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right)$$

15. Show that

$$\begin{vmatrix} x-y & 1 & x \\ y-z & 1 & y \\ z-1 & 1 & g \end{vmatrix} = \begin{vmatrix} x & 1 & y \\ y & 1 & z \\ z & 1 & x \end{vmatrix}$$

16. Show that

$$\begin{vmatrix} a^2 & 2ab & b^2 \\ b^2 & a^2 & 2ab \\ 2ab & b^2 & a^2 \end{vmatrix} = (a^3 + b^3)^2$$

17. Find the adjoint of the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 3 & -5 \end{pmatrix}$$

Verify $A\,(\text{Adj } A) = (\text{Adj } A)\,A = |A|\,I_2$

18. Find the adjoint of the matrices

(i) $A = \begin{pmatrix} 1 & 4 & 5 \\ 3 & 2 & 6 \\ 0 & 1 & -3 \end{pmatrix}$, (ii) $\begin{pmatrix} 1 & 0 & -1 \\ 3 & 4 & 5 \\ 0 & -6 & 7 \end{pmatrix}$

and verify that $A\,(\text{Adj } A)\,A = |A|\,I_3$

19. If $A = \begin{pmatrix} -1 & -2 & -2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix}$ show that $\text{Adj } A = 3A'$

20. If $A = \begin{bmatrix} 1 & 2 & 1 \\ 5 & 2 & 3 \\ 1 & 1 & 2 \end{bmatrix}$, verify $A\,(\text{Adj } A) = |A| \cdot I = (\text{Adj } A) \cdot A$

21. Use the Cramer's rule to solve the following equations:

(a) $x_1 - 2x_2 + x_3 = 3$

$2x_1 + x_2 + 3x_3 = 12$

$x_1 + x_2 + x_3 = 6$

(b) $3x_1 - x_2 - x_3 = 10$

$2x_1 + x_2 + 3x_3 = 12$

$2x_1 + x_2 - 2x_3 = 5$

22. Solve the following equations by matrix inversion method

$x + 2y - z = 5$

$3x - y + 2z = 9$

$5x + 3y + 4z = 15$

23. Solve completely by using Cramer's rule

$$2x - 3y = 3 \ , \ 4x - y = 11.$$

Solve following system of equations by Gauss Jordan Method

$$2x - 3y + 5z = 11$$

$$5x + 2y - 7z = -12$$

$$-4x + 3y + z = 5$$

24. The prices of 3 commodities A, B and C in a shop are Rs. 5, Rs. 6 and Rs. 10 respectively. Customer X buys 8 units of A, 7 units of B and 6 units of C . Customer Y buys 6 units of A, 7 units of B and 8 units of C. Show in matrix notation, the prices of the commodities, quantities bought and the amount spent.

25. Two types of food, 1 and 2 have a vitamin content in units per kg given by the following table:

	Vitamin A	**Vitamin B**
Food 1	3	7
Food 2	2	9

Express the vitamin content of 5 kg of food 1 and 6 kg of food 2 as a matrix product and evaluate it. If food 1 costs 30 paise per kg and food 2 costs 35 paise per kg, express the cost of 5 kg, 6 kg of foods 1, 2 respectively as a matrix product and evaluate it.

[Hint: $\begin{pmatrix} 5 & 6 \end{pmatrix} \begin{pmatrix} 3 & 7 \\ 2 & 9 \end{pmatrix} = \begin{pmatrix} 27 & 89 \end{pmatrix}$, i.e., 27 units of vitamin A and 89 units of vitamin B. $\begin{pmatrix} 5 & 6 \end{pmatrix} \begin{pmatrix} 30 \\ 35 \end{pmatrix} = \begin{pmatrix} 360 \end{pmatrix}$ i.e. the cost is Rs. 360]

26. A motor corporation has two types of factories each producing buses and trucks. The weekly production figures at each type of factory are as follows:

	Factory A	**Factory B**
Buses	20	30
Trucks	40	10

The corporation has 5 factories A and 7 factories B. Buses and trucks sell at Rs. 50,000 and Rs. 40,000 respectively. Express in matrix from, and hence evaluate:

(i) The total weekly production of buses and trucks.

(ii) The total market value of vehicles produced each week.

(Ans: (i) $\begin{pmatrix} 20 & 30 \\ 40 & 10 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 310 \\ 270 \end{pmatrix}$ i.e., 310 buses, 270 trucks

(ii) $\begin{pmatrix} 50000 & 40000 \end{pmatrix} \cdot \begin{pmatrix} 20 & 30 \\ 40 & 10 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix}$

$= \begin{pmatrix} 50000 & 40000 \end{pmatrix} \cdot \begin{pmatrix} 310 \\ 270 \end{pmatrix} = \begin{pmatrix} 2,63,00,000 \end{pmatrix}$

i.e., the total weekly value = Rs. 2,63,00,000)

27. In a certain coal mine, the amounts of Grade 1 and Grade 2 coal (in tonnes) obtained per shift from each of two teams, A and B are given by the following table:

	Grade 1	**Grade 2**
Team A	4000	2000
Team B	1000	3000

Team A has worked 5 shifts per week and team B has worked 4 shifts per week. Grade 1 coal sells at Rs. 9 per tonne and Grade 2 coal sells at Rs. 8 per tonne. Find:

(i) The total amount of coal mined each week,

(ii) The market value of the coal mined each shift,

(iii) The market value of the coal mined each week.

[Ans: (i) (24,000 ; 22,000) tons of Grade 1 and Grade 2 respectively.

(ii) $\begin{pmatrix} 52.000 \\ 33,000 \end{pmatrix}$ (iii) $\left[\begin{pmatrix} 5 & 4 \end{pmatrix} \begin{pmatrix} 4,000 & 2,000 \\ 1,000 & 3,000 \end{pmatrix} \begin{pmatrix} 9 \\ 8 \end{pmatrix} \right]$

28. A builder develops a site by building 9 houses and 6 bungalows. On the average one house requires 16,000 units of materials and 2,000 hours of labour; one bungalow requires 50,000 units of materials and 4,800 hours of labour. Labour costs Rs. 5 per hour and each unit of material costs on the average Rs. 10 Express in matrix form and hence evaluate:

(i) The total materials and labour used in completing the site

(ii) The cost of building a house and a bungalow

(iii) The total cost of developing the site

[Ans: (i) $\begin{pmatrix} 9 & 6 \end{pmatrix} \begin{pmatrix} 16000 & 2000 \\ 50000 & 4800 \end{pmatrix}$

(ii) $\begin{pmatrix} 16000 & 2000 \\ 50000 & 4800 \end{pmatrix} \begin{pmatrix} 10 \\ 5 \end{pmatrix}$

(iii)

29. Two television companies TV_1 and TV_2 both televise documentary programmes and variety programmes. TV_1 has two transmitting stations and TV_2 has three transmitting stations. All stations transmit different programmes. On an average the TV_1 stations broadcast 1 hour of documentary and 3 hours of variety programmes each day, whereas each TV_2 station broadcasts 2 hours of documentary and $1\frac{1}{2}$ hours of variety programmes each day. The transmission of documentary and variety programmes costs approximately Rs. 50 and Rs. 200 per hour respectively. Express in matrix form and hence evaluate:

(i) The daily cost of transmission from each TV_1 and TV_2 station.

(ii) The total number of hours daily which are devoted to documentary and to variety programmes by both corporations.

(iii) The total daily cost of transmission incurred by both corporations.

$$\left[\text{Ans: (i)} \begin{pmatrix} 1 & 3 \\ 2 & 1\frac{1}{2} \end{pmatrix} \times \begin{pmatrix} 50 \\ 200 \end{pmatrix} = \begin{pmatrix} 650 \\ 400 \end{pmatrix}\right.$$

i.e., Rs. 650, Rs. 400 per day respectively for each TV_1, TV_2 station.

(ii) $\begin{pmatrix} 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 1\frac{1}{2} \end{pmatrix} = \begin{pmatrix} 8 & 10\frac{1}{2} \end{pmatrix}$

i.e., 8 hours documentary and $10\frac{1}{2}$ hours variety.

(iii) $\text{Rs.} \begin{pmatrix} 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 1\frac{1}{2} \end{pmatrix} \begin{pmatrix} 50 \\ 200 \end{pmatrix} = \text{Rs} \cdot \begin{pmatrix} 2 & 3 \end{pmatrix} \begin{pmatrix} 650 \\ 400 \end{pmatrix} = \text{Rs} \cdot 2{,}500\left.\right]$

30. A firm produces five qualities of its product which needs the following materials:

Quality	Materials Needed			
	M_1	M_2	M_3	M_4
A_1	6	6	10	8
A_2	3	4	12	6
A_3	4	5	15	8
A_4	2	2	12	5
A_5	3	2	10	

If the firm has to produce, respectively, 3, 22, 20, 12 and 7 units of the five qualities find the amounts of different materials required by writing their requirements as a row vector.

[Ans: (169, 194, 658, 324)]

31. A publishing house has two branches. In each branch, there are three offices. In each office, there are 6 peons, 8 clerks and 10 typists. In one office of a branch, 12 salesmen are also working. In each office of other branch 4 head - clerks are also working. Using matrix notation find (i) the total number of posts of each kind in all the offices taken together in each branch, (ii) the total number of posts of each kind in all the offices taken together from both branches.

32. $$A = \begin{array}{c} \\ I \\ II \\ III \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \\ 14 & 16 & 18 \end{pmatrix} \end{array}, \quad B = \begin{pmatrix} 4 & 6 & 8 \\ 10 & 12 & 14 \\ 16 & 18 & 20 \end{pmatrix}, \quad C = \begin{pmatrix} 6 & 10 & 14 \\ 18 & 32 & 26 \\ 30 & 34 & 38 \end{pmatrix}.$$

Matrix A shows the stock of 3 types of items I, II, III in three shops A_1, A_2, A_3. Matrix B shows the number of items delivered to three shops at the beginning of a week. Matrix C shows the number of items sold during that week. Using matrix algebra, find

(i) The number of items immediately after the delivery

(ii) The number of items at the end of the week

33. The following matrix gives the vitamin content of food items, in conveniently chosen units:

$$\begin{array}{c} \text{Vita min :} \\ \text{Food I} \\ \text{Food II} \\ \text{Food III} \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} \cdot 5 & \cdot 5 & 0 & 0 \\ \cdot 3 & 0 & \cdot 2 & \cdot 1 \\ \cdot 1 & \cdot 1 & \cdot 2 & \cdot 5 \end{pmatrix} \end{array}$$

If we eat 5 units of food I, 10 units of food II, and 8 units of food III, how much of each types of vitamin we have consumed? If we pay only for the vitamin content of each food, paying 10 paise, 20 paise, 25 paise, 50 paise respectively for units of the four vitamins, how much does a unit of each type of food costs? Compute the total cost of the food eaten.

$$\left[\text{Ans: } \begin{pmatrix} 6\cdot3 & 3\cdot3 & 3\cdot6 & 5\cdot0 \end{pmatrix};\ \begin{bmatrix} 15 \\ 13 \\ 33 \end{bmatrix};\ \text{Rs. } 4\cdot69\right]$$

34. A manufacturing unit produces three types of products A, B, C. The following matrix shows the sale of products in two different cities.

$$\begin{pmatrix} \text{A} & \text{B} & \text{C} \\ 1200 & 900 & 600 \\ 900 & 600 & 300 \end{pmatrix}$$

If cost price of each product A, B, C is Rs. 1000; Rs. 2000; Rs. 3000 respectively and selling price Rs. 1500 ; Rs. 3000 ; Rs. 4000 respectively, find the total profits using matrix algebra only.

35. The production of a book involves several steps, first it must be set in type then it must be printed and finally it must be supplied with covers and bound. Suppose that type setter charges Rs. 6 per hour, paper costs $\frac{1}{4}$ paisa per sheet, that the printer charges 11 paise for each minute that his press runs, that the cover costs 28 paise, and the binder charges 15 paise to bind each book. Suppose now that a publishers wishes to print a book that requires 300 hours of work by the typesetter, 220 sheets of paper per book and five minutes of press time per book.

4.14 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics for Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. S.V.S. GIRIJA

Lesson - 5

COLLECTION OF DATA

Objectives:

After studying this lesson you should be able to:

- Importance of data
- Difference between primary and secondary data
- Know how collect the data
- Questionnaire preparation
- Editions of primary and secondary data
- Significance of census and samples

Structure:

5.1 Introduction:

The intelligent use of the appropriate tolls can reduce an otherwise highly complex problem to one of manageable dimension. Of course, the use of analytical tools, presupposes complete grip of the procedure involved and assumptions underlying each of such techiniques. An appropriate modelling adoptable to any dynomic environment would be helpful to management to supplement their 'native' decision making process with more quantitative scientific analysis and planning. The

primary role of statistics is to provide decision makers with methods for obtaining and analysing information to help make decisions. However, to take a decision in any business situation we need data. Facts about different dimensions of business when expressed in quantitative form can be called as data.

5.2 Primary and Secondary Data:

Based on the source of availability of data the data is generally classified.

The collection of data refers to purpositive gathering of information relevant to the requirement of decision making when data are collected for an investigation by actual observation, measurements and direct recording it is called 'Primary Data'. Data which are not collected originally by the investigator but are obtained from other sources published or unpublished then the data is called "Secondary Data". The secondary data constitute the chief material on which statistical work is carried out in many investigations. In fact before collecting primary data it is desirable that one should go through the existing literatures and learn what is already known of the general area in which the specific problem falls. This can help in getting an idea about possible pitfalls and avoiding duplication of efforts and waste of resources. The process of gathering primary data is called 'Collection' of statistics and the process of gathering secondary data from different published sources is known as compilation of statistics.

The difference between primary and secondary data is only of degree. Data which are primary in the hands of one become secondary in the hands of others. Thus data are primary in the hands of collecting agency whereas for the rest of the world they are secondary. Secondary data offers the following advantages

(i) It is highly convenient to use information which someone else has collected. Thus there is no need to setup the organisation of data collection & Editing, tabulating the data collected.

(ii) Secondary data can be obtained with less effort and less cost.

(iii) One some subjects it may not be possible to collect primary data like, census data cannot be collected by an individual or research organisation, but can only be collected from government publication.

5.3 Methods of collecting primary data:

Primary data may be obtained by applying any of the following methods:

i. Direct personal interview

ii. Indirect oral investigation

iii. Information from correspondents

iv. Mailed questionnaire method

v Schedules sent thorugh enumerators

(i) Direct Personal Interviews: As the name suggests the investigator collects the information personally from the source concerned. It is necessary that in such case the investigator has a sense of observation, polite, courteous and has the knowledge of the language the respondent speaks.

(ii) Indirect Oral investigation: In this method if collecting data the investigator contacts third parties or witnesses capable of supplying them necessary informations.

This method is generally used in those cases where the information to be obtained is of complex nature, the area to be covered is vast or the informants are reluctant to respond if approached directly. For example in an enquiry regarding addiction to durgs, the addicts may nt be inclined to respond. In such case the investigator has to approach their friends, neighbours, relatives, dealers of drugs etc. In a similar manner clues about thefts and murders are obtained by the police by interrogating third parties who are supposed to have knowledge about the problem under investigation. Enquiry committees and commissions appointed by government generally adopt this method . This Success of this method depends upon the character of persons who are interviewed and of efficiency of the investigator who interviews such persons. The correctness of the information obtained depends on a number of factors like:

1. The knowledge of the informants. If the people who are contacted do not know the full fact of the problem under investigation, the purpose of investigator would not be served by this method.

2. Nature of the informant. If the informats are biased or prejudiced also correct information cannot be collected.

3. The honesty of the interviewers who are collecting the information. The informations may be twisted because of bribery, nepotism, threatenings etc. As a result correct conclusion cannot be arrived at.

4. The capability of the respondent expressing himself correctly and giving the true account.

When this method is employed for obtaining information. The investigator should not depend upon the information supplied by one man but a number of persons should be interrogated.

The major merits of this method are:

1. **Wide Scope:** This method may be employed for collecting data, where the scope of enquiry is wide. Most commissions and committees adopt this method as they have to conduct an extensive enquiry.

2. **Saving of Time and Money:** This method results in savings of time and cost as those persons only are interviewed who know the full fact.

3. **Secret Informations:** This method is suitable in those cases where informants hesitate to supply informations when contacted directly.

The main limitations are:

1. **It Consumes Time:** This method is quite time-consuming if the people refuse to cooperate or supply needed information and the investigator has to convince the persons to supply informations.

2. **Personal Bias:** This method is exposed to bias of both sides of the interview. The informant may not be able to express himself as well as the interviewer may be biased and he may interview those people who may be known to him.

3. **Poor Investigators:** The success of this method depends upon the qualities of investigator. If the investigators are ill trained they cannot be justice to their job.

4. **Suitability:** This method is suitable when the enquiry is exhaustive in nature and where the indirect source of information are required to be tapped either because direct source does not exist or cannot be relied upon.

The success of this method depends to a large extent on the character and efficiency of the investigator. The investigator should be polite, tactful and must have full knowledge of the time of enumeration, the area to be covered, the persons to be interviewed and the meaning of the questions to be asked to the informants.

The major advantages of this method are:

1. **Uneducated Respondents:** it is the most satisfactory method of data collection where the respondents are uneducated. The investigator can explain the purpose of the enquiry and the meaning of the questions to illiterate persons.

2. **Accurate Information:** The information received is more reliable, as the accuracy of the statements can be checked by the enumerator with the help of supplementary questions, whenever necessary.

3. **More Response:** This method eliminates to a great extent the problem non response as the enumerators go personally to obtain information. This respondends respond well as they are interviewed at their residence and also according to their own convenience.

4. **Supplementary Information:** The investigator may also collect supplementary informations which may be employed in the analysis of related problems.

This method is exposed to various limitations as follows:

1. **Expensive:** This method is quite expensive because a large number of investigators are to be employed for data collection. Usually the enumerators are paid persons.

2. **Untrained and Careless Investigators:** If the investigators are not properly trained or not trained at all and are careless, then they would collect unwanted data which may do more harm than good to the enquiry.

3. **Time Consuming:** This method is time consumingas the investigation have to go to the informats personally and if they are not available and their place the investigator has to make repeated visits.

4. **Biased Investigators:** The investigator might be a biased one and man not enter the answers given by the respondents truthfully. He may twise or suppress the information provided by the informants.

5. **Variation in the Answer Obtained:** Where there are many enumeration they may interpret various terms in the questionnaire according to the own understanding. also the variation in the personalities of this interviewers will cause variation in the answer obtained.

6. **Suitability:** This method is most suitable where finance and trained enumerators are available to cover a wide field and where some significant is attached to the accuracy of results obtained.

In order to get accurate data by this method it is advisable to keep the following points in mind:

1. The questionnaire should be pretested to find out if the questions are proper and will be answered int he desired manner.
2. The enumerators should be properly trained, they should know the exact scope of the enquiry. They should be courteous and explain the object of enquiry with patience.
3. The questionnaire or schedules duly filled in should be scrutinised to detect any apparent inconsistency in the information provided by the respondent.

(iii) Information from correspondents: The investigator collects the data through local agents or correspondents in different parts of the field of enquiry under this method. Thecorrespondents submit the informations. collected through the agents, to the central office where the data are processed. This method is generally adopted by newspapers or periodicals and also various departments of government in those cases where regular information is to be collected from a wide area. For example in construction of wholesale price index regular prices of different commodities are regularly obtained from correspondents appointed in different areas. This method is particularly sutable for crop estimates.

The main advantages of this method are:

1. Cheap and appropriate for extensive investigation.
2. Quick results - It gives rough and approximate results very quickly where high degree of precision is not necessary.
3. Saves from botheration - this saves from the botherations usually associated with statistical investigation of other types.

The important limitations are:

1. Personal Bias - This method is susceptible to personal bias. The data are collected by the correspondents in their own fashion and according to their own likings.
2. Not reliable - As the data collected by the correspondents may be prejudiced, it may not always ensure accurate results.
3. Suitability - This method is generally suitable for the cases where information needed is of regular nature and is collected from a wide area. In addition it suits where the rough and approximate estimates are desired.

(iv) Mailed Questionnaire: Collection of data through questionnaire is one of the most popular methods used these days. Under this method a list of questions pertaining to the survey (Known as questionnaire) are prepared and is sent by post to persons from whom the inforamtions are to be obtained. The questionnaire contains questions and provides space for answer. The informants send back the duly filled questionnaire within the stipulated time mentioned in the covering letter sent with the questionnaire. The success of the method depends upon the skill with which the questionnaire is drafted and the extent to which the informants are willing to co-opearate. Since the questions answers are sought through correspondence it lacks personal contact. Thus the form and tone of the questionnaire

should be so designed to supply as far as possible the missingpersonal element. Because of thelegal or administrative sanction information required by the government departmetns are obtained through this method comparatively easily. In other cases it is necessary to take informants into confidence so that they furnish correct informations. This method of data collection is satisfactory only in the cases like:

1) When the respondent has interest in the enquiry.

2) When the respondent is under legal compulsion to supply the information.

3) When the questionnaires are sent by an association to its members.

4) When the questionnaire is accompanied with some gifts.

To make the method more effective the following points can be considered.

1. The questionnaire should be so prepared that the respondents should not feel it as a burden or taxing affair.

2. The sample should be large to neutralise the chances of non-response.

3. Prepaid postage stamps should be affixed so that respondents can mail the questionnaire without spending anything.

4. It should be adopted in the cases where there is under legal compulsion to supply the information.

5. It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest inthe enquiry.

The principle merits of this method are:

1. Wide Scope: It can be employed where the scope of enquiry is very wide.

2. Convenient: The person giving the information can full-in the questionnaire at his convenience without being disturbed by the investigator at an inconvenient time.

3. Less Expensive: It is the least expensive method of data collection. Therefore most research workers and private organisations adopt this method.

4. Confidential Informations: Confidential informations may be given on a postal questionnaire which the informant may show hesitation to respond when he is contacted personally.

5. Expeditious: This method of collecting data is much more expeditious provided the respondents are willing to part with information and respond timely.

This method is exposed to the following limitations:

1. This method presumes that the informants are literate and understand the language of the questionnaire. This limits the scope of this method to certain investigations only.

2. Non response is a more serious problem in case of postal questionnaire. A large part of sample taken may not answer the questionnaire. Thus it involves uncertainty about the response and cooperation on the part of respondents may by difficult to presume.

3. There are every possibility that respondents may not understand the meaning of the questions and may supply wrong answers and it may be difficult to verify.

4. Suitability: This method is appropriate where informants are spread over a wide area i.e. in case of extensive survey and when the respondnets are educated and trustworthy.

(v) Schedule Sent Through Enumerators: Another method of collecting information is that of sending schedules through the enumerator or interviewer. Schedule is the name usually applied to a set of questions. Which are asked and filled in a face to face situation with another person, i.e. by the interviewer. The essential difference between the mailed questionnaire method and this method is that inthe former case the questionnaire is sent to the informants by post and in the latter the enumerator carries the schedule personally to the informant. The success of this method depends upon the character and the efficiency of the investigator. To get maximum information about the problem under study the investigator needs acquaint himself with local customs conditions and tradition so that he is in a position to identify himself with the personal from whom the information is sought.

Following are the advantages of Direct Personal Interview:

(i) Encouraging Response: Response is more encouraging as the investigation approaches the informants personally and most people hesitate least to part with information, when approached personally.

(ii) Accurate Data: As the nature of enquiry is intensive and conducted personally, results obtained by this method are generally accurate and reliable. The investigator can clear up doubts of the informants about certain questions. In case the interviewers apprehends that informant is not giving accurate information, he may cross examine him and thereby try to obtain the information.

(iii) Supplementary Informations: It is possible through personal interview to colelct supplementary informations about the informants personal characteristics and environment which may be employed in the analysis and interpretation of results.

(iv) Handling Delicate Situations: The senstive informations can be collected carefully and tactfully and handled effectively by a personal interview than by any other method of investigation.

The important limitations of this method are:

1) Limited Scope: If the number of persons to be interviewed is large and they are spread over a wide area this method cannot be useful as it requires personal attention of the investigator.

2) Expensive: This method is very expensive because a large number of investigators have to be employed for collecting data. Individuals and small institutions cannot employ this method for data collection for wnat of funds.

3) Personal Bias: The chances of personal bias and prejudice are greater in this method which may do a lot of harm to the investigation. The personal likes and dislikes of the investigator may defeat the purposes of the plan.

4) Time Consuming: This method usually consumes more time as the interview can be held only at tbhe convenience ofthe informants. In order to familiarise with the local conditions, customs and lanugage in order to observe the phenomena properly investigator has to spend more time in the preparatory work.

5) Untrained Investigators: The interviewers have to be throughly trained and supervised, otherwise they may not be able to obtain informations. If the investigators are poorly trained or untrained they may include unwanted materials and omit required materials, thereby spoiling the entire work.

6) Suitability: Despite the above mentioned disadvantages this method is favoured for obtaining information in intensive investigations. It is also recommended in case where it is requied to get the correct and reliable information. This method gives very satisfactory results if the scope of enquiry is narrow and intensive at the same time and the investigators are dependable and unbiased.

5.4 Design in a questionnaire:

The constructions of a good schedule or questionnaire is essential for collection of primary data. The value of results depend greatly on the adquancy of questionnaire. The drafting of a questionnaire is the most difficult task and this job should be done by the experts as it is a specialised hob and needs lot of skill and experience. Though there is not hard and fast rule for preparing a questionnaire. Yet some broad characteristics are essentially to be followed for construction of any questionnaire.

1. **Get-up of the questionnaire and a covering letter:** The questionnaire should be made attractive and interesting thorugh proper presentation and layout. Every questionnaire should contain a convering letter containning the aims and objectives of the enquiry and the use that would be made of the information collected along with an appeal for seeking help and co-operation of the persons who are in a position to supply the information. The convering the letter also should contain the assurance as regards to the confidentiality of the responses of the respondents.

2. **Size of the questionnaire should be small:** Unnecesary detaild should be avioded and only relevant questions should be asked to keep the number of questions to a minimum. However the investigator should keep in mind that there are sufficient number of questions to cover the scope of study comprehensively. There is no hard and fast rule about the number of questions in a questionnaire. The precise number of questioner would depend upon the object and scipe of the investigation. Fifteen or twenty five may be regardedas a fair number.

3. **Questions should be short and simple to understand:** There should be no ambiguity in the questions. Thus questions should not be confusing and should be capable of straight answer.

4. **Personal questions should be avoided:** It is advisable as far as possible to avoid questions, which an informant may be unwilling or reluctant to answer as they may be involve discolsure of private confidential or personal informations.

5. **Questions should be in a logical order:** The questions in a questionnaire should be so arranged that while reading thorugh the questions the respondent should be able to understand

the object of the enquiry. the questions if arranged logically helps tabulation and classification of data. The questions should not slop back and forth from one topic to other. Thus it is undesirable to ask a man what brand of toothpaste he uses before asking whether he uses toothpaste are not questions supplying identification and description of the respondents should come first follwed by major information questions. Two different questions worded differently may be included in the same subject to provide cross check on important points.

6. **Instructions to the Informants:** The questionnaire should provide necessary instruction to the informants. An instruction sheet containning the operational definitions of various terms and concepts used int he questionnaire should be attached to the questionnaire. instruction as refgards the unit of measurement also should be given. Also instruction about the time within which the questionnaire duly filled in should be sent and to which address should be given.

7. **Questions should be capable of objective answers:** The qeustionaire should be so designed that wherever possible questions about opinions should be avoided. A set of possible answers may be given against each question and respondent may be asked to tick the correct answer.

8. **Questions REgarding Calculations Should be Avoided:** Questions should not require calculations to be made. Like questions necessitating calculation of ratio percentages etc. to answer a question should be avoided.

9. **Pretesting of Questionnaire:** The questionnaire should be pretested with a group before going for collection of data. The advantages of pretesting is that the shortcomings of the questionnaire can be discovered and it can be revised.

5.5 Pretesting The Questionnaire:

Ultimately, designing the perfect survey questionnaire is impossible. However, researchers can still create effective surveys. To determine the effectiveness of your survey questionnaire, it is necessary to pretest it before actually using it. Pretesting can help you determine the strengths and weaknesses of your survey concerning question format, wording and order.

There are two types of survey pretests: Participating and Undeclared.

Participating pretests ductate that you tell respondents that the pretest is a practice ryn rather than asking the respondents to simply fill out the quesionnaire, participating pretests usually involve an interview setting where respondents are asked to explain reactionsto question form, wording and order. This kind of pretest will help you determine whether the questionnaire is understandable.

When conducting an undeclared pretest, you do not tell respondents that it is a pretest. The survey is given just as you intend to conduct it for real. this type of pretest allows you to check your choice of analysis and the standardization of your survey. If researchers have theresources to do more than one pretest, it might be best to use a participatory pretest first, then an undeclared test.

General applications of Pretesting:

Whether or not you use a participating or undecalred pretest, pretesting should ideally also test specifically for question variation, meaning, task difficulty adn respondent interest and attention. Your pretests should also include any questions you borrowed from other similar surveys even if

they have already been pretested, because meaning can be affected by the particular context of your survey. Researchers can also pretest the following: flow, order, skip, patterns, timing and overall respondent well - being.

Pretesting for reliability and validity:

Researchers might also want to pretest the reliability and validity of the survey questions. To be reliable, a survey question must be answered by respondents the same way each time. Researchers can assess reliability by comparing the answers respondents give in one pretest with answers in another pretest, Then, a survey question's validity is determined by how well it measures the conept(s) it is intended to measure. Both convergent validity and divergent validity can be determined by first comparing answers to another question measuring the same concept, then by measuring this answer to the participant's response to a question that asks for the exact opposite answer.

5.6 Editing of Primary and Secondary Data:

Once the data have been collected either from primary source or secondary source they need to be scrutinised or edited to detect possible errors or irregularities. The task of editing is a highly specialised one and require great care and attention. However it should be remembered, data collected from internal records or published source is relatively simple than the data collected from a survey. While editing the data the following considerations need attention:

1. Completness
2. Consistency
3. Accuracy
4. Homogencity

The secondary data also should be scrutinised because thedata may be inadequate, inaccurate or unsuitable. How ever, it is quite difficult to verify secondary data to find inconsistency probable errors or commissions.

The investigator must decide which source of data he can use when one is deputed for collection of data one is tempted to go for secondary data because of its obvious easiness. But for more reliability of data one prefers a primary data to a secondary.

5.7 Sources of Secondary Data:

Official Statistics: Official statistics are statistics collected by governments and their various agencies, bureaus and departments. These statistics can be useful to researchers because they are an easily obtainable and comprehensive source of information that usually covers long periods of time. However, because official statistics are often "characterized by unreliability, data gaps, over-aggregation, inaccuracies, mutual inconsistencies and lack of timely reporting". It is important to critically analyze official statistics for accuracy and validity. There are several reasons why these problems exist.

1. The scale of official surveys generally requires large numbers of enumerators (interviewers0 and, in order to reach those numbers enumerators contracted are often under-skilled.

2. The size of the survey area and research team usually prohibits adequate supervision of enumerators and the research process and

3. Resource limitations (human and technical) often prevent timely and accurate reporting of results.

Technical Reports: Technical reports are accounts of work done on research projects. They are written to provide research results to colleagues, research institutions, governments and other interested researchers. A report may emanate from completed research or on - going reserach projects.

Scholarly Journals: Scholarly journals generally contain reports of original research or experimentation written by experts in specific fields. Articles in scholarly journals usually undergo a peer review where other experts in the same field review the content of the article for accuracy, orginicality and relevance.

Literature Review Articles: Literature review articles assemble and review original research dealing with a specific topic. Reviews are usually written by experts in the field and may be the first written overview of a topic area. Review articles discuss and list all the relevant publications from which the information is derived.

Trade Journals: Trade journals contain articles that discuss practical information concerning various fields. These journals provide people in these fields with information pertaining to that field or trade.

Reference Books: Reference books provide secondary source material. In many cases, specific facts or a summary of a topic is all that is included. Handbooks, manuals, encyclopedias and dictionaries are considered refererence books.

where to find secondary data:

There are numerous sources of secondary data and information. Thefirst step on collecting secondary data is to determine which institutionsconduct research on the topic area or country in question.

Large surveys and country wide studies are expensive and time consuming to conduct, therfore they are usually done by governments or large institutions with a research orientation. Thus, government documents and official statistics are a good starting place for gathering secondary data however as previously stated the quality of the documents will vary depending on the country of study and the amount of resources dedicated to data collection.

University libraries are good sources of information and should be consulted. Also, it woyld be beneficial to establish contact with experts at local university departments that are dedicated to research on the topic areas that you are intereted in (e.g. Departments of agricultural sciences, public health, economics, anthropology, and sociology). These experts can be important sources of information on on-going reserach projects as well as for guiding you forward other sources of topic area information or individuals that can be contacted of information. This is particularly true when you are searching for regional or local-level information and data. In some cases, they might also have small libraries that provide additional information.

Evaluating the quality of your information sources:

One of the advantages of secondary data review and analysis is that individuals with limited research training or technical expertise can be trained to conduct this type of analysis. Key to the process, however is the ability to judge the quality of the data or information that has been gathered. The following tips will help you assess the quality of the data.

Determine the original purpose of the data collection. Consider the purpose of the data or publication. Is it a government document or statistic, data collected for corporate and/or marketing purposes, or the output of a source whose business is to publish secondary data (e.g. research institutions). Knowing the purpose of data collection will help to evaluate the quality of the data and discern the potential level of bias.

Attempt to ascertainthe credentials of the source(s) or author(s) of the information. What are the author's or source's credentials educational background, past works/writings or experience in this area? For example, the following sources are generally considered reliable sources of data and information research reports documenting findings from agricultural research published by the FAO or IFAD socioeconomic data reported by the World Bank and survey health data reported in USAID's Demographic Health Surveys.

Does it include a mthods section and are the methods sound? Does the article have a section that discusses the methods used to conduct the study? If it does not you can assumethat it is a popular audience publication and should look for additional supporting information or data. If the research methods are discussed, review them to ascertain the quality of the study. If you are not a research methods expert have someone else in your country office review the methods section with you.

What is the date of publication? When was the source published? is the source current or out-of-date? Topic areas of continuing or rapid development such as the sciences, demand more current information.

Who is the intended audience? Is the publication aimed ata specialized or a general audience? If the source too elementary aimedat the general public?

What is the coverage of the report or document? Does the work update other sources substantiate other materials/reports that you have read or add new information to the topic area?

Is it a primary or secondary source? Primary sources are the raw material of the research process, they represent the records of research or events as first described. Secondary sources are based on primary sources. These sources analyze describe, and synthesize the primary or original source. If the source is secondary does it accurately relate information from primary sources?

Importantly, is the document or report well referenced? When data and/or figures are given, are they followed by a footnote endnote which which provides a full reference for the information at the end of the page or document or the name and date of the source (e.g. Burke 1997)? Without proper reference to the source of the information it is impossible to judge the quality and validity of the information reported.

5.8 Precautions in use of secondary data:

Since secondary data have already been obtained it is highly desirable that before the investigator uses it he should make a proper scrutiny of the data. Secondary data should not be accepted at their face value. Statistics collected by others cannot be depended fully as they may contain some pitfalls or limitations and unless they have been throughly secrutinised they should not be used. Thus before using the secondary data the investigator should confirm whether the following characteristics are present with the secondary data that is desired to be used.

1. **Suitability to the purpose of investigation:** It is essential for the analyst to satisfy whether the secondary data conforms to the purpose of study while using the same.

   If any doubt develops in the mind of the analyst about the nature and scope of secodnary data then such data should not be used. The suitability of data can be judged in the light of nature and scope of investigation.

2. **Adequacy of data for the investigation:** If the data are found suitable for the investigation they should be tested for adequacy. Adequacy of data is to be judged in the light of the rquirements of the survey and the geographical area covered by the available data. The adequacy of the data also should be judged as regard to the degree of accuracy achieved in the data.

3. **Reliability of Data:** In order to determine the reliability of the published data it is better to enquire about the collecting agency, the methodology adopted in collecting and compiling the data, sampling procedure followed, degree of accuracy achieved etc.

4. **Accuracy of Data:** Analyst should also ensure about the accuracy of the secondary data as the accuracy of conclusions drawn depends mainly on the accuracy of collected data. For judging the accuracy of colelcted data the analyst should see

   (a) Whether the data collecting agency was unbiased or biased.

   (b) Whether questionnaire were properly designed or not.

   (c) Whether the investigators were properly trained or not.

5.9 Census and Samples:

1. A manufacturer of electronic fuses needs to know the maximum amps at which a any one of a batch of fuses will burn out. Testing every one to destruction will leave none to sell, so a few of them are examined to get an idea of the properties of the batch as a whole.

2. The government of a large country needs to know what proportion of the population will be eligible for benefits in a health program directed at preventing a particular disease. It is not possible to examine everyone in the country for susceptibility to the disease (some of whom will have died and others who will have born during the examination process) so a much smaller group of individuals is examined in order to gage the likely proportion in the population that are susceptible.

3. A local authority, responsible for providing an emergency response team for car crash victims on a collection of highways, needs to know the likely monthly frequency and location of crashes

in order to establish the size of and rsources for the response team. The location and number of crashes per month in recent history is used as a guess of what the likely locations and frequencies would be.

Examining a much smaller and very real sub group of population is the sample. When every element in the population is identified and the relevant characteristic recorded the resultant data set is referred to as A Census, it constitutes a complete record of the population of interest. It is not necessarily an infinite list (the complete batch of fuses could certainly all be examined, the population of a country could certainly all be counted ata point in time) through sometimes it is (the number of places that accidents could take place in a highway system is infinite and the theoretical frequency with which they occur on average over a given period of time is certainly obscure and unobservable). When for various reasons (economic, feasibility or practical) a census cannot be taken the population of interest is something about which we can only conjecture.

The sample is something tangible that we do observe and use to explore conjectures about the population of interest i.e., we use the sample to tell us something about a population which we cannot examine directly via a census.

The simplest, most effective form of sampling is simple random sampling where in all of the elements of the sample are each independently and randomly drawn from the population. Agencies that collect data often collect "representative" samples for reasons of economy and unless great care is taken the results emerging from such samples can be misleading. Two types of representative sampling are cluster sampling which divides the population into clusters or groups and randomly selects a small set of clusters within which a complete census is taken and stratified sampling splitting the population into mutually exclusive groups or strata (by age, location, gender or profession for example) and taking a random sample from each strata. For example if one draws an equal number of people from each of the provinces in India (stratification by location) in order to calculate the average height of people in India, and if height is related to location so that the further west one goes, the taller people tend to be, then a straight average across all of the samples will misrepresent the average height in India.

5.10 Exercises:

1. Distinguish between primary and secondary data.
2. Explain the methods of collecting primary data.
3. Explain the prepartion of questionnair desigin.
4. What are the methods for testing questionnair.
5. Explain the editing of primary and secondary data.
6. What are the sources of secondary data.
7. What are the precautions in use of secondary data.
8. What is difference between census and samples.

5.11 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Wirter

Prof. M. KOTESWARA RAO

Lesson - 6

PRESENTATION OF DATA

Objectives:

After studying this lesson, you should be able to:

- Importance of classification of data
- Know the necessity of classifying data and various types of classification
- Frequency ditribution construction
- Significance of graphs

Structure:

6.1 Introduction:

Successful use of collected data is the next step in the statistical investigation to classify type data. The data contained in schedules or questionnaires are in a form which does not give an idea about the salient features of the problem under study. They are not directly fit for analysis as well. For the purpose of comparison, analysis, interpretation it is essential that data are to be put in condensed form. As long as collected data remain unorganised it is not possible to analyse the behaviour of data. Thus there arises the need to condense and simplify the raw data into such a form that the features of the data may be brought out. The procedure employed to reduce and to simplify raw data is called classification and tabulation. For the purpose of analysis and interpretation data have to be divided into homogeous graphs and are presented in a condensed form by the help of classification and tabulation of data.

6.2 Classification of Data:

The classification of data depends generally on the nature of data but more specifically on the purpose for which the data is being processed. For example, the data on the consumption of

a particular variety of fast food can be classified based on regions to find the geographical popularity and suitability.

Some common types of classifications are:

(a) Region - wise or area - wise classification

(b) Chronological classification

(c) Qualitative classification

(d) Quantitative classification

(a) Region - wise or area - wise classification: Under this classification data are arranged according to Geographical area. For example the sales executive of Aurofood, a wheat based Food Products unit would consider the sales in different regions of the country. Such classification would help him to plan the company's product distribution programme.

(b) Chronological classification: Under this classification, data would be arranged according to the time of its occurrence. For example, the ministry of Industries would be interested in understanding the Annual growth in demand for electric goods over the past 10 years before the issue of new licenses to industrialists.

(c) Qualitative classification: When data is to be arranged according to some qualitative attributes, this classification is resorted. For example, the productivity of skilled versus unskilled workers in an organisation, educational levels of children belonging to different categories of castes would be done in this way. Qualitative attributes would be of much help to an advertising executives.

(d) Quantitative classification: When data is classified according to some characterstics that can be measured, it is called quantitative classification. For example the customers of a particular variety of brand can be classified against their income levels. The quantitative data is characterised by different numerical values.

Variable: A variable in statistical terminology stands for any measurable quantity. Basing on the size of value classification of data is carried out.

(i) Discrete variable

(ii) Continuous variable

Discrete data refers to quantitative data that is limited to certain numerical values of a variable. For example, the size of ready-made garments, sizes of shoes are likely to possess certain specified values.

On the other hand, the continuous data takes all values of a variable. For example, the quantity of cloth purchased by customers, volume of sales, volume of production whose values are likely to range any length.

6.3 Objectives of Classification:

The main objectives of classifying data are:

i) Simplification of Raw Data: The data in the raw shape are difficult to understand. Through

classification homogeneous figures are grouped together, thus helping in understanding the data.

ii) **Facilitate Comparison:** The technique of classification facilitates comparison of data within the class between the classes of the some series as well as of different series.

iii) **Depicts salient features of the data:** The technique of classification throws light upon the significant features of the data and one can understand the significance of data at a glance.

iv) **Makes data more intelligible:** The process of classification differentiates the homogeneous figures from heterogeneous ones and also simplifies the statistical calculation like mean, median, mode, standard deviations etc...

v) **Saves space and time:** As the data are condensed and presented in a compact form it saves time and space.

vi) **Eliminates unnecessary details:** It eliminates the unnecessary details found in the raw data and gives prominence to the important information gathered.

vii) **Easy to interpret:** It enables the statistical treatment of the material collected and help in interpreting the data.

A Classification to be regardedas an ideal one should have the following characteristics:

1) It should be unambiguous

2) It should be stable to facilitate comparison

3) It should be flexible so as to adjust to change condition

4) It should cover the whole data

6.4 Frequency Distribution:

Tabulated Data:

When data is arranged in groups or classes accross some convenient ranges of observations, such arrangement in a tabular form in called frequency distribution. In frequency distribution the data isrepresented by distinct groups known as classes. The number of observations which fall in each class is known as class frequency. When data is described by the continuous variable it is called continuous data distribution and when it is described by discrete variable it is called discrete data distribution.

For example, the number of units of children's ready-made dress sold according to the size of the dress could be classified as discrete distribution.

An Example of Discrete Data

Size of ready-made dress	6.5	9.5	10.5	11.0	12.0	14.0
Number of units sold (Frequency):	36	20	26	50	30	30

On the other hand, if a personnel manager classifies the wages received by the skilled employees in his organisation it results incontinuous data distribution.

Wages (Rs.) Per Month	Number of Workers (Frequency)
0 - 1000	20
1000 - 1500	36
1500 - 2000	27
2000 - 3000	14
above 3000	3

a) Construction of Discrete Frequency Distribution:

The process of preparing a frequency distribution is quite easy. In case of discrete frequency distribution, countthe number of times each variable is repeated. It constitutes thefrequency. The task of counting is done traditionally with the help of tally bars. A tally bar is simply one single indication of count, which would be helpful to count the occurrence of the observation in that class of distribution. Five tally bars are shown as blocks. The frequency refers to the total number of Tally bars.

To construct a descrete frequently distribution, consider a sample of 50 customers in a large departmental store. The number of packets of noodles purchased by them are:

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

To present this data in adiscrete freqiency distribution let us construct tally bars.

No.of packets purchased	Tally bars	Frequency
1	𝍸 III	8
2	𝍸 𝍸 II	12
3	𝍸 𝍸 III	13
4	𝍸 II	7
5	𝍸 II	7
6	III	3
		50

b) Construction of a continuous frequency distribution: If a variable assumes any value within a particular range, then continuous frequency distribution is prepared. The data is divided into certain classes and then the frequency of each class is obtained. In constructing the frequency distribution for continuous data, the following terms are frequently used.

Class boundaries: The highest and the lowest values that can be included in a class.

Class-interval: Class interval represents thewidth of a class. This is nothing but difference between upper and lower limits of a class.

Frequency: The number of observations falling within a particular class.

Dtermination of Classes:

No specific rule: In determining the number of classes and class intervals for a given data distribution, the number of classes should not be too small or too large. Class width in multiples of 5s or 10s preferred.

Types of class intervals:

Open - End: There are two ways of making class intervals. One of it is open-end. In open-end distribution, the lower limit of the very first class interval and upper limit of the last class are not defined. For example, when our interest of analysis is limited to certain categories of income classes of respondents, then the extreme values whcih entail with smaller frequencies would be kept open-end.

Income Classes		
Less than	-	Rs. 1,500
Rs. 1,500	-	Rs. 2,500
Rs. 2,500	-	Rs. 3,500
Rs. 3,500	-	Rs. 4,500
Rs. 4,500	-	Above

Closed - End: Another type of class interval is the exclusive type, where in the class intervals are so arranged that the upper limit of one class is the lower limit of the next class. For example, profits earned by selected public limited enterprises can be shown as

Profits (Rs. lakhs)	No.of Enterprises
0 - 5	20
5 - 10	30
10 - 15	45
15 - 20	35
20 - 25	10
	140

Subsequent to an in-plant training to 50 apprentices in an organisation, their performances in an aptitude test are as follows:

29	30	32	37	43	54	55	47	38	62
37	39	56	54	38	49	60	37	28	27
22	21	37	33	28	42	56	33	32	59
40	47	29	65	45	48	55	43	42	40
32	33	47	36	35	42	43	55	53	48

In order to form the frequency distribution for the above data, we have to consider the maximum and minimum values and their difference and if we divide by 10, it forms 5 class intervals.

Aptitude test scores	Tally Marks	Frequency
20 - 30	𝍸 II	7
30 - 40	𝍸 𝍸 𝍸 I	16
40 - 50	𝍸 𝍸 𝍸	15
50 - 60	𝍸 IIII	9
60 - 70	III	3
		Total: 50

6.5 Presenting Data Through Graphs:

Chartings of frequency distributions which cover both diagrams and graphs are useful because they enable a quick interpretation of the data. Graphical presentations are often in the form of

(i) Pie Diagram

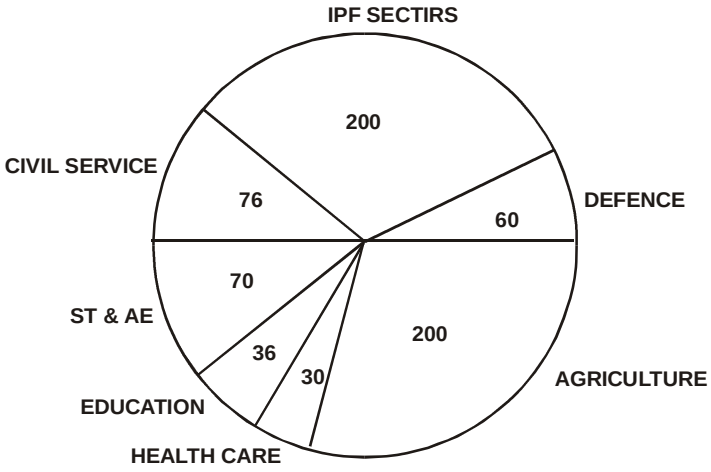
(ii) Bar Diagram or Histogram

(iii) Frequency Polygon

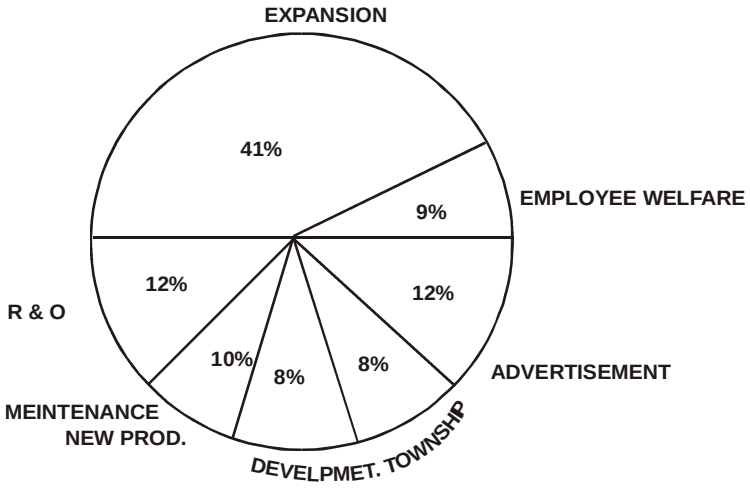
(iv) Ogive or cumulative frequency curve

(a) Pie Diagram: Pie chart is an effective way of presenting statistical data. This method is used to show how total has been divided. An entire circle, represents the total amount available and pieces are proportional to the amount of the total they represent. In order to understand the proportions of components are converted in relative terms and expressed as percentages. Select pie charts are given under.

(a) Break-Down of a Government Budget (Amount in Crores)

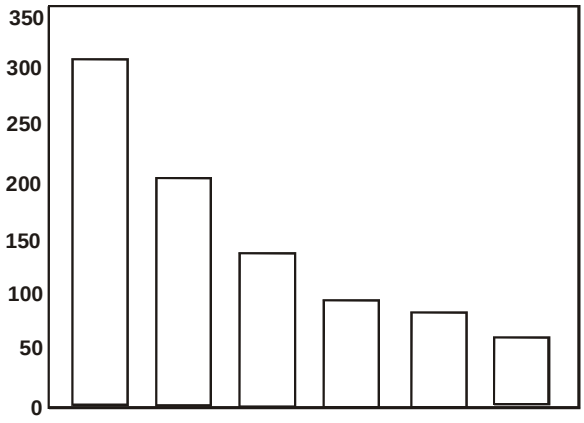


(b) Break Down of a Manufacturing firm's expenditure

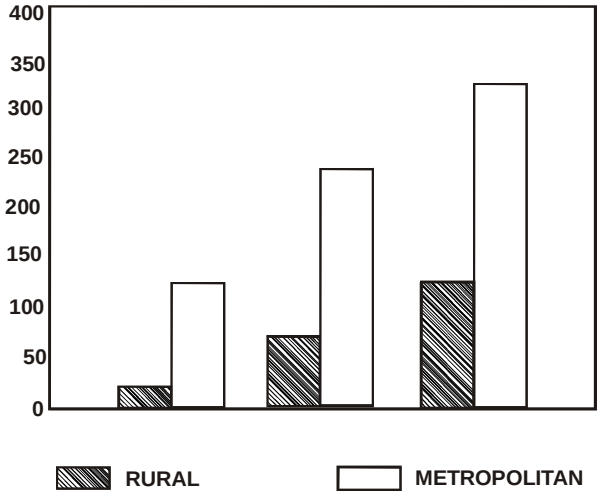


(b) Bar diagram and Histogram: Bar diagrams are the eassiest and most adaptable generalpurpose charts. Often the business and economic data are presented through bar diagrams. These pictures are unidimensional as the size of the data are presented by the length of the bar. The bars genrally are presented as equidistant rectangular bars. The width has no relationship with the measurements. In order to draw a bar diagram, take the chracteristic (or attribute) which is under considerationon X-axis and its corresponding value on Y - axis. It is a general practice to mention the value depicted by the bar on the top of the bar.

Bar Diagram for Two Variables



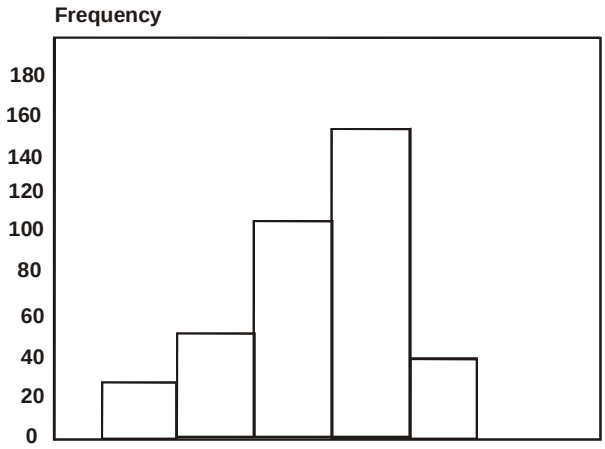
Bar Diagram for two variables



There are different varieties of Bar graphs. One such is the Multiple Bar Diagram. It is used when comparisons are to be made between two or more sets of related statistical data. To illustrate consider credit disbursed by commercial banks in Rural and Metropolitan areas.

c) Histogram: One of the popular and most commonly used methods of graphic representation of continuous frequency distribution is the histogram. It consists of a series of adjacent vertical rectangles with heights proportional to corresponding frequency of each class and width equal to the width of class interval. To construct a histogram, take class intervals of the variable on X - axis and the frequiences on Y-axis. To illustrate, let us consider the following frequency distribution.

Histogram for even class intervals

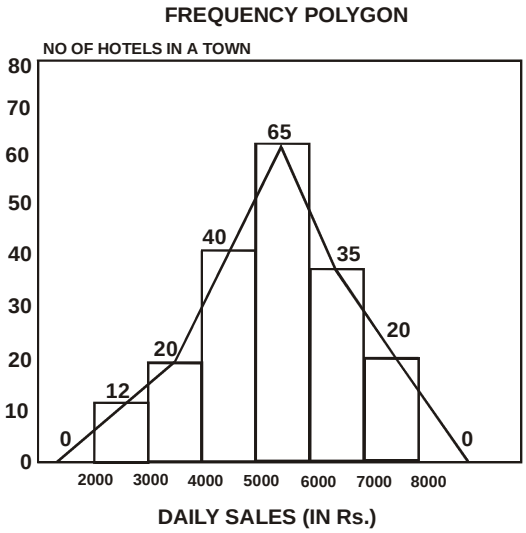


No.of Hospital Beds Used in a Year

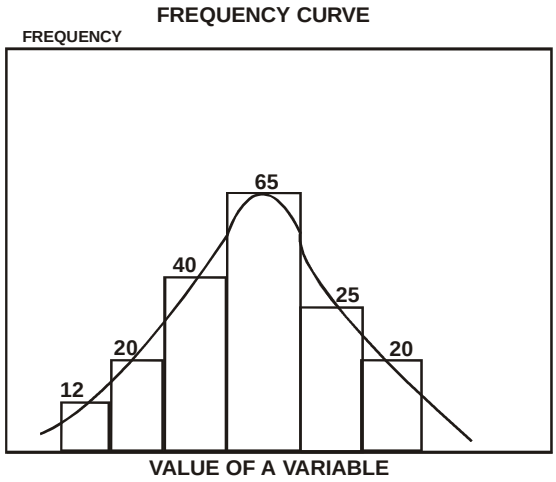
The administrator - cum - Medical - officer - in charge of a large AP based Hospital is worried about effective provision of facilities in the light of growing demand in the recent past. In order to verify the use of Hospital beds, he has classified last one year's data and presented it in the above histogram.

The above graph shows that for a large number of days 125 to 250, hospital beds are in use indicating he need for provision of the necessary facilities to such a strength of patients.

d) Frequency Polygon: Frequency polygon is a graphical representation fo frequency distribution in the form of a curve. It is constructed by taking the mid-points of upper horizontal side ofeach rectangle of a histogram. Frequency polygon is an improved version of histogram and it provides a continuous curve. It approximates at both the ends even without actual data, thus proviging and unbroken curve.

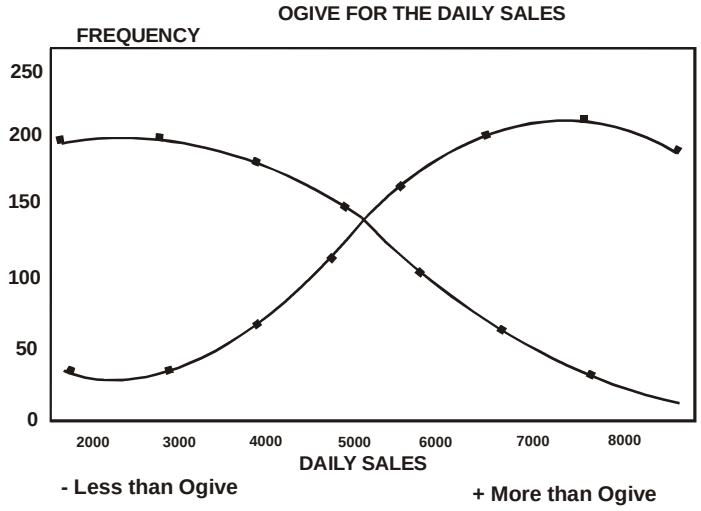


Instead of joining the mid-points of different rectangles by straight lines, if we draw a smooth curve touching all the points, such curve is known as frequency curve.



The frequency polygon and frequency curves have a clear advantage over Histograms. These curves would provide a clear picture about the symmetry of the data distribution. In turn, helps in comparing different frquency distributions.

e) Ogive: Ogive curve is a graphic representation of a cumulative frequency polygon. The ogive is less-than or more-than based on the type of cumulation of frequencies attempted to. For constructing the less-than-ogive the upper limit of various classes are taken on the X-axis and the cumulated frequencies, cumulation made from the lowest class isnterval, are plotted on Y-axis. On the other hand, a more than ogive is constructed by taking the lower limits on X axis and the cumulative frequencies, cumulated from below, are plotted on Y-axis. The shape of less than ogive would be rising and looks like elongated S and in case of more than ogive would also be elongated S but turned upside down. The ogive curve would be particularly useful for graphic computation of Median, Quartiles, Deciles. A superimposition of both Less-than-Ogive and more-than-Ogive would help us to determine the Median at their intersecting point. Further, Ogive would be useful to determine graphically the proportion of observations below or above a given value of a variable.



Let us construct the less - than and more-than-ogive for the problem of Daily Sales of Hotel establishments located in a town.

Daily Sales (Rs.)	No.of Hotel Establishments (f)
2000 - 3000	12
3000 - 4000	20
4000 - 5000	40
5000 - 6000	65
6000 - 7000	35
7000 - 8000	20
Total	**192**

Calculation of cumulative frequencies

Daily Sales (Rs.)	Cumulative Frequencies	Daily Sales (Rs.)	Cumulative Frequencies
Less than 3000	12	More than 2000	192
Less than 4000	32	More than 3000	180
Less than 5000	72	More than 4000	160
Less than 6000	137	More than 5000	120
Less than 7000	172	More than 6000	55
Less than 8000	192	More than 7000	20

6.7 Exercises:

1) You have been asked by the Director of Marketing to make a presentation at next week's Board meeting of BPL Sanyo Utilities and Appliances Ltd. The presentation concerns the company's advertising budget for the past year and the projected budget for the next year. You are given the following data.

Mode	This year's Expense (Rs.)	Next Year's Budget (Rs.)
News Papers	35,00,000	40,00,000
Television	60,00,000	80,00,000
Trade Publications	25,00,000	20,00,000
Miscellaneous	15,00,000	10,00,000

2) T.V. Quiz-time authorities wish to select a Nationalteam for participation in SAARC Quiz competitions. They have conducted an Aptitute Test for a maximum of 75 marks for 25 aspirants. Their score are as follows:

28	35	61	29	36	48	57	67	69	50	48	40	47
42	41	37	51	62	63	33	31	32	35	40	38	

Construct a less than ogive and find how many candidates can be selected if they wish to send top 10 percent aspirants.

3) Explain the purpose and methods of classification fo data giving suitable examples.

6.8 Summary:

Presentation of data is provided through tables and graphs. The frequency ditribution may show actual relative of cumulative frequency.

6.9 Reference Books

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: M. KOTESWARA RAO

Lesson - 7

MEASURES OF Central Tendency

Objectives:

After going through this lesson, you will learn:

- The concept of central tendency
- The computation of mean, median, quartiles, mode, geometric mean and Harmonic means.
- The relationship among various averages

Structure:

7.1 Introduction:

Data collected for any statistical enquiry will be generally large in size. Human mind can not grasp its significance. As a first step in statistical analysis the collected data are classified and tabulated. The process of classification and tabulation simplifies the complaexities of the data, but does not reveal the inherent characteristics of the distribution. Hence further analysis of classified and tabulated data is needed. There is a need to find out a few summary measures, which can reveal the basic features of the data.

7.2 Arithmatic Mean:

Significance of measurs of central tendency an average is a single figure which sums up the characteristics of a whole group of figures. An average is a representative figure which is

"gist", if not the substance of statistics, since an average sums up the characteristics of the whole group. Its value always lies between the minimum and maximum values. Generally it is located in the centre of middle of he distribution. As this single value (average) has a tendency to be some where at the centre and with in the range of all values, it is also known as "Measure of central tendency".

Central tendency or average, the before refers to point on the scale of values where most of the item tend to cluster.

Importance of average:

Average occupies a very important place in statistics. That is why Bowley called this. Science as a 'science of averages'. Averages are widely used than any other measures. Many other techniques of statistical analysis depend on averages. If decribe the group better than any other value in the group. If provides a concise and clear picture of the data. Properties of a good measure of central tendecny an average is a single figure which represents the data, hence it should have certain properties.

1. **It should be regidly defined:** An average should be regidly defined so that there can be no scope of any personal bias.
2. **It should be based on all observations:** An average calculated or located should be based on all the items of series, other wise it can not be said to be a representative.
3. **It should be easy to understand and simple to calculate:** The average should be simple for comprehension, easy to calculate other-wise its use is limited.
4. **It should not be affected much by extreme observations:** A few very large or very small items should not affect the values of good average.
5. **It should be capable of further a;gebraic treatment:** Beside the above a good average should be able to represent as many characteristics of the data as possible and its value should be nearest to the most of the items, should be capable of expression in obsolute numerical terms.

Objects and Functions of Average:

The following are the objectives of the study of average.

1. **To get one single value that describe the characteristics of the group:** An average is a simple and precise indicater of the central tendency of the data. The purpose of an average is to represent a group of individual values in a simple and concise manner. It aids the human mind in grasping the significance and characteristics of a mass of complete data. For example: If an average age of an Indian is 50 years it will give an idea about health conditions etc.
2. **To facilitate Comparsion:** An average provides a common denominator for comparing one set of data with others. Average facilitates comparision and further analysis for example. If we want to study the economic condition of two countries, it is necessary to reduce the income of the people of two countries into a single value i.e., per capita income (average) then only comparision can be made.

3. **To trace precise relationship:** Average is useful to establish relationship between different groups in quantitative terms. For example, if we say that he income of an average ammerican is more than the average income of an Indian, it is vague, something abstract. There is no definiteness. It will be relatively more precise if the respective incomes are expresed in numbers (in quantitative tersm) i.e., averages.

4. **To know about group:** Average contiaions the characteristics of the data. It describes the data. It helps in obtaining an idea about the group form which it is calculated.

5. **To help in decision making:** Averages are useful in setting standand's. For example business man be often interested to know about average wage, avarage output, average sales, average profit etc.... This averages are valuable ine estimating planning and decision making.

The following are the various types of average in use:

I. Positional averages:

1) Median ..... (M) Symbol used

2) Mode ..... (Z) Symbol used.

II. Mathematical averages:

3) Arithemetic average $\bar{X}$ symbol used

4) Geometric Mean GM symbol used

5) Harmonic Mean HM symbol used

6) Quadratic Mean QM symbol used

III. Commerical Averages:

7) Moving Average

8) Progressive Average

Arthmetic Mean:

Mean or arithmetic Mean is the most widely used and the most generally of all averages. The calculation of mean is a straight forward operation.

It is defined as "the quotient that resutls when the sum of all the items in the series is divided by the number of items."

Symbolically Mean or $\bar{X} = \frac{X_1 + X_2 + X_3 + \cdot\ \cdot\ \cdot + X_n}{N}$

$$= \frac{\sum X}{N}$$

X = Mean, $X_1, X_2 =$ values, N = Total number of values, $\sum =$ Summation of sigma.

Merits:

1. It is easy to calculte and simple to understand.
2. It is regidly defined.
3. It is based on all the observations of the series.
4. Each independent item plays an equal part in the determination of the mean.
5. It is amenable to algebraic treatment.
6. It facilitates comparision
7. It possesses fairly good sampling stabilities.
8. It causes kess fluatuations in sampling.

It contains all the required charateristics of a good average.

Demerits:

Though it is common, familiar and good average it has certain demerits.

1. The greatest disadvantages of mean is that, it induly effected by extreme items.

   For example: The average salary of three persons Rs. 30, 300, 3000.

$$\frac{30 + 300 + 3000}{3} = \frac{3330}{3} = \text{Rs} \cdot 1110$$

   This average is not correct representative of the data, an non-existent in the series.

2. It gives greater importance to bigger items and less importance to small items. It has an upward bias.

   For example: Wages Rs. 4, 5, 5, 6, 80 average is Rs. 20 It is effected by the presence of 80 average is pooled up.

   For example: Salaries Rs. 50, 400, 300, 500, 600 average is = 370

   It is not effected by the presence of small item 50 here average is not pooled down.

3. Sometimes the average may not be actual item in the series. For example the average of 5, 3, 2, 6 is 4 which not in the series hence can not be representative of the data.
4. Average in certain circumstances may give misleading and observed results. For example No of children for three families 4, 3, 6. Avearage is 4, 3 which is redicules.
5. If any of the item is ignored or lost, the accuracy of mean will be effected.

Example:

The following table gives the monthly sales of ten firms.

Firms	1	2	3	4	5	6	7	8	9	10
Sales (Rs.)	1200	1000	900	1600	500	1500	1100	600	800	300

Calculate arithmetic mean of sales

Solution:

Calculation of Arthmetic Mean

Net sales be denoted by the symbol X

S.No.	X (Sales Rs.)
1	1200
2	1000
3	900
4	1600
5	500
6	1500
7	1100
8	600
9	300
10	300
N = 10	$\sum X = 9,500$

$$\bar{X} = \frac{\sum X}{N}$$

$$\bar{X} = \frac{9500}{10} = 950$$

Average Sales = 950 Rs.

Continuous Series: In continuous series, arithmetic mean may be computed either by direct method or by short cut method.

Direct Method:

$$\bar{X} = \frac{\sum fm}{N}$$

m = mid point of the class interval, f = frequency, fm = m x f, N = Total.

Example:

Calculate Mean from the following: (Direct Method)

Marks:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Students:	8	12	30	20	20	10

Solution: Calculation of Mean

Marks	**Students**		
	f	**m**	**f m**
0 - 10	8	5	40
10 - 20	12	15	180
20 - 30	30	25	750
30 - 40	20	35	700
40 - 50	20	45	900
50 - 60	10	55	550
	N = 100		$\sum f m = 3120$

$$\bar{X} = \frac{\sum f m}{N}$$

$$\bar{X} = \frac{3120}{100} = 31 \cdot 2 \text{ Marks}$$

The above problem is worked out by short cut method.

Shortcut Method:

$$\bar{X} = A + \frac{\sum f d}{N}$$

A = Assumed Mean, d = deviations from assumed mean (m - A),

f = frequency $f d = (f \times d)$, N = Total of frequencies.

Procedure:

1. Take an assumed mean
2. Obtain deviations from assumed mean
3. Multiply with frequencies
4. Apply the above formula

Solution: Calculation of Mean

Marks	Students	m	d = m - 25 A = 25	f d
0 - 10	8	5	- 20	- 160
10 - 20	12	15	- 10	- 120
20 - 30	30	25	0	0
30 - 40	20	35	+ 10	+ 200
40 - 50	20	45	+ 20	+ 400
50 - 60	10	55	+ 40	+ 300
	N = 100			$\sum f d = 620$

$$\bar{X} = A + \frac{\sum f d}{N}$$

$$\bar{X} = 25 + \frac{620}{100}$$

$$\bar{X} = 31 \cdot 2 \text{ Marks}$$

Step Deviation Method: In case of grouped or continuous series with class intervals of equal magnitude the calculations can be simplified by taking a common facator in the mid value of the class.

Then the formula for Mean is as follows:

$$\bar{X} = A + \frac{\sum f d'}{N} \times C \; ; \; C = \text{Common Factor}$$

Example:

Calculate Mean from the following:

Life in hours	Number of Bulbs
0 - 400	4
400 - 800	12
800 - 1200	40
1200 - 1600	41
1600 - 2000	27
2000 - 2400	13
2400 - 2800	13
	150 (Andhra 71)

Solution:

calculation of mean

$$d = \frac{M - A}{100} = \frac{M - 1400}{100}$$

$$A = 1400$$

X	f	m	d'	$f\,d'$
0 - 400	4	200	- 12	- 48
400 - 800	12	600	- 8	- 96
800 - 1200	40	1000	- 4	- 160
1200 - 1600	41	1400	0	0
1600 - 2000	27	1800	+ 4	+ 108
2000 - 2400	13	2200	+ 8	+ 104
2400 - 3800	13	1600	+ 15	+ 156
	N = 150			$\sum f\,d' + 64$

$$\bar{X} = A + \frac{\sum f\,d'}{N} \times C$$

$$\bar{X} = 1400 + \frac{64}{150} \times 100$$

$$\bar{X} = 1400 + \frac{6400}{150}$$

$$\bar{X} = 1400 + 42 \cdot 6 = 1442 \cdot 6 \text{ Hours}$$

Step deviation is taken C = Common factor, C = 100

5. Mathematical Properties of Arithmetic Mean:

i) The products of arithmetic mean X and the total number of values N which the mean is based is equal to the sum of all given values.

ii) The algebraic sum of deviation of the given values from the arithmetic mean in equal to zero. Thus it is regarded as the point of balance.

iii) The sum of square of deviations is minimum when taken from the arithmetic mean.

6. Weighted Arithmetic Mean:

Simple mean assumes equal importance to all the values of sizes of items in a given distribution. But in practice greater importance to some observations are given to compared to others. When a firm wishes to ascertain the average cost of produceing one unit the weightages are considered in terms of number of skilled labour hours used and unskilled labour hours used.

In weighted average the component items are being multiplied by certain values known as 'weights' and the aggregate multiplied product is divided by the total sum of their weights instead of total number of observations.

$$\bar{X} = \frac{W_1X_1 + W_2X_2 + \cdots + W_nX_n}{W_1 + W_2 + \cdots + W_n}$$

$$= \frac{\sum WX}{\sum W}$$

Example: An examination was held to decide the award of scholarship. The weights given to the various subjects were different. Marks of three applicants were as follows:

Subjects	Weights	Marks		
		A	B	C
Statistics	4	60	65	63
Accountancy	3	64	70	63
Economics	2	56	63	58
Principles of Management	1	80	52	70

If the candidate setting heighest marks is to be awarded the scholar ship, who should get it?

Calculation of weighted average mean marks of A B C.

Marks	Weights		Marks	Weights		Marks	Weights	
X A	W A	XWA	XB	WB	XWB	XC	WC	XWC
60	4	240	65	4	260	63	4	252
40	3	192	70	3	210	65	3	195
56	2	112	63	2	126	58	2	116
58	1	80	52	1	52	70	1	70
	$\sum WA =$ 10	$\sum XWA =$ 624		$\sum WB =$ 10	$\sum XWB =$ 648		$\sum WC =$ 10	$\sum XWC =$ 633

$$\bar{X}\,WA = \frac{624}{10} = 62\cdot 4,\ \bar{X}\,WB = \frac{648}{10} = 64\cdot 8,\ \bar{X}\,WC = \frac{633}{10} = 63\cdot 3.$$

7.3 Median:

Median is the mid-item of a series when it is arranged either an accending or deccending order of magnitude. It divides the data into two parts. In one part the values will be less than the median and in the other part the values wil be more than the median. It is located by inspection. It is a positional average.

It is defined as "that value which divides a distribution so that an equal number of item is on either side of it".

Symbolically $\prod = \left(N + \frac{1}{2}\right)^{th}$ item.

= denotes number of items.

Median is the size or value of the middle item. For example marks of 5 students 40, 20, 50, 70, 30. If they are placed in order 20, 30, 40, 50, 70. The median would be the size of third item $\left(N + \frac{1}{2}\right)^{th}$ item. It means $\left(5 + \frac{1}{2}\right) =$ 3rd item 40. If the number of item in a series is even the median will be the average value of the middle two items.

Merits:

1. It is easy to calculated and simple to understand.
2. It's position is based on all the observations.
3. Its value is not effected by extreme items on either end.
4. Usually it is an actual item from the series.
5. It is an appropriate measure of central tendency in such qualitative characteristics which can be marked such as intelligence beauty etc.

Demerits:

1. It is neither so familiar nor so widely as mean.
2. It is not adoptable to alegebraic treatment
3. The computation of median requires the arrangement of the data in order. It is very often a combersome job.
4. It is erratic if the number of items is small
5. It is not useful in those cases where large weight is to be given to extreme item for it gives equal weight to all items.
6. Median is more likely to be effected by the fluctuation of sampling than arithmetic mean.

1. Individual Series:

$$M = \text{size} \left(\frac{N+1}{2}\right)^{th} \text{ item}$$

Procedure:

1. Arrange the items either in ascending order or descending order
2. Locate the $\left(\frac{N+1}{2}\right)^{th}$ item. It is the median
3. In case the number of items is even the average of the middle items should be taken.

Example:

Find out median from the following:

Wages: Rs. 40, 15, 60, 75, 100, 10, 95, 110, 42.

Solution:

Wages

10

15 M = size of $\left(\frac{N+1}{2}\right)^{th}$ item

40

42 $= \left(\frac{9+1}{2}\right) = 5\text{th item}$

60

75

95 M = 60 hence Rs. 60 is the Median Wage

100

110

2) Continuous Series:

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

In continuous series the value of $\left(\frac{N}{2}\right)^{th}$ item is taken and not $\left(\frac{N+1}{2}\right)^{th}$ item.

Procedure:

1. Arrange the data in ascending order.
2. Find out cumulative frequencies.
3. Findout the class inwhich median lies by $\left(\frac{N}{2}\right)$
4. Having discovered the class use the following formula to find out Median Value.

$$M = L + \frac{\frac{N}{2} - Cf}{f} X$$

L = Lower limit of the median class

$\frac{N}{2}$ = Half of the total frequency

C f = Cumulative frequency

f = Frequency of the median class

i = Class interval of the median class

Example:

Find out Median from the following:

Marks	**Students**
0 - 10	25
10 - 20	40
20 - 30	65
30 - 40	110
40 - 50	70
50 - 60	55
60 - 70	20
70 - 80	15

Solution:

Calculation of Median

(Marks) Class	Students (f)	C f
0 - 10	25	25
10 - 20	40	65
20 - 30	65	130
30 - 40	110	240
40 - 50	70	310
50 - 60	55	365
60 - 70	20	385
70 - 80	15	400

M = size of $\left(\frac{N}{2}\right)^{th}$ item $\frac{400}{2} = 200^{th}$ item

200^{th} item is in the class 30 - 40

$$M = L + \frac{\frac{N}{2} - CF}{f} \times i$$

$$M = 30 + \frac{200-130}{110} \times 10$$

$$M = 30 + 6 \cdot 3 = 36 \cdot 3 \text{ marks}$$

8. Mathematical Properties of Median:

(i) The sum of the deviations of the items from median ignoring signs is the least. For example the median of 4, 6, 8, 9, 12 is 8 (ignoring the signs) the deviation from 8 totals to 4 + 2 + 0 + 1 + 3 = 10.

(ii) Deviation taken from any other value will be more than 10 like, if we take 6 and take the deviation from the deviation totals to 2 + 0 + 2 + 3 + 6 = 13.

7.4 Quartiles:

The quartiles is a measure which divides the series into four equal parts when the series is being arranged in the order of magnitude, There are three quartiles which divide a series into four equal parts. The first quartile is the value of the variable which divides the series into 25 percent of

items below it and 75 percent above it. The second quartile which divides the series into two equal parts which is also the median. The third quartile is the value of the variable that divides the series in a way so that 75 percent of items are below it and 25 percent of items exceed the value.

For grouped data, the following formulas are used for quartiles.

$$Q_j = L + \frac{j\,N/4 - PCf}{f} \times i \quad \text{for } j = 1, 2, 3.$$

Where L is lower limit of the quartile class, PCf is the preceding cumulative frequency to the quartile class, f is the frequency of the quartile class and i is the size of the quartile class.

Deciles:

Decile is a measure which divides the series into 10 equal parts. Thus there are 9 deciles in a series. For grouped data, the following formulas are used for deciles.

$$D_K = L + \frac{KN/10 - PCf}{f} \times i \quad \text{for } K = 1, 2, \ldots\ldots\ldots\ldots,9$$

Where the symbols have usual meaning and interpretation.

Percentiles:

Percentiles is a measure which divides the series into 100 equal parts. Thus there are 99 percentiles in a series. For grouped data, the following formulas are used for percentiles.

$$P_i = L + \frac{1N/100 - PCf}{f} \times i \quad \text{for } i = 1, 2, \ldots\ldots\ldots, 99$$

Example:

Find the median, Q_1 and Q_3 4th decile and 60th percentile for the following distribution:

Marks:	0 - 4	4 - 8	8 - 12	12 - 14	14 - 18	18 - 20	20 - 25	25 - 30
No.of Students:	10	12	18	7	5	8	4	6

Solution:

Here the class - intervals are not all equal. To find any partition value, there is no need to make them equal.

Marks	**No.of Students (f)**	**C.f.**
0 - 4	10	10
4 - 8	12	22
8 - 12	18	40
12 - 14	7	47
14 - 18	5	52
18 - 20	8	60
20 - 25	4	64
25 - 30	6	70
		N =70

Calculation of Median $\frac{N}{2} = 35,\ f = 18,\ PCf = 22,\ i = 4,\ L = 8\cdot$

Median or $Q_2 = L + \frac{2 \times \frac{N}{4} - PCf}{f} \times i$

$$= 8 + \frac{35 - 22}{18} \times 8 = 10\cdot 89$$

Calculation of Q_1, $\frac{N}{4} = 17\cdot 5$, f = 12, PCf = 10, i = 4, L = 4.

$$Q_1 = L + \frac{\frac{3N}{4} - PCf}{f} \times i = 4 + \frac{17\cdot 5 - 10}{12} \times 4 = 6\cdot 5$$

Calculation of D_4: $\frac{4N}{10} = 28,\ i = 4,\ L = 8,\ f = 18,\ PCf = 22\cdot$

$$\therefore D_4 = L + \frac{\frac{4N}{10} - PCf}{f} \times i = 8 + \frac{28 - 22}{18} \times 4 = 9\cdot 33$$

Calculation of P_{60}: $\frac{60N}{100} = 42,\ L = 12,\ i = 2,\ f = 7,\ PCf = 40\cdot$

$$\therefore D_{60} = L + \frac{\frac{60N}{100} - PCf}{f} \times i$$

$$= 12 + \frac{42 - 40}{7} \times 2 = 12\cdot 57$$

7.5 Mode:

Mode is the most common figure of the data or the most repetitive figure of the data. Mode isthe most fashionable. Figure of the data. Mode is an average which is conceptually very useful. Mode is perhaps what most people in mind when they think about an average. Average size of shoes, marks of the class, wage etc. In the words of zizit mode is the value occuring most frequency in a series of items and around which other items are destributed most densely.

Croxton and cowden define mode as follows:

"The mode of a distribution is the value at the point around which the items tend to be must hearily concentrated. It may be regarded as the most typical of a series of values.

Mode can be located by inspection by arranging the data in the order of magnitude. Location of mode in individual series sometimes becomes difficult. Series are to be arranged in order and than it is to be located. Where number is large data are to be condensal into frequency distribution.

In discrete series, if there is regularity and homogeneithy in the series mode can be located by inspection alone. The size having the highest frequency will be recokoned as mode.

In case frequency distribution not being regular then group process can be resorted by taking frequencies sets 12, 23, 123, 234, 345. The grouping is done to ensure that every item has equal influence in grouping process. Ater grouping is over an analysis is made to find out the item. Which repeats largest number of items. In continuous series after finding out the class interval in which mode lies, mode value is to be interpolated.

Mode also can be determined graphically. In binomial and multimodal distributions modal value can be determined by making use of the following formula. There is an emperical relationship between mean, median and mode.

Mode = (3 Median - 2 Mean)

Merits:

1. It is easy to locate and simple to understand.
2. It is by definition the most usual or typical value. It is simple and precise.
3. It is not affected by extremely large or small items.
4. For determination of mode, it is not necessary to known the value of all items.

Mode is having practical importance. In the modern age where large scale production is the order of the day, mode is the most popular measure used for many purpose.

Individual Series:

Calculation of mode in individual series:

The value occurring the maximum number of times is the modal value - arrange the data in ascending order.

Example:

Find out mode from the following:

Marks: 20, 24, 15, 20, 10, 20, 28, 20, 15, 20, 12, 20.

Solution: Arrange the data in order

10, 12, 15, 15, 20, 20, 20, 20, 20, 20, 24, 28.

As 20 is repeated six times, mode is 20.

Mode is continuous series:

Procedure:

1. By preparing grouping table and analysis table or by inspection find out the modal class.
2. Determine the value of mode by applying the following formula.

$$Z = L_1 + \frac{f_1 - f_2}{2f_1 - f_0 - f_2}\left(l_2 - l_1\right)$$

where L_1 = Lower limit of the modal class

L_2 = Upper limit of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the modal class preceding the modal class

f_2 = Frequency of the modal class succeeding the modal class

Example:

Calculate the model wage freom the following data:

Wages (Rs):	0 - 100	100 - 200	200 - 300	300 - 400	400 - 500	500 - 600
No.of Workers:	8	12	25	15	10	6

Solution:

ℓ_1 = Lower limit of model class = 200

f_1 = Frequency of the model class = 25

f_0 = Frequency of the model class preceding the model class = 12

f_2 = Frequency of the model class succeeding the model class = 15

d_2 = upper limit of model class = 300

$$\text{Mode} = 200 + \frac{25 - 12}{2 \times 25 - 12 - 15} \times 100$$

$$= 200 + \frac{13}{23} \times 100 = 200 + 56 \cdot 52$$

$$= \text{Rs } 256 \cdot 52$$

Locating the mode graphically:

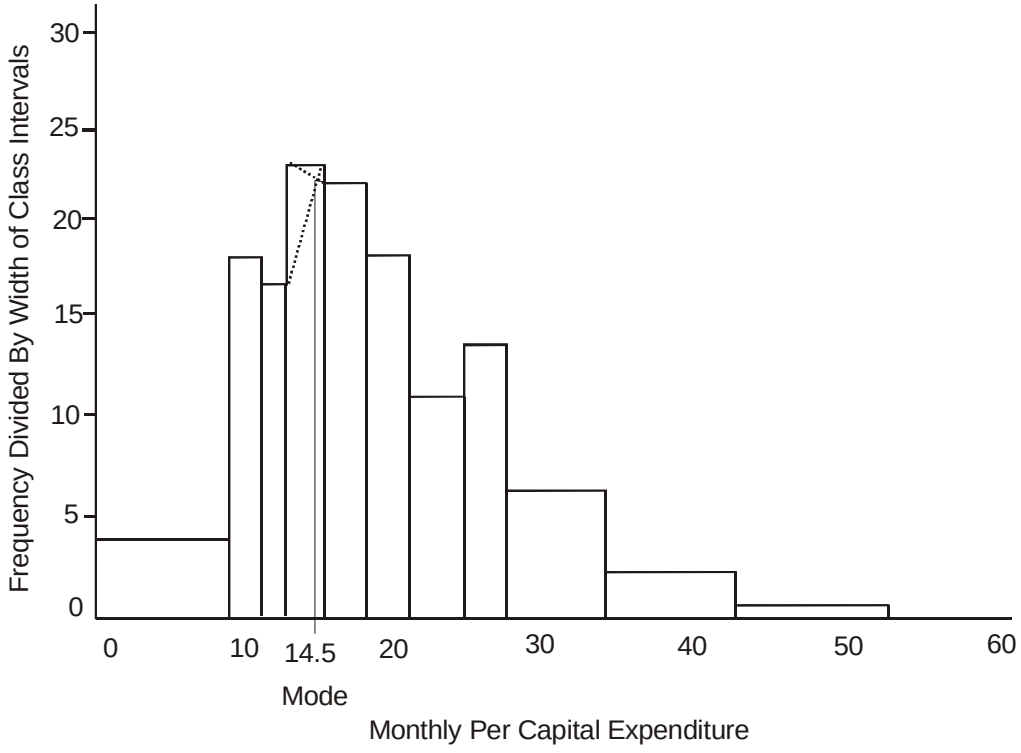
Prepare a histogram, consider a bar having maximum height and the bars to the left and right adjoining to it. Join the top left corner of the highest bar to the top left corner of the right side bar and top right corner ofthe highest bar to the top right corner of the bar to its left. The absciss a value of the point of inter section of the joinning lines is the mode.

Example:

Distribution of housholds according to their per capita expenditure in Rs.

Class interval	Frequency
0 - 8	27
8 - 11	54
11 - 13	34
13 - 15	46
15 - 18	63
18 - 21	54
21 - 24	34
24 - 28	52
28 - 34	39
34 - 43	22
43 - 55	7
55 - 65	25

Computation of Mode



7.6 Geometric Mean:

Geometric Mean is the 'n th' root of the product of in items of a series. Geometric mean is obtained by multiplying the values of items together and extracting the root of the product corresponding to the number of items.

For example:

Geometric mean of the items of 4, 6, 8, 10 is obtained by taking the 4th root of the product.

$$\sqrt[4]{4 \times 6 \times 8 \times 10} =$$

Symbolically $G \cdot M \cdot = \sqrt[n]{X_1 \times X_2 \times X_3 \times \cdot \cdot \cdot X_n}$

Where X_1, X_2, X_3 are the values of items and n = the number of items. As the calculation is difficult in case of larger items the use of logarithms make the task easy.

$$\text{G.M. = Antilog of } \left[\frac{\sum \log X}{N} \right]$$

Merits:

1. It can be calculted with mathematical precision.
2. It is not influenced by the extreme items to the same extent as arithmetic mean.
3. It takes into considaration all the items.
4. It is amenable to algebraic treatment.
5. It gives less weight to big item and mean weight small items.
6. It is particularly useful when dialing with ratios.

Demerits:

1. The geometric mean is not a popular measure of central tendency
2. The average man does not understand its measurement. It is difficult to calcualate.
3. It connot be determined when there are positive and negative values in the series.
4. If one value of the item is zero, the geometric mean is zero.

Geometric mean is used in index numbers. It is also usually employed in the study of social and economic phenomena, when it is required to give greater weight to smaller items.

G.M.

Individual Series:

$$G \cdot M \cdot = \sqrt[n]{X_1 \times X_2 \times X_3 \times \cdot \cdot \cdot \times X_n}$$

N = Number of item, X = Variable.

$$G \cdot M \cdot = \text{Anti}\log \text{ of } \left\{ \frac{\sum \log X}{N} \right\}$$

Procedure:

1. Convert the given values of the variable (X into logarithms)
2. Obtain the total of $\log X = \sum \log X$
3. Divide the $\sum \log X$ by N, get the antilog for the quotient so obtained, it gives a geomentric mean.

Example:

Find out Geometric mean

The monthly incomes of 10 families in a locality are given below:

Rs. 85 70 15 75 500 8 45 250 40 and 36

Solution:

Calculation of G.M.

Family	Income	Logarithms
1	85	1.9294
2	70	1.8451
3	15	1.1761
4	75	1.8751
5	500	2.6990
6	8	0.9031
7	45	1.6532
8	250	2.379
9	40	1.6021
10	36	1.5563
		$\sum \log X 17 \cdot 6373$

$$G \cdot M \cdot = \text{Anti}\log \left\{ \frac{\sum \log X}{N} \right\}$$

$$\text{Antilog of } = \frac{17 \cdot 6373}{10}$$

Antilog of = 1.7637

G.M. Rs. 58.03

Calculate of geometric mean - Discrete Series.

Procedure:

Find out the logarithm of each item (log X) multiply the logarithm of each item by its frequency (f log X)

Summate the products (f log X)

Divide the sum of the products by the total frequencies $\frac{(f \log X)}{N}$

Find out the anti - logarithm of the quotient obtained above. It gives geomentric mean.

$$G \cdot M \cdot - \text{Anti}\log \frac{\sum f \log X}{N}$$

Example:

Find out geomentric mean from the following:

Mark (size)	6	7	8	9	10	11	12
Students (Frequency)	8	12	18	26	16	12	8

Solution:

Calculation of G.M.

X	Logarithms	f	f x log X
6	0.7782	8	6.2256
7	0.8451	12	10.1412
8	0.9031	18	16.2558
9	0.9542	26	24.8092
10	1.0000	16	16.0000
11	1.0414	12	12.4968
12	1.0792	8	8.6336
		N = 100	$\sum f \log X = 94 \cdot 5622$

G.M. Antilog of $\left\{ \frac{\sum f \times \log X}{N} \right\} \frac{94 \cdot 5622}{10} = 94566 - 8 \cdot 822$

Calculation of geometric mean (Continuous series):

(a) Find out the mid-values of the class intervals.

(b) Find out the logs for mid - values

(c) Multiply logs with the respective class frequencies

(d) Obtain total of $f \times \log X = \sum f \log X$

(e) Divide the $\sum f \log m$ by N

(f) Find out antilog of the quotient that is geometric mean.

Example:

Find out geometric mean for the following data:

Age:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Persons:	14	23	27	21	15

Solution:

Calcultaion of G.M.

Age	**Person X**	**Mid -Value X**	**log X**	**f x log X**
0 - 10	14	5	0.6990	9.7860
10 - 20	23	15	1.1761	27.0503
20 - 30	27	25	1.3979	37.7433
30 - 40	21	35	1.5441	32.4261
40 - 50	15	45	1.6532	24.7980
	N = 100		$\sum f \log m$	131·8037

G.M. = Antilog of $\left\{ \frac{\sum f \log X}{N} \right\} = 131 \cdot 8037$

Antilog of 1.3180 = 20.8

7.7 Harmonic Mean:

Harmonic Mean is "the reciprocal of the arithmetic mean of the reciprocals of the observations".

If the given numbers are a, b, c, ...x.

Symbolically H.M. = $\frac{N}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \cdots + \frac{1}{x}}$

For example: Harmonic Mean of 4, 6, 8, is

$$\frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{8}} = \frac{3}{\frac{1}{2}\,\frac{1}{4}} = \frac{3\times 24}{13} = \frac{72}{13} = 5\cdot 5$$

Harmonic mean is a type of statistical average capable of application only with in a restricted field. It is better than arithmetic mean, where the weights are to be considered.

Merits:

1. It is determination
2. It is based on all the observations
3. It gives less weight to large items and more weight to small items
4. This average is useful in the case of series having wide dispersion
5. This is useful in calculating rates

Demerits:

1. It is difficult to calculate and also difficult to understand
2. It can be calculated only when all items are known
3. It is meaningless when the observations include positive and negative values, when one or more values are zero
4. It is not suitable for general application
5. It may not be an actual items in the series

Individual Series:

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

Procedure:

1. In individual series, obtain reciprocal for all items.
2. Add reciprocals = $\sum \text{Reciprocals}$
3. Divide the total by the number (N)
4. Find out reciprocal of the above

Example:

Find out Harmonic Mean

Marks obtained by 7 students 35, 30, 15, 20, 10, 5, 25.

Solution:

Calculation of Harmonic Mean

S.No.	X	Reciprocals $\frac{1}{X}$
1	35	0.0285
2	30	3.0333
3	15	0.0666
4	20	0.0500
5	10	0.1000
6	5	0.2000
7	25	0.0400
	$\sum \frac{1}{X}$	0.5184

Harmonic Mean $\frac{N}{\sum \frac{1}{X}}$

$$H \cdot M \cdot = \frac{7}{0 \cdot 5184} = 13 \cdot 6$$

Discrete Series:

Calculation of Harmonic Mean

Procedure:

1. Take the reciprocals of the sizes of the variable X
2. Multiply these reciprocals by the respective frequencies
3. Get the total i.e. $\sum \left(f \times \frac{1}{X} \right)$

$$H \cdot M \cdot = \frac{N}{\sum \left(f \times \frac{1}{X} \right)}$$

Example:

Find out Harmonic Mean from the following:

Marks:	10	20	25	40	50
Students:	10	15	25	10	5

Solution:

Marks(X)	**Students (f)**	**Reciprocal** $\left(\frac{1}{x}\right)$	$f\frac{1}{X}$
10	10	1000	1.000
20	15	0500	0.7500
25	25	04000	1.0000
40	10	02500	0.2500
50	5	02000	0.1000
	N = 65		$\sum f \times \frac{1}{X}\ 3\cdot1000$

$$H\cdot M\cdot = \frac{N}{\sum\left(f \times \frac{1}{X}\right)}$$

$$H\cdot M\cdot = \frac{65}{3\cdot100} = 20\cdot968 = 20\cdot97$$

Harmonic Mean continuous series:

Procedure:

1. Obtain mid values (X)
2. Find out reciprocals for mid values (X)
3. Multiply these reciprocals by the respective for frequencies
4. Get the total $\sum\left(f \times \frac{1}{X}\right)$
5. Apply the formula $\frac{N}{\sum\left(f \times \frac{1}{X}\right)}$

Example:

Find out Harmonic Mean from the following:

Age:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Persons:	5	10	7	3	2

Solution:

Age (X)	**X**	**Persons f**	**Reciprocal** $\frac{1}{X}$	$f \times \frac{1}{X}$
0 - 10	5	5	2000	1.000
10 - 20	15	10	0666	0.667
20 - 30	25	7	0400	0.280
30 - 40	35	3	0285	0.085
40 - 50	45	2	0222	0.044
		N = 27	$\sum f \times \frac{1}{X}$	2.076

$$H \cdot M \cdot = \frac{N}{\sum \left(f \times \frac{1}{X} \right)}$$

$$= \frac{27}{2 \cdot 076} = 13$$

7.8 Relationship among mean, median and mode:

There exists a definite relationship among three measures of central tendency namely mean, median and mode. In case of the frequency distribution is perfectly symmetrical the values of mean, median and mode are identical i.e.. mean = median = mode.

As the frequency distribution departs from symmentry these values differ. How ever they still maintain a definite relation to each other. A distribution may be asymmetrical either to the left or to the right.

(a) When the distribution is skewed to the left median is less than mode and mean is less than median. i.e.: Mean < Median < Mode

(b) When the distribution is skewed to the right the mean is greater than median and more. Their relation ship may be put as mean > median > mode.

It is evident from the above that median always lies between mean and mode. When mode is minimum the distribution is skewed towards left. For a moderately skewed distribution karl pearsons has given the relationship among 3 measures of central tendency as given below:

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

mode = 3 Median - 2 Mean and Median = Mode + 2/3 (mean - mode)

The relationship is based ont he fact that distance between mean and median is half of the distance between mode and median. The figures below show the relative position of mean, median and mode for frequency distributions which are moderately asymmetrial.

7.9 Summary:

We have discussed five alternative averages as measures of central tendency. We have also noted situations where a specific average is more appropriate than others. The arithmetic mean is widely used and understand as a measure of central tendency. The concept of weighted arithmatic mean. G.M. and H.M. are useful for specific type of applications. The Median is generally a more representative measure for open - end distribution. The mode should be used when the most demanded or customary value is needed.

7.10 Exercises:

1. List the various measures of central tendency studied inthe lesson and explain the difference between them.
2. Discuss the mathematical properties of arithametic mean and median
3. Review of each of the measure of central tendency their advantages and disadvantages.
4. Explain how you decide which average to use in a particular problem
5. What are quartiles ? Explain and illustrate the concepts of quartiles, deciles and percentiles.
6. How do you account for the predominant choice of arithmetic mean of statistical data as a measure of central tendency? Under what circumstances would it be appropriate to use mode or median?
7. Define median and mode with examples. Show how can they be calculated in case of discrete value?
8. Under what circumstances the Geometric Mean and Median are considered to be most suitable measures for describing the central tendency of a frequency distribution.
9. Calculate the Harmonic Mean of the following series of monthly expenditure of a batch of students.

   Rs. 120, 130, 75, 10, 45, 0.5, 0.4, 500, 150.

10. Calculate Arithmetic Mean

Sales (in 000)	No.of Firms
0 - 500	3
500 - 1000	24
1000 - 1500	55
1500 - 2000	98
2000 - 2500	120
2500 - 3000	95
3000 - 3500	51
3500 - 4000	39
4000 - 4500	15
	Total 500

11. Calcualte the Median and Mode of the following data:

Sales (in 000)	No.of Firms
410 - 419	14
420 - 429	20
430 - 439	42
440 - 449	54
450 - 459	45
460 - 469	18
470 - 479	7

12. An incomplete distribution is given below:

Variable	Frequency
0 - 10	10
10 - 20	20
20 - 30	?
30 - 40	40
40 - 50	?
50 - 60	25
60 - 70	15
	170

13. Monthly income of families in rupees in a locality are given below calculate Geometric Mean.

    Rs. 85, 70, 15, 75, 500, 8, 45, 250, 40, 36.

14. Calculate the geometric mean weight of the following data:

Weights (in Kgs)	No. of Workers
130	3
135	4
140	6
145	6
146	3
148	5
149	2
150	1
157	1

15. Find lower quartile, 8th decile and 56th percentile and mode for the following distribution:

Class	Frequency
1 - 3	6
3 - 5	53
5 - 7	85
7 - 9	56
9 - 11	21
11 - 13	16
13 - 15	4
15 - 17	4

7.11 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof. M. KOTESWARA RAO

Lesson - 8

MEASURES OF VARIATION AND SKEWNESS

Objectives:

After going through this lesson, you will learn:

- The concept of measures of dispersion
- The computation of range, quartile deviation, Mean deviation, varience, standard deviation and skewness.

Structure:

8.1 Introduction:

Experience tells that in many situations, the spread of values is different but their central values are same. All the more, a central value provides no information aboutthe scattering of values in a set of data. Hence, certain measures are evolved which reflect ont he scattering of values in numerical terms are known as measures of dispersion.

Types of dispersion:

(i) Range, (ii) Interquartile Range and Quartile Deviation, (iii) Mean Deviation, (iv) Median absolute deviation, (v) Variance, (vi) Standard deviation and (vii) Coefficient of variation.

Requisities of dispersion:

i) It should be based on all the observations.

ii) Its unit should be some as the unit of measurement of items.

iii) It should be rigidly defined.

iv) It should follow general rules of mathematics.

v) It should not be subjected to complicated and tedious calculations.

Uses of dispersion:

i) It tells about the reliability of a measure of central value.

ii) It makes possible to compare two series of data in respect of their variability.

iii) Measure of dispersion provides the basis for the control of variability.

iv) It has a wide application in almost all fields of statistics.

8.2 Range Interquartile Range:

The difference between the largest and smallest values of a set of data is called its range. Range is shown as smallest value(s) - largest value (L).

Merits:

i) It is the easiest measure of dispersion.

ii) It can always be found out usually i.e., it involves no calculations.

iii) It is one of the largely used measure of dispersion.

Demerits:

i) It depends on two extreme values ofa series. Thus, it gives no information about the observations lying between smallest and largest values.

ii) It is highly susceptible to sampling fluctuations.

iii) It is not suitable for further mathematical treatment.

iv) Addition or deletion of a single value may change the entire complex of range.

Coefficient of Range:

It is a pure number given as the ratio of difference between the largest and smallest values to the sum of the largest and smallest values of a set of data. Numerically, coefficient of range is $(L-S)/(L+S)$. Lesser the coefficient of range, better it is.

Individual Series:

Range (R) = L - S

L = Largest value S = Smallest Value

Coefficient of Range (CR) = $\frac{L-S}{L+S}$

Example:

Find out Range and Coefficent of Range from the following wages (Rs).

70, 40, 52, 30, 65, 20.

Solution:

Range = L - S

= 70 - 20 = 50

Coefficinet of Range $= \frac{L - S}{L + S}$

$$= \frac{70 - 20}{70 + 20} = \frac{50}{90} = 0 \cdot 55$$

Descrete Series:

Example:

Mraks	1	2	3	4	5	6	7	8	9	10
Students	5	10	20	30	50	20	10	5	3	2
	Total 155									

Solution:

Range = L - S = 10 - 1 = 9

Coefficient = $\frac{L - S}{L + S} = \frac{10 - 1}{10 + 1} = 0 \cdot 82$ approx

Continuous Series:

Example:

Find out range and coefficinet of range.

Marks	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Students	5	10	7	5	4	2

Solution:

In marks highest value 70, smallest is 10.

So range = L - S = 70 - 10 = 60

Coefficient of Range $= \frac{L - S}{L + S} = \frac{70 - 10}{70 + 10} = \frac{60}{80} = 0 \cdot 75$

Interquartile Range:

It is equal to the difference between the upper and lower quartiles. Symbolically it is equal to $(Q_3 - Q_1)$. This measure of dispersion tells about the range of the middle 50% values of a set of data. In this measure, lower 25% and upper 25% values are excluded. It is nto a good measure of dispersion as it tells nothing about the dispersion of values around average. It hardly fulfils any of the requisites of a good measure of dispersion.

8.3 Quartile Deviation (Q.D.):

It is half of the interquartile range i.e. $(Q_3 - Q_1)/2$. It is an absolute measure of dispersion. Hence, to compare two series, a relative measure known as coefficient of quartile deviation is given which is symbolically expressed as $(Q_3 - Q_1)/(Q_3 + Q_1)$.

Merits:

i) It is easy to calculate and understand.

ii) It can be calculated in case of open end frequency distributions as well.

iii) It is not effected by 25% upper and 25% lower extreme values.

Quartile Deviation or Semi - Inter - Quartile Range:

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

$$\text{Coefficient of Quartile Deviation (CQD)} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Individual Series:

Example:

Form the following data find out quartile deviation and its coefficient.

Months	J	F	M	A	M	J	J	A	S	O	N	D
Sales (000 Rs.)	70	40	20	60	55	90	10	80	25	5	8	9

Solution:

Arrange the data in order

5 8 9 10 20 25 40 55 60 70 80 90

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{th} \text{ item } = \frac{12+1}{4}$$

= 3.25th item = 9 + 0.25 = 9.25

$$Q_2 = \text{size of } \left(\frac{3(N+1)}{4}\right)^{th} \text{ item } = \frac{3(12+1)}{4}$$

= 9.75th item 60 + 7.5 = 67.5

$$Q \cdot D. = \frac{Q_3 - Q_1}{2} = \frac{67 \cdot 5 - 9 \cdot 25}{2} = \frac{58 \cdot 25}{2} = 29 \cdot 12$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{67 \cdot 5 - 9 \cdot 25}{67 \cdot 5 + 9 \cdot 25} = 0 \cdot 7$$

Example:

Find out quartile deviation and its coefficient

Wages	10	20	30	40	50	60	70	80
Workers	20	45	110	200	60	36	22	10

Solution:

Calculation of Quartiles

X (Wages)	f (Workers)	C f
10	20	20
20	45	65
30	110	175
40	200	375
50	60	435
60	36	471
70	22	493
80	10	503

$$Q_1 = \text{size of } \left(\frac{N+1}{4}\right) \quad \frac{503+1}{4} = 126^{th} \text{ item}$$

$$Q_3 = \text{Size of } 3\,\frac{N+1}{4} \quad \frac{3(503+1)}{4} = \frac{3(503+1)}{4} = 378^{th} \text{ item}$$

$$Q_1 = 30 \qquad Q_3 = 50$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{50 - 30}{2} = 10$$

$$CQD = \frac{Q_3 - Q_1}{Q_4 + Q_1} = \frac{50 - 30}{50 + 30} = \frac{20}{80} = 0 \cdot 25$$

Continuous Series:

Example:

Find out quartile deviation (QD) and its coefficient.

Wages in Rs.	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45
No.of Workers	2	10	25	16	7

Solution:

Calculation of Quartiles

Wages(X)	**No.of. Workers (f)**	**C f**
20 - 25	2	2
25 - 30	10	12
30 - 35	25	37
35 - 40	16	53
40 - 45	7	60

$$Q_1 = \text{Size of } \left(\frac{N}{4}\right)^{th} \text{ item } = \frac{60}{4} = 15^{th} \text{ item}$$

Q_1 is in the class 30 - 35

$$Q_1 = L + \frac{\frac{N}{4} - Cf}{f} \times i$$

$$= 30 + \frac{15 - 12}{25} \times 5$$

$$= 30 + \frac{15}{25} = 30 \cdot 6$$

$$Q_3 = \text{Size of } \frac{3(N)}{4} = \frac{3(60)}{4} = 45^{th} \text{ item}$$

Q_3 is in the class 35.40

$$Q_3 = L + \frac{\frac{3N}{4} - Cf}{f} \times (i)$$

$$= 35 + \frac{45 - 37}{16} \times 5$$

$$= 35 + \frac{40}{16}$$

$$= 35 + 2 \cdot 5 = 37 \cdot 5$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{37 \cdot 5 - 30 \cdot 6}{2} = 3 \cdot 45$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{37 \cdot 5 - 30 \cdot 6}{37 \cdot 5 + 30 \cdot 6} = 0 \cdot 101$$

8.4 Mean Deviation:

Range and quartile deviation are positional measures of dispersion, where as mean deviatiion is a measure of dispersion which is based on all values of a set of data. It is defined as the average of the absolute deviations taken from an average usually, the mean, median or mode. It is usually denoted by δ. To clarify whether the average used in mean deviation is mean, median or mode a suffix is attached to δ such as $\delta \bar{x}$ δM or δMO.

The formula for calculating mean deviation of n

Observations $X_1, X_2, \cdots, X_n$ is

$$M \cdot D. = \frac{1}{n} \sum_i \left| X_i - A \right|$$

for i = 1, 2, . . . , n

Also 'A' may be any choosen constant out of mean, median and mode.

For a frequency distribution in which the variate value x_i occurs f_i times $(i = 1, 2, \cdots, K)$ the formula for mean deviation is,

$$M \cdot D \cdot = \frac{1}{n} \sum_i f_i \left| X_i - A \right|$$

where $\sum_i f_i = n$ and A as defined above.

Here, it is worth emphosising that mean deviation is minimum about the median. That is why median is commonly used as an average value about which the mean deviation is calculated.

Coefficient of Mean Deviation:

Mean deviation has the same unit of measurement as that of the variable x. It two series have different units of measurement, the series cannot be compared. Hence for comparing any two series, an unitless measure is given known as coefficient of mean deviation. It is the ratio of mean deviation to the average 'A' used in calculating it. Its formula is,

$$\text{Coefficient of M.D.} = \frac{\text{Mean Deviation}}{\text{A}} \times 100$$

It is multipled by 100 to express coefficient of mean deviation is percentage

Merits:

1. It utilises all the observations of the set.
2. It is simple to calculate and understand.
3. It is least affected by extreme values.

Demerits:

1. The foremost weakness of mean deviation is that in its calculation negative differences are considered positive without any sound reasoning.
2. It is not amenable to further algebraic treatment.
3. It can not be calculated in case of open end frequency distribution.

Example:

Individual Series:

Calculate mean deviation from meand and median and also coefficients of the following data:

Marks - 54, 80, 57, 52, 49, 45, 72, 57, 47.

Solution:

Mean deviation from Mean			Mean deviation from Median	
S.No.	X	$(dX)_{57}$	(series are arranged in order)	$(dM)_{54}$
1	54	3	45	9
2	80	23	47	7
3	57	0	49	5
4	52	5	52	2
5	49	8	54	0
6	45	12	57	3
7	72	15	57	3
8	57	0	72	18
9	47	10	80	26
	$\sum X 513$	76		73

$$\bar{X} = \frac{\sum X}{N} = \frac{513}{9} = 57$$

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

$$\delta\bar{X} = \frac{\sum\left(d\bar{X}\right)}{N}$$

$$M = \left(\frac{9+1}{2}\right) 5^{th} \text{ item} = 54$$

$$\delta\bar{X} = \frac{76}{9} = 8\cdot44$$

$$\delta M = \frac{\sum(dm)}{N} = \frac{73}{9} = 8\cdot1$$

Coefficient of mean deviation

$$\frac{\delta\bar{X}}{\bar{X}} = \frac{8\cdot4}{57} = 0\cdot15$$

Coefficient of mean deviation from median

$$\frac{\delta M}{M} = \frac{8\cdot1}{54} = 0\cdot15$$

Example:

Find the mean deviation of the following distribution:

Age Group	15 - 24	25 - 34	35 - 44	45 - 54	55 - 64
No observed	4,000	16,000	28,000	33,000	28,000

Solution:

Calculation of mean deviation from Mean

Age Group	**Mid Value (X)**	$\mathbf{d = \frac{X - A}{C}} = \frac{X - 39\cdot5}{10}$	**f**	**f d**	$\lvert d\rvert = \left(X - \bar{X}\right)$ $\left(\bar{X} = 45\cdot5\right)$	$\mathbf{f\lvert d\rvert}$
15 - 24	19.5	- 2	4000	- 8000	26	1,04,000
25 - 34	29.5	- 1	16000	- 16000	16	2,56,000
35 - 44	39.5	0	28000	0	6	1,68,000
45 - 54	49.5	+ 1	33000	+ 33000	4	1,32,000
55 - 64	59.5	+ 2	28000	+ 56000	14	3,92,000
			N = 1,09,000	$\sum f d =$ 65,000		$\sum f\lvert d\rvert =$ 10,52,000

$$\bar{X} = A + \frac{\sum fd}{N} \times C$$

A = 39.5, f d = 65,000 ; N = 109.000, C = 10.

$$\bar{X} = 39 \cdot 5 + 65 \cdot 000/1,09 \cdot 000 \times 10$$

= 39.5 + 5.96

= 45.46 = 45.5

Mean deviation from Mean = $\frac{f|d|}{N}$

$f|d| = 10,52,000$; $N = 1,09,000$

$$M \cdot D \cdot = \frac{10,52,000}{1,09,000} = 9 \cdot 65$$

Note: In this problem deviations are taken from Arithmetic mean, approximating it as 45.5

8.5 Variance:

The average of the square of the deviations taken from mean is called variance. The population variance is generally denoted by σ^2 and its estimate (sample variance) by S^2. For N population values $X_1, X_2, \cdots, X_N$ having the population mean μ, the population varience,

$$\sigma^2 = \frac{1}{N} \sum_i (X_i - \mu)^2$$

For i = 1, 2, ..., N

Where $\mu = \sum_i X_i / N$

An estimate of σ^2 based on n sample values $x_1, x_2, \cdots, x_n$ the sample varience.

$$S^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

for i = 1, 2, ..., n

and $\bar{x} = \sum_i x_i / n$

For a frequency distribution of sample values $x_1, x_2, \cdots, x_k$ having frequencies $f_1, f_2, \cdots, f_k$ respectively. The sample varience,

$$S^2 = \frac{1}{n-1} \sum_i f_i (x_i - \bar{x})^2$$

$$= \frac{1}{n-1}\left\{ \sum_i f_i x_i^2 - \frac{\left(\sum_i f_i x_i\right)^2}{n} \right\}$$

where $n = \sum_i f_i$

Merits and demerits of varience:

It prossess all the requisites of dispersion except that its unit is square of the unit of measurement of variate values. Hence, many times it becomes difficult to actually adjudge the magnitude of variate. Also variance is sensitive to extreme values. Variance is the backbone of statistics.

Absolute and Relative Dispersion:

If the unit of a measure of dispersion is in same terms as that of the observations of a series, it is called absolute measure of dispersion, e.g., height in cms, weight in kilograms, income in rupees etc. In this case two series having different units of dispersion, can not be compared. Hence, for comparison of two series having different units of measurement, one requires an unit less measure of dispersion. Such measures are termed as relative measures of coefficient of dispersion. For this a measure of dispersion is usually divided by mean used in its calculation and multiplied by 100. This provides the measure expressed in percentage which is fit for comparision of any two or more series.

Standard Deviation:

The positive square root of the variance is called standard deviation. The idea of standard deviation was first given by Karl Pearson in 1893.

Symbolically,

$\sigma = \sqrt{\sigma^2}$ (population)

$S = \sqrt{S^2}$ (sample)

It fulfils all the requisites of dispersion except that it is sensitive to extreme values. That is why it is known as standard deviation.

Median Absolute Deviation (MAD):

Standard deviation is affected by extreme values. Hence, the median absolute deviation is an alternative measure of dispersion. Median absolute deviation is defined as the median of the absolute deviation taken from median. It is seldom used as it is not easily an enable to further algebraic treatment. Moreover, it is not involved in distribution function.

The formula for median absolute deviation is,

$$MAD = \text{Median} \left| X_i - X_{Md} \right|$$

The variance or standard deviation remains the some as they are independent of change of origin.

The variance of the new set of values will be $\frac{1}{d^2}$ times the original value of variance.

Thevariance of the transformed set of values will be d^2 times the varience of the original set of values.

Coefficient of variation and its importance:

Coefficient of variation (C.V.) is the ratio of the standard deviation and the mean. Usually it is expressed in percentage. The formula for coefficient of variation is,

$$C \cdot V \cdot = \frac{S \cdot D.}{\text{mean}} \times 100$$

It is a relative measure and is most suitable to compare any two series. As we know, the size of measure of dispersion also depends on the size of measurement. Hence, it is very appropriate measure of dispersion to compare two series which differ largely in respect of their means. All the more, a series or a set of value having lesser coefficient of variation as compared to the other is more consistent.

Combined Variance:

Suppose $\sigma_1^2, \sigma_2^2, \overline{X_1}, \overline{X_2}$ and N_1, N_2 are the variances, means and sizes of two groups of values respectively. Also let $\overline{X_{12}}$ be their combined mean. The combined variance of two groups is given by the formula.

$$\sigma_{12}^2 = \frac{N_1\left\{\sigma_1^2 + \left(\overline{X_1} - \overline{X_{12}}\right)^2\right\} + N_2\left\{\sigma_2^2 + \left(\overline{X_2} - \overline{X_{12}}\right)^2\right\}}{N_1 + N_2}$$

$$= \frac{N_1\left(\sigma_1^2 + d_1^2\right) + N_2\left(\sigma_2^2 + d_2^2\right)}{N_1 + N_2}$$

Where $\overline{X_1} - \overline{X_{12}} = d_1$ and $\overline{X_2} = \overline{X_{12}} = d_2$. The formula for combined varience can be extended to any number of groups.

Advantages:

Many times we know the means and variance of individual series or groups of data of known sizes. Then for some statistical analysis, their pooled variance is required. By the formula for pooled variance, it can easily be obtained without original data. Also a lot of time and labour is saved.

Advantages of standard deviation:

(i) Standard deviation carries great importance in sampling methods.

(ii) It is least sensitive to sampling fluctuations.

(iii) With the help of standard deviation, it is possible to as certain the area under the normal curve.

(iv) It has great utility in testing of hypotheses which other measures of dispersion hardly do.

Individual Series

$$\sigma = \sqrt{\frac{\sum d^2}{N}} \quad \text{where } d = x - \bar{x}$$

Discrete Series

$$\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2} \quad \text{where } d = x - A$$

Continuous Series

$$\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2} \times C \qquad \text{where } d = \frac{x - A}{C}$$

Example:

The prices of commodity in two towns are given below. In which town the prices are more consistent.

A Town Rs.	B Town Rs.
20	10
22	20
19	18
23	12
16	15

Solution:

Calculation of Standard Deviation

A Town

X	$(X - \bar{X})$ d	d^2
20	0	0
22	+ 2	4
19	- 1	1
23	+ 3	9
16	- 4	16
$\sum$ x 100		$\sum d^2$ 30

$$\bar{X} = \frac{\sum X}{N} = \frac{100}{5} = 20$$

$$\sigma = \sqrt{\frac{\sum d^2}{N}}$$

$$\sigma = \sqrt{\frac{30}{5}} = \sqrt{6} = 2 \cdot 44$$

$$C \cdot V \cdot = \frac{\sigma}{X} \times 100 = \frac{2 \cdot 44}{20} \times 100 = 12 \cdot 2\%$$

B Town

X	$(X - \bar{X})$ d	d^2
10	- 5	25
20	+ 5	25
18	+ 3	9
12	= 3	9
15	0	0
$\sum$ X 75		$\sum d^2$ 68

$$\overline{X} = \frac{\sum X}{N} = \frac{75}{5} = 15$$

$$\sigma = \sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{68}{5}} = \sqrt{13 \cdot 6} = 3 \cdot 6$$

$$C \cdot V. = \frac{\sigma}{\overline{X}} \times 100 = \frac{3 \cdot 6}{15} \times 100 = 24\%$$

Prices in A town are more stable than B town.

Example:

Calculate the standard deviation from the following data:

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

Size of item	**d = X - A = X - 9**	**Frequency**	**f d**	**$f d^2$**
6	- 3	3	- 9	27
7	- 2	6	- 12	24
8	- 1	9	- 9	9
9	0	13	0	0
10	+ 1	8	+ 8	8
11	+ 2	5	+ 10	20
12	+ 3	4	+ 12	36
		N = 48	f d = 0	$\sum fd^2 = 124$

$$\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2} = \sqrt{\frac{124}{48} - \left(\frac{0}{48}\right)^2}$$

$$\sum f d = 0 \quad \sum f d^2 = 124 \quad N = 48$$

$$\sigma = \sqrt{\frac{124}{48}} = \sqrt{2 \cdot 58}$$

$$\sigma = 1 \cdot 6$$

Example:

Lives of two models of refrigerators manufactured in a factory are given below:

Life (in years)	Refrigirators Centre	
	Model A	Model B
0 - 2	5	2
2 - 4	16	7
4 - 6	13	12
6 - 8	7	19
8 - 10	5	9
10 - 12	4	1

What is the average life of each model of these refrigerators? which model has more uniformity?

Solution:

For finding out the average life compute arithmetic mean. For determining the model which has greater uniformity compare the coefficient of variation.

Model A:

Calcultion of mean and coefficeint of variation

Life (No.of years)	f	m	$d' = \frac{m-5}{2}$	$f\,d'$	$f\,d'^2$
0 - 2	5	1	- 2	- 10	+ 20
2 - 4	16	3	- 1	- 16	16
4 - 6	13	5	- 0	0	0
6 - 8	7	7	1	+ 7	7
8 - 10	5	9	2	+ 10	20
10 - 12	4	11	3	+ 12	36
	N = 50		$f\,d_1 = +3$		$f\,d'^2 = 99$

$$\overline{X} = A + \frac{\sum f\,d'}{N} \times C$$

$A = 5, \sum f\,d' = 3, N = 50, C = 2 \cdot$

$$\overline{X} = 5 + \frac{3}{50} \times 2 = 5 + 0 \cdot 12 = 5 \cdot 12$$

Average life of Model A = 5.12

$$\sigma = \sqrt{\frac{\sum f d'^2}{N} - \left(\frac{\sum f d'}{N}\right)^2} \times C$$

$\sum f d'^2 = 99, \sum f d' = 3, N = 50, C = 2 \cdot$

$$\sigma = \sqrt{\frac{99}{50} - \left(\frac{3}{50}\right)^2} \times 2$$

$= \sqrt{1 \cdot 98 - 0 \cdot 0036} \times 2 \quad = \sqrt{1 \cdot 9764 \times 2}$

= 1.406 x 2 = 2.81

$$C \cdot V \cdot = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{2 \cdot 81}{5 \cdot 12} \times 100 \quad = 54 \cdot 9\ \%$$

Model B:

Calculation of Mean and Coefficient of Variation

Life (No.of hours)	f	m	$d' = \frac{m-5}{2}$	$f d'$	$f d'^2$
0 - 2	2	1	- 2	- 4	8
2 - 4	7	3	- 1	- 7	7
4 - 6	12	5	0	0	0
6 - 8	19	7	1	19	19
8 - 10	9	9	2	18	36
10 - 12	1	11	3	3	2
	N = 50			$f d' = 29$	$f d'^2 = 79$

$$\bar{X} = A + \frac{\sum f d'}{N} \times C$$

$A = 5, \sum f d' = 29, N = 50, C = 2 \cdot$

$$\bar{X} = 5 + \frac{29}{50} \times 2$$

$$= 5 + 16 = 6 \cdot 16$$

Average life of model B = 6.16

$$\sigma = \sqrt{\frac{\sum f d'^2}{N} - \left(\frac{\sum f d'}{N}\right)^2}$$

$$\sum f d'^2 = 79, \ \sum f d' = 29, \ N = 50, \ C = 2 \cdot$$

$$\sigma = \sqrt{\frac{79}{50} - \left(\frac{29}{50}\right)^2} \times 2$$

$$= \sqrt{1 \cdot 58 - 0 \cdot 3364} \times 2 = \sqrt{1 \cdot 2436} \times 2 = 1 \cdot 115 \times 2 = 2 \cdot 23$$

$$C \cdot V. = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{2 \cdot 23}{6 \cdot 16} \times 100 = 36 \cdot 2\%$$

B model has much uniformity

Example:

The following are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	**Girls**
Number	100	50
Mean Weight	60 Kgs	45 Kgs
Variance	9	4

i) Find out standard deviation of the combined data

ii) Which of the two distributions is more variable

Solution:

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 \times N_2}}$$

$$d_1 = \left(\overline{X_1} - \overline{X_{12}}\right)$$

$$d_2 = \left(\overline{X_2} - \overline{X_{12}}\right)$$

$$X_{12} = \frac{N_1\,\overline{X_1} + N_2\,\overline{X_1}}{N_1 + N_2}$$

$N_1 = 100,\ \overline{X_1} = 60,\ N_2 = 50,\ \overline{X_2} = 45 \cdot$

$$= \frac{(100 \times 60) + (50 \times 45)}{100 \times 50}$$

$$= \frac{6000 \times 2250}{150} = \frac{8250}{150} = 55 \text{ Kg}$$

$$d_1 = (60 - 55) = -5$$

$$C_2 = (45 - 55) = -10$$

$$\sigma_{12} = \sqrt{\frac{(100 \times 9) + (50 \times 4) + (100 \times 25) + (50 \times 100)}{100 + 50}}$$

$$= \sqrt{\frac{900 + 200 + 2500 + 5000}{150}}$$

$$= \sqrt{\frac{8600}{150}} = \sqrt{37 \cdot 33} = 61 \text{ Kgs}$$

ii) For finding out which of the two distributions is more variable we have to compare the coefficient of variation of the two distributions.

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

Boys: $C.V. = \frac{3}{60} \times 100 = 5$

Girls: $C \cdot V. = \frac{2}{45} \times 4 \cdot 44$

Since the coefficeint ofvariation is greater for boys than girls hence the first distirbution i.e., that of boys is more variable.

8.6 Skewness:

Lack of symmetry of tails (about mean) of a frequency distribution curve is known as skewness. Symetry of tail means that the frequency of the points at equal distances on both sides ofthe centre of the curve on X - axis is same. Also, the area under the curve at equidistant intervals on both sides of the centre is also equal. Departure from symmetry leads to skewness. It is adjudged by the elongation of the right and left tails of the curve.

Positive and Negative Skewness:

If the left tails of the frequency curve is more elongated than right tail, it is known as negative skewness and a reverse situation leads to positive skewness.

Different Formulae for measuring Skewness:

(i) Bowley's formula for measuring skewness interms of quartiles is:

$$S_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

(ii) Kelley gave the formula in terms of percentiles and deciles Kelley's absolute measure of Skewness are

$$S_k = P_{90} + P_{10} - 2P_{50}$$

$$= D_9 + D_1 = 2\ D_5$$

The formula are not practically used. Instead, it is measured as coefficient of skewness which is given as,

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

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

Kelley's formula are seldom used.

(iii) Karl Pearson's measure of Skewness

$$S_k = \frac{\text{mean} - \text{mode}}{S \cdot D.}$$

(iv) Karl Pearson's formula for a wide class of frequency distributions in terms of moments is,

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

β_1 gives only the measure of skewness but not the direction of Skewness.

Example:

From the following data compute quartile deviation and the co-efficient of Skewness.

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

Solution:

Calculation of Skewness

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

(i) Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$$Q_1 = \text{size of } \left(\frac{N}{4}\right)^{th} \text{ Item}$$

$\frac{100}{4}$ th item = 25th item

It lies in the class 7.5 to 10.5

$$Q_1 = L_1 + \frac{N/4 - Cf}{f} \times i$$

$L_1 = 75$ $N/4 = 25$ $Cf = 14$ $i = 3$ $f = 24$

$$Q_1 = 7 \cdot 5 + \frac{25 - 14}{24} \times 3$$

$$= 7\cdot5 + \frac{33}{24} = 7\cdot5 + 1\cdot37 = 8\cdot87$$

$Q_3 =$ size of $3\left(\frac{N}{4}\right)$ th item, 75th item

It is lies in the class 10.5 to 13.5

$$Q_3 = L_1 + \frac{3\times N/4 - Cf}{f}\times i$$

$L_1 = 10\cdot5, \quad N = 100, \quad Cf = 38, \quad i = 3\cdot$

$$Q_3 = 10\cdot5 + \frac{75-38}{38}\times 3$$

$$= 10\cdot5 + \frac{111}{38} = 10\cdot5 + 2\cdot92$$

$$= 13\cdot42$$

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

$$= \frac{13\cdot42 - 8\cdot87}{2} = \frac{4\cdot55}{2} = 2\cdot275$$

(ii) Bowley's coefficient of Skewness $= \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$

Median = Size of $\frac{N}{2}$ th item or $\frac{100}{2}$ th item = 50th item

It lies in the class 10.5 to 13.5

Median $L_1 + \frac{\frac{N}{2} - Cf}{f}\times i$

$L_1 = 10\cdot5$, $N/2 = 50$, $Cf = 38$, $f = 38$, $C = 3\cdot$

$$M = 10\cdot5 + \frac{50-38}{38}\times 3$$

$$= 10 \cdot 5 + \frac{36}{38} \quad = 10 \cdot 5 + 0 \cdot 947 \; = 11 \cdot 447$$

$$SK_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

$$= \frac{13 \cdot 42 + 8 \cdot 87 - 2 \times 11 \cdot 447}{13 \cdot 42 - 8 \cdot 87}$$

$$= \frac{22 \cdot 29 - 22 \cdot 89}{4 \cdot 55} = \frac{-0 \cdot 6}{4 \cdot 55} = -0 \cdot 13$$

Kelly's Measure of Skewness:

Example:

Calculate percentile coefficient of skewness from the following positional measure given below:

Solution:

$$P_{90} = 101, \; P_{10} = 58 \cdot 12, \; M = 79 \cdot 06.$$

$$S_K = \frac{P_{90} + P_{10} - 2M}{P_{90} - P_{10}}$$

$$= \frac{101 + 58 \cdot 12 - 2(79 \cdot 06)}{101 - 58 \cdot 12}$$

$$= \frac{159 \cdot 12 - 158 \cdot 12}{42 \cdot 88} = \frac{1}{42 \cdot 88}$$

$$S_K = +0 \cdot 023$$

8.7 Summary:

In this lesson, we have shown how the concept of range, quartile deviation, Mean deviation, variance, stnadard deviation and skewness. Also concept of coefficient of variation was introduced and compare relative variations of different data.

8.8 Exercises:

1. Discuss the importane of Measuring variability for managerial decision making.

2. What are different measurees of dispersion? How would you calculate them from a given frequency distribution? Briefly discuss the relative merits of different measures of disersion.

3. Why is Standard Deviation regarded as superior to other measures of dispersion? Explain with example.

4. What is coefficient of variation? What are the relative advantages fof a relative measure to that of an absolute measure?

5. How does Skewness differ from Dispersion? What is the objective of measuring Skewness?

6. Find out coefficient of quartile deviation.

Height in inches	53	55	57	59	61	63	65	67	69
Students	25	21	28	20	18	24	22	18	23

7. Following are the prices of shares of a company from Monday to Friday.

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Price	670	678	750	705	720

Compute the value of range and interpret the value.

8. Calculate the coefficient of range from the following data:

Sales (Rs. Lakhs)	No.of Days	Sales (Rs. Lakhs)	No.of Days
30 - 40	12	60 - 70	19
40 - 50	18	70 - 80	13
50 - 60	20	80 - 90	8

9. Compute the range and the quartile deviation for the following data:

Monthly Wage (Rs.)	No.of Workers
700 - 800	28
800 - 900	32
900 - 1000	40
1000 - 1100	30
1100 - 1200	25
1200 - 1300	15

10. Find out Quartile deviation

Weekly Wages in Rupees	No.of Wages Earners
35 - 36	14
36 - 37	20
37 - 38	42
38 - 40	54
40 - 41	45
41 - 42	21
42 - 43	8

11. Calculate the appropriate measure of dispersion from the following data:

Weekly Wages in Rupees	No.of Wage Earners
Lessthan 35	14
35 - 37	62
38 - 40	99
41 - 43	18
Over 43	7

12. Find out mean deviation from mean and its coefficient.

Marks: 30 45 80 40 50 65 80 90 60 40

13. Find out mean deviation from the following:

Marks	10	15	20	30	40	50
Students	8	12	15	10	3	2

14. Calculate coefficient of mean deviation

Marks less than	10	20	30	40	50
Students	5	13	25	28	30

15. Find out standard deviation and coefficient of variation Index No. of shares 108, 107, 105, 105, 106, 107, 104, 103, 104, 101.

16. Calculate coefficient of variation

The following are the runs scored by two batsmen A and B in ten innings.

A	101	27	0	36	82	45	7	13	65	14
B	97	12	40	96	13	8	85	8	56	15

Who is the more consistent batsman?

17. From the prices of shares X co and Y co are given below:

State which Co's share is more stable in value.

Share X Co. Rs:	55	54	52	53	56	58	52	50	51	49
Share Y Co. Rs:	108	107	105	105	106	107	104	103	104	101

18. Find only one of the measures of dispersion from the following data:

Profit (in Rs.)	No.of Firm
5000 to 6000	10
4000 to 5000	15
3000 to 4000	30
2000 to 3000	10
1000 to 2000	5
0 to 1000	4
- 1000 to 0	6
- 2000 to - 1000	8
- 3000 to - 2000	10

19. Calculate the standard deviation from the following:

Marks more than	0	10	20	30	40	50	60	70
No.of Students	100	90	75	50	25	15	5	0

20. Fifty items sold in debt A had a mean price of Rs. 30. 75 items sold in debt B had a mean price of Rs. 20 Find out the mean price of commodities sold in debt A and B.

21. Mean and standard deviation of two distribution of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the mean and stnadard deviation of all the 250 items taken together.

22. In the following distribution the mean is 132. Find out standard and deviation.

Income	Persons
100 - 110	2
110 - 120	4
120 - 130	7
130 - 140	?
140 - 150	5
150 - 160	2
160 - 170	1

Find out the missing frequency.

23. Find Bowley's coefficient of Skewness for the following frequency distribution.

No.of Children per family	0	1	2	3	4	5	6
No.of Families	7	10	16	25	18	11	8

24. Calculate the coefficient of Skewness for the following distribution of daily wages.

Wages (Rs.)	No.of Workers
4.5	35
5.5	40
6.5	48
7.5	100
8.5	125
9.5	87
10.5	43
11.5	22

25. Calculate Pearson's co-efficient of Skewness from the table given below:

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

26. In a certain distribution the following result were obtained.

$\overline{X} = 45$, Median 48,. Coefficient o Skewness = - 0.4

Find out standard deviation.

27. The weekly wages earned by 100 workers of a factoru are set out in the following table. Compute Bowley's coefficient of Skewness.

Weekly Wages:	25 - 35	35 - 45	45 - 55	55 - 65	65 - 75
No.of Workers:	12	16	25	15	13
Weekly Wages:	75 - 85	85 - 95	95 - 105	105 - 115	
No.of Workers:	10	5	3	1	

28. For a group of 50 Male workers the mean and standard deviation of their weekly wages are Rs. 63 and Rs. 9 respectively. For a group of 40 Female workers these are Rs. 54 and Rs. 6 respectively. Find the standard deviation of the combined group of workers and also combined mean.

8.9 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: M. KOTESWARA RAO

Lesson - 9

BASIC CONCEPTS OF PROBABILITY

Objectives:

After reading this lesson you should be able to

- Importance of probability in decission making.
- Understand the different approaches to probability.
- Calculate probabilities in different situations.
- To learn the rules of probability when events are independent and interdependent.

Structure:

9.1 Introduction:

The word 'probability' or 'chance' is very commonly used in day - to - day conversation. The moment the word chance is used to indicate an element of uncertainty. Thus the theory of probability provides a numerical measure of the element of uncertainty. The theory of probability is one of the most useful and interesting branches of modern mathematics. It is becoming prominent by its application in many fields of learning such as business, insurance, statistics, biological sciences, physical sciences, Engineering, etc....

9.2 Basic Concepts:

9.2.1 Random Experiment: It is an experiment which if conducted repeatedly under homogeneous condition does not give the same result. The result may be any one of the possible outcomes, the experimetn is called a random trail or random experiment.

The outcomes are known as elementary events and a set of outcomes is an event. Thus an elementary a event is also an event.

Example 1: A coin is tossed and get head it does not mean if we toss it again we get a head. That is the outcome will be any one of the possible outcomes (head or tail)

9.2.2 Sample Space: The set of all possible simple evetns in a trial is called a sample space for the trial. Each element of a simple space is called a sample point. Any sub set of a sample space is an event. It is generally denoted by E. Thus a simple event is a sample - point.

Sample space is denoted by S.

Eample 2: Two coins are tossed. Then the possible simple events of the trial are HH, HT, TH, TT.

$\therefore$ The sample space of the trial S = {HH, HT, TH, TT}

If an event (E) is that either two heads or two tails appear then E = {HH, TT}.

Clearly the elements of E, are sample points and $E \leq S$.

9.2.3 Event and Trial: Any particular performance of a random experiment is called a trial and outcome or combination of outcomes are termed as events.

Example 3:

(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) In an experiment which consider of the throw of a six - faced die and observing the number of points that appear, the possible outcomes are 1, 2, 3, 4, 5, 6.

In the same experiment, the possible events could also be stated as 'odd number of points', 'Even number of points', 'Getting a point greater than 4', and so on.

Event is called simple if it corresponds to a single possible outcome of the experiment otherwise it is known as a compound or composite event. Thus in tossing of a single die the event of getting '6' is a simple event but the event of getting an even number is a composite event.

9.2.4 Exhaustave Events or Cases: The total number of possible outcomes of a random experiment is known as the exhaustive events or cases.

Example 4:

(i) In tossing of a coin, there are two exhaustive cases viz, head and tail (the possibility of the win standing on an edge being ignored).

(ii) In drawing two cards from a pack of cards, the exhaustive number of cases is ${}^{52}C_2$, since 2 cards can be drawn out of 52 cards in ${}^{52}C_2$ ways.

9.2.5 Favourable Events or Cases: The number of outcomes of a random experiment which result in happening of a desired event are called favourable cases.

Example 5:

(i) In drawing a card from a pack of cards the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

(ii) In throwing of two dice, the number of cases favourable to getting the sum 5 is (1, 4), (4, 1), (2, 3), (3, 2) i.e., 4.

9.2.6 Mutually Exclusive Events: Events are said to be mutually exclusive, if the happening of any one of the events in a trial excludes the happening of any one of the others i.e., if no two or more of the events can happen simultaneously in the same trial.

Example 6:

(i) The toss of a coin the events 'head' or 'tail' are mutually exclusive because if head comes, e cannot get tail and if tail comes, we cannot get head.

(ii) Similarly in throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others, in the same trial is ruled out.

9.2.7 Equally Likely Events: The outcomes are aid to be equally likely or equally probable if none of then are expected to occur in preference to the other. In tossing of a coin all the outcomes of head or tail are equally likely if the coin is not biased.

Example 7:

(i) In a random toss of an unbiased or uniform coin head and tail are equally likely events.

(ii) In throwing an unbiased die, all the six faces are equally likely to come.

9.2.8 Independent Events: Several events are said to be independent if the happending (or non happening) of an event is not affected by the supplementary knowledge concerning the occurence of any number of the remaining events.

Example 8:

(i) When a die is thrown twice, the result of the first throw does not effect the result of the second throw.

(ii) In tossing an unbiased coin, the event of getting ahead int he first toss is independent of getting a head in the second, third and subsequent throws.

9.2.9 Collectively Exhaustive Events: The total number of possible out comes of a random experiment is called "Collectively exhaustive events" for the experiment.

Example 9: In toss of a single coin, exhaustive number of cases is 2, in a throw of a dice exhaustive number of cases is 6, and in case of throw of two dice, exhaustive number of cases are $6^2 = 36 \cdot$

9.2.10 Concepts of Permutations, Combinations and Their use in Probability:

(a) Permutations: Permutations refer to separate arrangement of different objects contained in a set of elements.

Example 10: If seven alphabets A, B, C, D, E, F, G are to be arranged by taking two letters at a time without containing same letter, (like AA, BB, etc...) the following permutations are possible.

AB	AC	AD	AE	AF	AG
BA	BC	BD	BE	BF	BG
CA	CB	CD	CE	CF	CG
DA	DB	DC	DE	DF	DG
EA	EB	EC	ED	EF	EG
FA	FB	FC	FD	FE	FG
GA	GB	GC	GD	GE	GF

Hene there are $7 \times 6 = 42$ permutations.

The number of different permutations of 'n' different objects taken 'r' at a time without repetition is

$$^{n}P_{r} = n\left[(n-1)(n-2) \cdot \cdot \cdot (n-r+1)\right]$$

(b) Combinations: A combination is a selection of objects considered without regard to their arrangements. In a permutation the order of the grouped items is important i in combination the order does not matter. Combinations are arrangements of items where in order is not important and duplication of components is inadmissible.

Example 11: If letters A, B, C, D, E are to be arranged in rows, but the same letters are not to be used at the same time.

The permutations will be $n \times (n-1) \cdot \cdot \cdot = 5 \times 4 = 20$

AB	AC	AD	AE
BA	BC	BD	BE
CA	CB	CD	CE
DA	DB	DC	DE
EA	EB	EC	ED

In case of combinations

AB	AC	AD	AE
BC	BD	BE	
CD	CE		
DE			

A combination of 'n' different objects taken 'r' at a time, denoted by $^{n}C_{r}$ or $\binom{n}{r}$, a selection of only 'r' objects out of 'n' objects, without any regard to the order of arrangements.

$$^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{2!(5-2)!} = \frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times1} = 10 \quad \text{(or)}$$

$$^{n}C_{r} = \frac{^{n}p_{r}}{r!} = \frac{20}{2\times1} = 10$$

Example 2:

1. Out of 4 officers and 10 clerks in a business firm, a committee consisting of 2 officers and 3 clerks is to be formed. In how many ways can this be done it.

(a) Any Officer and any clerk can be included

(b) One particular clerk must be on the committee

(c) Two particular officers cannot be on the committee

Solution:

(a) 2 officers out of 4 car be selected in:

$= {}^{4}C_{2}$ ways

3 clerks out of 10 can be selected at

$= {}^{10}C_{3}$ ways

Total number of possible selections

$= {}^{4}C_{2} \times {}^{10}C_{3} = 720$

(b) One particular clerk must be in the selection committee

2 officers out of 4 can be selected

$= {}^{4}C_{2}$ ways

= additional clerks out of a can be selected

$= {}^{3}C_{2}$ ways

Total number of possible selections

$= {}^{4}C_{2} \times {}^{9}C_{2} = 216$ ways

(c) When two particular officers should not be included:

2 officers out of remaining 2 can be selected in ${}^{2}C_{2}$ ways

3 clerks out of 10 can be selected in ${}^{10}C_{3}$ ways

Total number of possible selections

$${}^{2}C_{2} \times {}^{10}C_{3} = 120 \text{ ways}$$

Example 13: A bag contains 4 white, 5 red and 6 green balls. Three balls are drawn at randim what is the chance that a red and a green balls are drawn?

Solution:

Total number of balls in a bag = 4 + 5 + 6 = 15 balls

3 balls can be drawn out of 15 in $= {}^{5}C_{3}$ ways

Possibility of drawing a white balls $= {}^{4}C_{1}$

Possibility of drawing a red ball $= {}^{5}C_{1}$

Possibility of drawing a green ball $= {}^{6}C_{1}$

The total number of favourable cases for drawing three balls subsequently is ${}^{4}C_{1} \times {}^{5}C_{1} \times {}^{6}C_{1}$

P (drawing red and green balls in three draws) $= \dfrac{{}^{4}C_{1} \times {}^{5}C_{1} \times {}^{6}C_{1}}{{}^{15}C_{3}} = \dfrac{24}{91}$

9.2.11 Concept of Probability: The probability of a given event is an expression of likelihood or chance of occurrence of an event. A probability is a number which ranges between zero to one. The general rule of the happening of an event is that if the event can happen in 'm' ways and fail to happen in 'n' ways. The probability of the happening of the event is $P = {}^{m}/_{m+n}$ i.e., number of cases favourable to the event divided by number of exhaustive cases. However, broadly speaking there are four different schools of thought on the concept of probability.

9.3 Different Approaches to Probability Theory Axioms:

9.3.1 Mathematical or Classical or 'A Priori' Probability:

If a random experiment or a trial results in 'n' exhanstive, mutually exclusive and equally likely outcomes (or cases), out of which 'm' are favourable to the occurrence of an event E, then the probability 'p' of occurrence (or happening) of E, usually denoted by P(E) is given by

$$p = p(E) = \frac{\text{number of favourable cases}}{\text{total number of exhaustive cases}} = \frac{m}{n}$$

This definition was given by janes bernoulli who was the first person to obtain a quantitaitve measure of uncertainty.

Example:

(a) What is the probability of getting 'head' in a throw of a coin?

(b) If who coins are tassed once, what is the probability of getting.

(i) both heads?

(ii) At least one head?

Solution:

(a) When are coin is tassed, the outcomes are two head or tail

n = 2

out come 'head; is favourable (m) = 1

Hence, $P(\text{Head}) = 1/2$

(b) In case of two coins possible out comes are

(i) Two heads (H, H)

(ii) One head, one Tail (H, T)

(iii) One Tail in one coin and Head in second coin (T, H)

(iv) Two tails (TT)

So $n' = 4$

Favourable cases are : both Heads in only one case (m = 1)

Hence $P(H\ H) = 1/4$

(c) In case HH, HT, TH we get at least one head

P (at least one head) = 3/4 or 1 - 1/4

Example:

A ball is drawn at random from a box containing 6 white, 8 red and 10 green balls. Determine the probability of a ball drawn.

(i) White, (ii) Red (iii) Green (iv) Not Red (v) Red or Green

Solution:

Total number of balls = 6 + 8 + 10 = 24

n = 24

(i) Probability of drawing a white ball $= {}^{6}\!/_{24} = {}^{1}\!/_{4}$

(ii) Probability of drawing a red ball $= {}^{8}\!/_{24} = {}^{1}\!/_{3}$

(iii) Probability of drawing a green ball $= {}^{10}\!/_{24} = {}^{5}\!/_{12}$

(iv) Probability of drawing a ball which is not red $= {}^{16}\!/_{24} = {}^{2}\!/_{3}$

(v) Probability of drawing red or green ball $= \frac{8+10}{24} = \frac{18}{24} = \frac{6}{8} = \frac{3}{4}$

9.3.2 Relative Frequecny Probability: Under this approach the probability of an event represents the proportion of times, under identical circumstances, the outcome can be expected to occur. The value is the relative frequency of occurrence.

If the experiment is repeated for a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event 'A' happens, is called the probability of the occurrence of "A". The posterior probability or empirical probability of an event is thus

$$P = \frac{\text{Relative Frequency}}{\text{Number of trials}}$$

Example:

The following data relates to the length of life of wholesale grocers in a particular city.

Length of life (Yrs)	Percentage of Wholesalers
0 - 5	65
5 - 10	16
10 - 15	9
15 - 25	5
25 and above	5
	100

(i) During the period studied, what is the probability that an entrant to this profession will fail within five years?

(ii) That he will survive at least 25 years?

(iii) How many years would he have to survive to be among 10 percent longest survivers?

Solution: Total number of cases = 100

(i) The number of favourable cases to the condition that an entrant to the profession will fail within fire years = 65. Hence

$$\text{P(Entract will fail in the profession with in 5 years)} = \frac{65}{100} = 0 \cdot 65$$

(ii) The number of favourable cases to the condition of above 25 years of life is 5 and hence

$$\text{P(Survival after 25 years)} = \frac{5}{100} = 0 \cdot 05$$

(iii) The number of favourable cases for a wholesales with more than 15 years age is 5 + 5 = 10, Hence

$$\text{P (Survival after 15 years)} = \frac{10}{100} = 0 \cdot 01$$

Example: The following gives a distribution of monthly wages of 2000 employees of a firm.

Wages (in Rs.)	No.of Workers
Below 280	18
280 - 320	236
320 - 360	956
360 - 400	400
400 - 440	284
440 - 480	70
480 above	36

If an individual is selected at random from the above groups, what is the probability that his wages are (i) under Rs. 320? (ii) above Rs. 400 ? and (iii) Between Rs. 320 and Rs. 400?

Solution:

(i) Total number of wage earners = 2000

Total wage earners below Rs. 320 = 18 + 236 = 254

$$\text{P (Individual selected is under Rs. 320)} = \frac{254}{2000} = \frac{127}{1000}$$

(ii) No.of wage earners earning wages of Rs. 400 and above per month

= 284 + 70 + 36 = 390

P (Individual getting wages above Rs. 400) $= \frac{390}{2000} = \frac{39}{200}$

(iii) No.of wage earners in the wage group of Rs. 320 to Rs. 400 are 956 + 400 = 1356.

P (Individual in the range of Rs. 320 to Rs. 400) $= \frac{1356}{2000} = \frac{339}{500}$

9.3.3 Subjective Probability: The probabilities (chances) of occurrence of the corresponding events are assigned by individuals and are based on their personal judgement, wisdom, intuition and expertise. These probabilities are called the subjective probabilities and represent the degree of belief and the confidence one has in the occurrence of the respective event.

9.3.4 Axiomatic Definition of Probability: The classical definition of probability breaks down when we do not have a complete priori analysis i.e., when the outcomes of the trial are not equally likely or when the total number of trials is infinite or when the enumeration of all equally likely events is not possible. So the necessary of the following definition viz., statistical or empirical definition of probability.

9.3.5 Von Misses's Statistical Definition of Probability: 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 occur to the number 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.

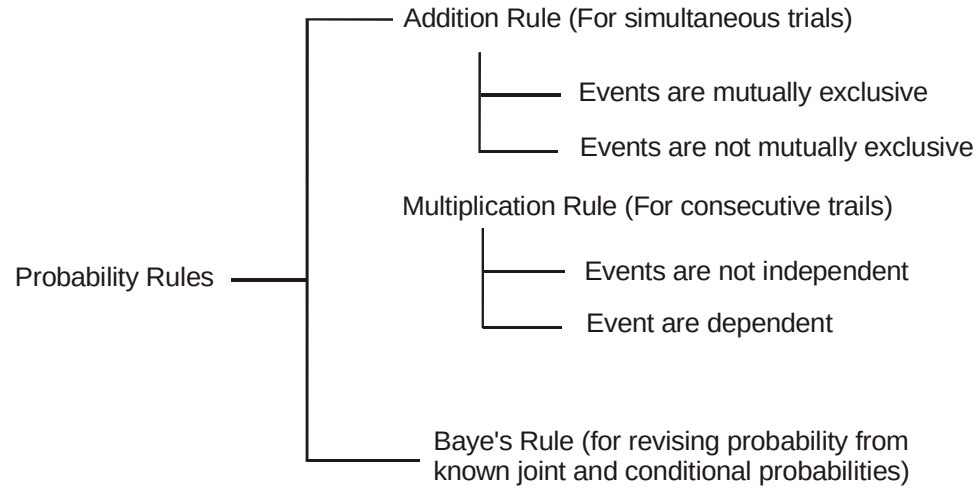
Symbolically if in 'N' trials an event E happens M times, then the probability of the happening of E, denoted by P(E) is given by

$$P(E) = \lim_{N \to \infty} \frac{M}{N}$$

9.3.6 Improtance of Probability: Though probability theory started with the game of chances i.e., from gambling. It has assumed a greate importance in almost all fields of study. It is the foundation of the classical decision procedures of estimation and testing. Highlighting the importance of probability theory Ya Lun Choce has beautifully pointed out that "statistics as a method of decision making under uncertainty is founded on probability theory". Since probability is at once the language and measure of uncertainty and the risks associated with it. Before learning statistical decision procedure the reader must acquire an understanding of probability theory.

9.4 Calculating Probability in Complex Situations:

9.4.1 Probability Rules: The solutions to many problems involving probabilities require a through understanding of some basic rules which govern the manipulations of probabilities. They are generally called probability rules.



9.4.2 Addition Theorem on Probability:

If 'S' is a sample space, and A , B are any events in 'S' then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: A bag contains 4 green, 6 black and 7 white balls. A ball is drawn at random what is the probability that it is either a green or a black ball?

Solution: Let s be the sample space associated with the drawing of a ball from a a bag containing 4 green, 6 black and 7 white balls.

$$\therefore n(s) = {}^{17}C_1 = 17$$

Let E_1 denote the event of drawing a green ball and E_2 that of a black ball.

$$\therefore n(E_1) = 4,\ n(E_2) = 6 \cdot$$

$$\therefore P(E_1) = \frac{n(E_1)}{n(s)} = \frac{4}{17},\ P(E_2) = \frac{n(E_2)}{n(s)} = \frac{6}{17}$$

E_1, E_2 are mutually exclusive i.e., $E_1 \cap E_2 = \phi$

$\therefore$ Probability of drawing either a green or a black ball $= P(E_1 \cup E_2)$

$$= P(E_1) + P(E_2) = \frac{4}{17} + \frac{6}{17} = \frac{10}{17}$$

Example: A card is drawn from a well shuffled pack of cards. What is the probability that it is either a spade or an ace

Solution: Let 's' is the sample space of all the simple events

$$\therefore \quad n(s) = 52$$

Let 'A' denote the event of getting a spade and B denotes the event of getting an ace

Then $A \cup B =$ Theevent of getting a spade or an ace.

$A \cap B =$ The event of getting a spade and an ace.

$$P(A) = \frac{13}{52}, \quad P(B) = \frac{4}{52}, \quad P(A \cap B) = \frac{1}{52}$$

By addition theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Example: Three students A, B, C are in running race. A and B have the same probability of winning and each is twice as likely to win as C. Find the probability thta B or C wins.

Solution: $A \cup B \cup C = s =$ Sample space of race

By data $P(A) = P(B)$ and $P(A) = 2\,P(C)$

We have $P(A) + P(B) + P(C) = 1$

$$2\,P(C) + 2P(C) + P(C) = 1$$

$$\Rightarrow P(C) = \frac{1}{5}, P(A) = \frac{2}{5} \text{ and } P(B) = \frac{2}{5}.$$

The probability that B or C wins = $P(B \cup C)$

$$= P(B) + P(C) - P(B \cap C)$$

$$= \frac{2}{5} + \frac{1}{5} - 0 = \frac{3}{5}$$

9.4.3 Multiplication Theorem of Probability: in a random experiment if E_1, E_2 are two events such that $P(E_1) = 0$ and $P(E_2) = 0$ then

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_1/E_2)$$

$$P(E_2 \cap E_1) = P(E_2) \cdot P(E_1/E_2)$$

Example: A bag contains 4 white and 3 black balls. A ball is drawn out of it and replaced in the bag. Then a ball is drawn again. What is the probability that

(i) Both the balls drawn were black

(ii) Both were white

(iii) The first ball is black and second white

(iv) The first ball was white and second black

Solution: The events are independent and capable of simultaneous occurrence. The rule of multiplication would be applied.

The probability that

(i) Both are black $= \frac{3}{7} \times \frac{3}{7} = \frac{9}{49}$

(ii) Both are white $= \frac{4}{7} \times \frac{4}{7} = \frac{16}{49}$

(iii) First is black and second is white $= \frac{3}{7} \times \frac{4}{7} = \frac{12}{49}$

(iv) The first one is white and second black $= \frac{4}{7} \times \frac{3}{7} = \frac{12}{49}$

Example: The students A, B, C are given a problem in statistics, the probability of their solving the problem are $\frac{3}{4}$, $\frac{1}{4}$ and $\frac{2}{4}$ respectively. What is the probability that if all of then try the problem would be solved?

Solution: Probability that A solves and B and C fails $= \frac{3}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{18}{64}$

Probability that B solves and C fails $= \frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$

Probability that C solves A and B fails $= \frac{1}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{6}{64}$

Probability that A and B solve and C fails $= \frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{6}{64}$

Probability that A and C solve B fails $= \frac{3}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{18}{64}$

Probability that B and C solve and A fails $= \frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$

Probability that A, B, C all solve $= \frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{6}{64}$

The probability that any one of the above situations happen as

$$= \frac{18}{64} + \frac{2}{64} + \frac{6}{64} + \frac{6}{64} + \frac{18}{64} + \frac{2}{64} + \frac{6}{64} = \frac{58}{64}$$

Example: A bag cotnains 8 red and 6 blue balls. Two drawing of each 2 balls are made. Find the probability than the first drawing gives two red balls and second drawing gives 2 blue balls, if the balls drawn first are replaced before the second draw.

Solution: $E_1 =$ Event pf drawing 2 red balls in the forst draw from the bag containing 8 red and 6 blue balls.

$\therefore P(E_1) = \frac{{}^8C_2}{{}^{14}C_2}$ The two drawn balls are replaced.

$E_2 =$ Event of drawing 2 blue balls in the second draw fromt he bag.

$\therefore P(E_2) = \frac{{}^6C_2}{{}^{14}C_2}$

Now $E_1 \cap E_2 =$ Event of drawing 2 red balls inthe first draw and another drawing of 2 blue balls in the second draw after the balls are replaced.

Also E_1 , E_2 are independent.

$$\therefore P(E_1 \cap E_2) = P(E_1) \times P(E_2) = \frac{{}^8C_2}{{}^{14}C_2} \times \frac{{}^6C_2}{{}^{14}C_2} = \frac{8 \times 7}{14 \times 13} \times \frac{6 \times 5}{14 \times 13}$$

9.4.4 Types of probability:

Basically there are three types of probabilities.

1. Marginal, 2. Joint, 3. Conditional.

1. Marginal Probability: Marginal Probability are also otherwise called as conditional probabilities. Marginal prbabilities are the probabilities of single event which is expressed symbolically as P(A), the probability of the event of happening. Thus it is the simple probability of occurrence of an event.

2. Joint Probability: The probability of two or more independent events occuring together or in succession is the product of their marginal probabilities. Symbolically joint probability is dentoed as P(AB) and mathematically the joint probability is stated

$$P(AB) = P(A) \times P(B)$$

where

P(AB) - Probability of events A and B occuring together or in succession, known as joint probability.

P(A) - Marginal probability of event A occuring.

P(B) - Marginal probability of event B occuring.

3. Conditional Probability: Conditional probability is the probability that a second event (B) will occur if a first event (A) has already occurred. If the events are independent i.e., happening of A does not effect happening of B the conditional probability denoted by $P(A/B)$ will be P(A)

$$P(A/B) = P(A)$$

when the events are dependent

$$P(A/B) = {}^{P(AB)}\!/_{P(B)}$$

Example: A bag contains 5 red and 3 white balls. Two draws are made without replacement. What is the probability that both the balls are red?

Solution: Probability of drawing a red ball in the first draw is $P(A) = {}^{5}\!/_{8}$

Probability of drawinga red ball on the second draw given that first ball is red of $P(B/A) = 4/7$ (since only 7 balls are left and only 4 of then are red).

The combined probability of the two events is

$$P(AB) = P(A) \times P(B/A)$$

$$= 5/8 \times 4/7 = 20/56$$

Example: A bag contains 4 red and 6 green balls. Two draws of one ball each are made without replacement. What is the probability that one is red and the other is green?

Solution: Probability of drawing a red ball $P(A) = {}^{4}\!/_{10}$

Probability of drawing a green ball in the second draw given that the first draw has given a red ball $P(B/A) = 6/9$ (Since only a balls are left out of which 6 are green)

Probability of the combined event

$$P(AB) = P(A) \times P(B/A) = 4/10 \times 6/9 = 24/90$$

But it could also happen the first ball may be green and second ball is red.

Probability of drawing a green frist $P(B) = 6/10$ and red next (given green has been drawn) $P(A/B) = 4/9$

$$P(AB) = P(B) \times P(A/B) = {}^{6}/_{10} \times {}^{4}/_{9} = {}^{24}/_{90}$$

Now when any one of the two situations (red and green or green and red) can happen and both of them are mutually exclusive the required probability will be

$$P(A/B) = \frac{P(AB)}{P(B)}$$

$$= \frac{P(A)\ P(B/A)}{P(B)}$$

$$= \frac{P(A)\ P(B/A)}{P(A)\ P(B/A) + P(\text{Not } A)\ P(B/\text{Not } A)}$$

In asimilar way Bayes theorem can be extended to n events also. Application of Bayes theorem is apowerful method of evaluating new information in order to revise nor probabilities. When correctly used baye's theorem can be of tremendous aid in decision making.

9.5 Revising Probability Estimate:

Baye's Theorem:

$E_1, E_2, \cdots, E_n$ are 'n' mutually exclusive and exhanstive events such that $P(E_1) > 0\ (i = 1, 2, \cdots, n)$ in a sample space S and A is any other event in 'S' intersecting with every E_i (i.e., A can only occur in combination with every one of the events $E_1, E_2, \cdots, E_n$) such that P(A) > 0.

If E_i is any of the event of $E_1, E_2, \cdots, E_n$ where $P(E_1), P(E_2), \cdots, P(E_n)$ and $P\left[\frac{A}{E_1}\right] \cdot P\left[\frac{A}{E_2}\right] \cdots P\left[\frac{A}{E_n}\right]$ are known, then

$$P\left[\frac{E_k}{A}\right] = \frac{P(E_k) \cdot P(A/E_K)}{P(E_1) \cdot P(A/E_1) + P(E_2) \cdot (A/E_2) + \cdots + P(E_n) \cdot P(A/E_n)}$$

Example: In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body.

(a) What is the probability that mathematics is being studied.

(b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl?

(c) a boy?

Solution: Given P(Boy) = $P(B) = \frac{40}{100} = \frac{2}{5}$

$$P(G) = \frac{60}{100} = \frac{3}{5}$$

Probability that mathematics is studied given the student is a boy } $= P(M/B) = \frac{25}{100} = \frac{1}{4}$

Probability that mathematics is studied given the student is a girl } $= P(M/G) = \frac{10}{100} = \frac{1}{10}$

(a) Probability that maths is studied $= P(M) = P(G)\,P(M/G) + P(B) \quad P(M/B)$

$\therefore$ By total probability theorem

$$P(M) = \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{25}$$

(b) By Baye's theorem probability of maths student is a girl

$$= P(G/M) = \frac{P(G)\,P(M/G)}{P(M)} = \frac{3/5 \cdot 1/10}{4/25} = \frac{3}{8}$$

(c) Probability of maths student is a boy

$$= P(B/M) = \frac{P(B) \cdot P(M/B)}{P(M)} = \frac{2/5 \cdot 1/4}{4/25} = \frac{5}{8}$$

Example: The chance taht doctor A will diagnise a disease x correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease x, died what is the chance that his disease was diagnised correctly.

Solution: Let E_1 be the event that disease x is diagnised correctly by doctor A.

and E_2 be the event that a patient of doctor A who has disease x died?

Then $P(E_1) = \frac{60}{100} = 0\cdot6,\ P\left[\frac{E_2}{E_1}\right] = \frac{40}{100} = 0\cdot4$

$$P(\bar{E}) = 1 - 0\cdot6 = 0\cdot4\ ,\ P\left[\frac{E_2}{E_1}\right] = \frac{70}{100} = 0\cdot7$$

$\therefore$ By Bayes theorem $P\left[\frac{E_1}{E_2}\right] = \frac{P(E_1)\cdot P(E_2/E_1)}{P(E_1)\cdot P(E_2/E_1) + P(\overline{E_1})\cdot P(E_2/\overline{E_1})}$

$$= \frac{0\cdot6 \times 0\cdot4}{0\cdot6\times0\cdot4 + 0\cdot4\times0\cdot7} = \frac{6}{13}$$

Example: First box contains 2 black, 3 red, 1 white balls. Second box contains 1 black, 1 red, 2 white balls and third box contains 5 black, 3 red, 4 white balls of these a box find the probability that is from second box.

Solution: Let x, y, z be the first, second and third boxes.

$$\therefore\ p(x) = \frac{1}{3}\ ,\ p(y) = \frac{1}{3},\ p(z) = \frac{1}{3}\cdot$$

Let R be the event of drawing a red ball from a box

So $p(R/x) = \frac{3}{6}\ ,\ p(R/y) = \frac{1}{4}\ ,\ \ p(R/z) = \frac{3}{12}$

$\therefore$ By Baye's theorem, the required probability

$$= p(y/R)$$

$$= \frac{p(y)\cdot p(R/y)}{p(x)\cdot(R/x) + p(y)\cdot p(R/y) + p(y)\cdot p(R/z)}$$

$$= \frac{1/3 \times 1/4}{1/3\times3/6 + 1/3\times1/4 + 1/3\times3/12} = \frac{1}{4}$$

9.6 Summary:

Probability means the chance of occurrance of an event. Different approaches to probability measurement are given. In this lesson we have given addition theorem, multiplication theorem and Baye's theorem and given.

9.7 Exercise:

1. Define probability and explain the importance of this concept in statistics and with the help of a suitable example?
2. Explain with examples the concept of independence and mutually exclusive events in probability and with the help of a suitable example?
3. State the addition and multiplication therorems of probability for two mutually exclusive evetns? and with the help of a suitable example?
4. What is Bayes theorem? Explain with the help of a suitable example?
5. Define mathematical probability of an event A. If A and B are two events, can they be mutually exclusive as well as independent?
6. Find the probability that in a random arrangement of the letters of the word UNIVERSITY the two I's do not come together.
7. Three fair dice are thrown. What is the probability of getting a sum 6 or less on the three dice?
8. Five tickets are drawn N at random from a bag containing 50 tickets numbered 1, 2, 3,....., 50. The tickets are arranged in ascending order of magnitude $\left(x_1 < x_2 < x_3 < x_4 < x_5\right)$. Find the probability that $x_3 = 30$.
9. From 6 gentleman and 4 ladies a committee of 5 is to be formed. Find the probability that this can be done so as to always include at least one lady.
10. What is the probability that a card drawn at random from the pack of playing cards may be either a queen or a king.
11. A coin is tossed n times. What is the probbaility that the tail will present itself an odd number of times?
12. In a group there are 3 man and 2 women. Three persons are selected at random from this group. Find the probability that one man and two women or two men and one women are selected.
13. An integer is choosen at random from the first 200 positive integers. What is the probability that the integer chosen is divisible by 6 or by 8 ?
14. If $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$, $P(A \cap B) = \frac{1}{5}$ then find (i) $P(A \cup B)$ (ii) $P\left(A^c \cap B\right)$ (iii) $P\left(A \cap B^c\right)$ (iv) $P\left(A^c \cap B^c\right)$

15. A card is drawn from a well shuffled pack of cards. What isthe probability that it is either a spade or an ace.

16. A class has 10 boys and 6 girls. Three students are selected at random one after another. Find the probability that (i) First two are boys and third is girl, (ii) First and third of same sex and second is opposite sex.

17. Mr. X is selecrted for interview for 3 posts. For the first post there are 5 candidates, for the second there are 4 and for the third there are 6. If the selection of each condidate is equally likely, find the chance that Mr. X will be selected for atleast one post.

18. An anti-aircraft gun can take a maximum of 4 shots at an enemy plane moving away from it. The probabilities of hitting the plane at the first, second, third and fourth shots are 0.4, 0.3, 0.2 and 0.1 respectively. What is the probability thatthe gun hits the plane.

19. A die is tossed. If the number is odd what is the probbaility that it is prime.

20. In a certain town 40% have brown hair, 25% have brown eyes and 15% have both brown hair and brown eyes. A person is selected at random from the town.

    (a) If he has brown hair, what is the probbaility that he has brown eyes also?

    (b) If he has brown eyes, determine the probability that he does not hare brown hair?

    (c) Determine the probability that he has neither brown hair nor brown eyes.

21. Box I contains 1 white, 2 red, 3 green balls. Box II contains 2 white, 3 red, 1 green balls. Box III contains 3 white, 1 red, 2 green balls. Two balls are drawn from a box chosen at random. These are found to be one white and one red. Determine the probability that the balls so drawn came from box II.

22. Of the three men, the chances that a politician a business man or an academician will be appointedas a vice - chancellor (VC) of a university are 0.5, 0.3, 0.2 respectively. Probbaility that research is promded by these persons if they are appointed as VC are 0.3, 0.7, 0.8 respectively.

    (i) Determine the probability that research is promoted.

    (ii) If reserach is promoted, what is the probability that VC is an academician.

23. Companies B_1, B_2, B_3 produce 30%, 45%, and 25% of the cars respectively. It is known that 2%, 3% and 2% of the cars produced from B_1, B_2 and B_3 are defective.

    (a) What isthe probability that a car purchased is defective?

    (b) If a car purchased is found to be defective what is the probability that this car is produced by company B_3 ?

24. Suppose three companies x, y, z produce T.V.'s X produce twice as many as y while y and z produce the same number. It is known that 2% of x, 2% of y and 4% of y are defective. All the T.V.'s produced and put into one shop and then one T.V. is choosen at random.

(a) What is the probability that the T.V.'s is defective?

(b) Suppose a T.V. Chosen is defective what is probability that this T.V. is produced by company X?

25. A box contains 4 balls. Two balls are drawn from it and are found to be while find the probbaility that all the balls in the bag are white.

9.8 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: M. KOTESWARA RAO

Lesson - 10

DISCRETE PROBABILITY DISTRIBUTIONS

Objectives:

After reading this lesson you should be able to

- Understand the basic concepts of secondary variable and probability distribution.
- Importance of probability distribution is Business - decisions.
- Where to use the discrete probability distributions.
- Binomial and Poisson distribution applications.

Structure:

10.1 Introduction:

Probability distribution refer to mathematical models of relative frequencies of a finite number of observations of a variable by listing all the possible out comes of an experiment. These distributions may not fully agree with actual observations or the emperical distributions based on sample observations and it is likely that if the number of experiments in increased sufficiently the emprical distributions may approach closely to these theoritical probability distributions. Since, these distributions deals with expectations. more over, when experiments are not possible at all times, especially in case of business decisisons, these theoritical models will help the managerial decision making with certain amount of accuracy.

10.2 Basic Concepts:

10.2.1 Random Variable: We are considering a function whose domain isthe set of possible outcomes, and whose range is a subset of the set of reals such functionis called a random variable.

Intuitively by a random variable (r.v.) we mean a real numbes X connected with the outcome of a random experiment E.

For Example: If E consists of two tosses the random variable which is the number of heads (0, 1, 2)

Outcome (W)	H H	H T	T H	T T
Value of (X)	2	1	1	0

Thus to each outcome W, there corresponds a real number X(W). Since the points of the sample space s corresponds to outcomes, this means that a real number, which we denote by X(W), is defined for each $W \in S$.

For a mathematical and rigorou defination ofthe random variable, let us consider the probability space, the triplet (S, B, P) where S is the sample space i.e., space of outcomes B is the σ - field of subsets of in S, and P is a probability function on B.

Definition: A random variable (r.v.) is a function X(W) with domain 'S' and range $(-\infty, \infty)$ such that for every real number a, the event $[W : X(W) \leq a] \in B$.

The set of all possible values taken by a random variable X is called the image of X or range of X and it is a subset of real numbers R.

Notation: Random variables are usually denoted by capital letters of Engligh alphabets and particular values which the random variable takes are denoted by the corresponding small letters.

10.2.2 Types of Random Variables: Random variables are of two types.

(i) Discrete random variable

(ii) Continuous random variable

(i) Discrete random variable: A variable which can assume only a countable number of real values and for which the value which the variable takes depends on chance, is called a discrete random variable.

In otherwords, a real valued function defined on a discrete sample space is called a discrete random variable.

Example:

1) The random variabledenoting the number of telephone calls per unit time

2) The random variable denoting the number of successes in n trials

(ii) Continuous random variable: A random variable X is said to be continuous if it can take all possible values between certain limits.

Example:

1) The height X of a student in a particular class may be between 4 ft. and 6 ft.

$$X(S) = \{ x / 4 \le x \le 6 \}$$

This X is a continuous random variable.

10.2.3 Probability distribution function:

Definition: Let X be a random variable. The function F, defined for all real x by

$$F(x) = P(X \le x)$$

$$= P\{W : X(W) \le x\} \; ; \; -\infty < x < \infty$$

is called the distribution function (d.f.) of the r.v. (X)

A distribution function is also called the cumulative distribution function.

The domain of the distribution function in $(-\infty, \infty)$ and its range is [0, 1].

Properties of distributiob function:

1. If F is the distribution function of the random variable X and if a < b, then

(i) $P(a < x \le b) = F(b) - F(a)$

(ii) $P(a \le x \le b) = P(x = a) + [F(b) - F(a)]$

(iii) $P(a < x < b) = [F(b) - F(a)] - P(x = b)$

(iv) $P(a \le x < b) = [F(b) - F(a)] - P(x = b) + P(x = a)$

If $P(x = a) = P(x = b) = 0$ then

$$P(a < x \le b) = P(a \le x \le b) = P(a < x < b) = P(a \le x < b) = F(b) - F(a)$$

2. All distribution functions are monotonically increasing and lie between 0 and 1.

i.e., If F is the distribution function of the random variable X, then

(i) $0 \le F(x) \le 1$ (ii) $F(x) < F(y)$ when $x < y$

3. (i) $F(-\infty) = \underset{x \to -\infty}{Lt} F(x) = 0$

(ii) $F(\infty) = \underset{x \to \infty}{Lt} F(x) = 1$

10.3 Discrete Probability Distribution:

Suppose X is a discrete random variable with possible outcomes (values) $x_1, x_2, \cdots, x_n$.

The probability of each possible outcome x_i is p_i

$$p_i = P(X = x_i) = p(x_i) \quad \text{for } i = 1, 2, \cdots$$

If the numbers $p(x_i),\ i = 1, 2, \cdots$

Satisfy the two conditions

(i) $p(x_i) \geq 0$ for all values of i

(ii) $\sum p(x_i) = 1,\ i = 1, 2, \cdots$

The function p is called the probability mass function of the random variable X and the set $\{p(x_i)\},\ i = 1, 2, \cdots$ is called the discrete probability distribution of the discrete random variable X.

Example 1:

In tossing a coin two times

$$S = \{TT, HT, TM, HH\}$$

$P(X = 0) =$ Probability of getting two tails (no heads)

$$= P\{T, T\}$$

$$= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$P(X = 1) =$ probability of getting one head

$$= P\{(H, T) \cdot (T, H)\}$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$P(X = 2) =$ Probability of getting two head

$$= P\{(H, H)\}$$

$$= \frac{1}{4}$$

Thus the total probability 1 is distributed into three parts $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ according to whether $X = 0$ or 1 or 2.

This probability distribution is given in the following table

$\mathbf{X - x_i}$	0	1	2
$\mathbf{P(X = x_i)}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 2:

From a lot of 10 items containing 3 defeetives, a sample of 4 items is drawn a random. Let the random variable X denote the number of defective items in the sample. Find the probability distribution of X when the sample is drawn without replacement.

Solution:

Obviously X can taken the values 0, 1, 2 or 3.

Given total number of items = 10

Number of good items = 7

Number of defective items = 3

$$P(X = 0) = P(\text{no defective}) = \frac{^7C_4}{^{10}C_4} = \frac{7!}{4!\,3!} \times \frac{4!\,6!}{10!} = \frac{1}{6}$$

$P(X = 1) = P$ (one defective and 3 good items)

$$= \frac{^3C_1 \times {}^7C_3}{^{10}C_4} = \frac{3 \times 7!}{3!\,4!} \times \frac{4!\,6!}{10!} = \frac{1}{2}$$

$P(X = 2)\ P$ (2 defective and 2 good items)

$$= \frac{^3C_2 \times {}^7C_2}{^{10}C_4} = \frac{3}{10}$$

$P(X=3) = P$ (3 defective and 1 good item)

$$= \frac{{}^{3}C_{3} \times {}^{7}C_{1}}{{}^{10}C_{4}} = \frac{7}{{}^{10}C_{4}} = \frac{4!}{8 \times 9 \times 10} = \frac{1}{30}$$

$\therefore$ The probability distribution of random variable X is as follows:

X	0	1	2	3
$P(X)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

10.3.1 Discrete Distribution Function:

Suppose that X is a discrete random variable. Then the discrete distribution function or cumulative distribution function F(x) is defined by

$$F(x) = P(X \le x)$$

$$= \sum_{(i : x_i \le x)} p(x_i)$$

For example, if x_i is just the integer i.

So that $P(X=i) = p_i$; $i = 1, 2, \cdots$ then F(x) is a step function havign jump p_i at i, and being constant between each pair of integers.

Example: 3

If $p(x) = \begin{cases} \frac{x}{15}, & x = 1,2,3,4,5 \\ 0, & \text{else where} \end{cases}$

Find (i) $P\{X = 1 \text{ or } 2\}$ and ii) $P\left\{\frac{1}{2} < x < \frac{5}{2} \Big/ x > 1\right\}$

Solution:

(i) $P(X = 1 \text{ or } 2) = P(x=1) + P(x=2) = \frac{1}{15} + \frac{2}{15} = \frac{1}{5}$

(ii) $$P\left(\frac{1}{2} < x < \frac{5}{2} \Big/ x > 1\right) = \frac{P\left\{\left(\frac{1}{2} < x < \frac{5}{2}\right) \cap (x > 1)\right\}}{P(x > 1)}$$

$$= \frac{P\left\{(X = 1 \text{ or } 2) \cap (x > 1)\right\}}{P(X > 1)}$$

$$= \frac{P(X = 2)}{1 - P(X = 1)} = \frac{2/15}{1 - (1/15)}$$

$$= \frac{1}{7}$$

10.3.2 Bernoulli Process:

Any uncertain situation or experiment that is marked by the following three properties is known as a Bernowli process.

1. There are only two mutually exclusive and collectively exhaustive outcomes in the experiment.
2. In repeated observations of the experiment, the probabilities of occurerice of these events remain constant.
3. The observations are independent of one another.

10.4 Summary Measures and Their Application:

10.4.1 Expectation: Suppose a random variable X assumes the values $x_1, x_2, \cdots, x_n$ with respective probabilities $p_1, p_2, \cdots, p_n$. Then the expectation or expected value of X, denoted by $E(X)$, is defind as the sum of products of differnt values of X and the correspoinding probabilities.

i.e., $E(X) = \sum_{i=1}^{n} P_i\, x_i$

Similarly $E(x^r) = \sum_{i=1}^{n} p_i\, x_i^r$

In general, the expected value of any function g(x) of a random variable X is defined as

$$E\left\{g(x)\right\} = \sum_{i=1}^{n} p_i\, g(x_i)$$

10.4.2 Some Important Results on Expectation:

1. If X is a random variable and K is a constant then

$$E(X + K) = E(X) + K$$

2. If X is a random variable and a and b are constants, then

$$E(aX + b) = a\,E(X) + b$$

3. If X and Y are two discrete random variables, then

$$E(X + Y) = E(X) + E(y)$$

4. If X and Y are two independent random variables, then

$$E(XY) = E(X)\,E(Y)$$

Mean: The mean value μ of the discrete distribution function is given by

$$\mu = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i = E(X)$$

Variance: The Variance σ^2 of the discrete distribution function is given by

$$V(X) = \sigma^2 = E\left[\{X - E(X)\}^2\right]$$

$$V(X) = \sigma^2 = E(X^2) - [E(X)]^2, \quad \text{where } E(X^2) = \sum_{i=1}^{n} p_i x_i^2$$

Standard Deviation:

It is the positive square root of the variance

$$\therefore S\cdot D\cdot = \sqrt{\text{variance}} = \sqrt{\sigma^2} = \sigma$$

Some important result on variance:

(i) If X is a discrete random variable, then $v(ax + b) = a^2\,v(X)$

where $v(X)$ is variance of X and a, b are constants.

(i) If b = 0, then $v(ax) = a^2\,v(x)$

(ii) If a = 0, then $v(b) = 0$

(iii) If a = 1 then $v(x + b) = v(x)$

Example 4:

A random variable X has the following probability function.

x	0	1	2	3	4	5	6	7
p(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2 + k$

(i) Determine K

(ii) Mean

(iii) Variance

Solution:

(i) Since $\sum_{i=0}^{7} p(x) = 1$

$$k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k - 1 = 0$$

$$10\,k^2 + 9k - 1 = 0$$

$$10\,k^2 + 10K - k - 1 = 0$$

$$10k\,(k+1) - 1\,(k+1) = 0$$

$$(k+1)(10\,k - 1) = 0$$

$$(k+1) = 0 \ \text{(or)} \ (10k - 1) = 0$$

$$\Rightarrow k = \frac{1}{10} \quad \text{(or)} \quad k = -1$$

but $p(x) \geq 0$

$$\therefore k = \frac{1}{10}$$

x	0	1	2	3	4	5	6	7
p(x)	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{100}$	$\frac{2}{100}$	$\frac{17}{100}$

(ii) Mean

$$\text{Mean} = E(x) = \sum_{i=0}^{7} p_i\, x_i$$

$$= 0 + \frac{1}{10} + \frac{4}{10} + \frac{6}{10} + \frac{12}{10} + \frac{5}{100} + \frac{12}{100} + \frac{119}{100}$$

$$= \frac{23}{10} + \frac{136}{100} = \frac{366}{100} = 3 \cdot 66$$

(iii) Variance

$$\text{Variance} = \sum_{i=0}^{7} p_i\, x_i^2 - \mu^2$$

$$= \frac{1}{10} + \frac{8}{10} + \frac{18}{10} + \frac{48}{10} + \frac{25}{100} + \frac{72}{100} + \frac{833}{100} - (3\cdot 66)^2$$

$$= \frac{75}{10} + \frac{930}{100} - 13\cdot 3956$$

$$= \frac{750 + 930}{100} - 13\cdot 3956$$

$$= 16\cdot 8 - 13\cdot 3956$$

$$V(X) = 3\cdot 4044$$

10.5 Some Important Discrete Probability Distribution:

The value of the random variables and the corresponding probabilities are arranged such away that one can suit a mathematical function for the probabilities in terms of the value of the random variable. A good model requires sufficiently large number of observations.

10.5.1 Bernoulli's Distribution:

A random variable x which takes two values 0 and 1 with probability q and p respectively. i.e., $p(x=0) = q$, $q = 1-p$ is called a Bernoulli's discrete random variable and is said to have a Bernoulli's distribution.

The probability function fo Bernoulli's distribution can be written as

$$P(x) = p^x\, q^{1-x} = \begin{cases} p^x\,(1-p)^{1-x} & , \quad \text{for } x = 0, 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

The parameter p statisfies $0 \le p \le 1$

(i) Mean:

Mean of the Bernoulli's discrete random variable x is

$$E(x) = \sum x_i\, p(x_i) = 0 \times q + 1 \times p = p$$

$$E(x) \text{ or } \mu = p$$

(ii) Variance:

Variance of x is

$$v(x) = E(x^2) - [E(x)]^2 = \sum x_i^2 - p(x_i) - \mu^2 = 0^2 \times q + 1^2 \times p - p^2 = p - p^2$$

$$= p(1-p)$$

$$\sigma^2 = v(x) = pq$$

(iii) Standard Deviation:

Standard Deviation of x is

Standard Deviation $= \sqrt{\text{variance}}$

$$= \sqrt{\sigma^2} = \sqrt{pq}$$

$\therefore$ standard deviation $(\sigma) = \sqrt{pq}$

10.5.2 Binomial Distribution:

Binomial distribution was discovered by james bernoulli in the year 1700 and it is a discrete probability distribution.

Definition:

A random variable x has a Binomial distribution if it assumes oly non - negative values and its probability distribution is as follows:

$$P(X = x) = p(x) = \begin{cases} {}^nC_x \; p^x q^{n-x} & , \quad x = 0,1,2, \cdots, n \\ 0 & , \quad \text{otherwise} \end{cases}$$

This distribution contains two independent constants namely n and p (or q). They are called parameters of the binomial distribution. Sometimes, n is also known as the degrees of the distribution.

Mean:

The binomial probability distribution is given by

$$p(r) = {}^nC_r \; p^r \; q^{n-r}, \; r = 0, 1, 2, \cdots, n \text{ and } q = 1 - p$$

Mean of x,

$$\mu = E(X) = \sum_{r=0}^{n} r \; p(r)$$

$$\mu = E(X) = n\ p$$

$\therefore$ Arithmetic mean of the binomial distribution = n p

Variance:

Variance $v(x) = E(x^2) = [E(x)]^2$

$$= \sum_{r=0}^{n} r^2\ p(r) - \mu^2 = n\ p\ q$$

$$= n\ p\ q$$

$\therefore$ variance of the binomial distribution = n p q

Example 5:

It has been claimed that in 60% of all solar heat installations the utility bill is reduced by at least one-third. Accroding what are the probabilities that the utility bill will be reduced by at least one - third in four of five installations.

Solution:

P = probability that in the solar heat installations the utility bill is reduced by

one-third = 60% = 0.6

$$q = 1 - p = 1 - 0 \cdot 6 = 0 \cdot 4$$

r = 4, n = 5.

$$P(4) = {}^5C_4\ (0 \cdot 6)^4\ (0 \cdot 4)^{5-4} = 5(0 \cdot 6)^4\ (0 \cdot 4) = 0 \cdot 2592$$

Example:

If 3 of 20 tyres are defective and 4 of them are randamly closen for inspection, what is the probability that only one of the defective tyre will be included?

Solution:

Let P = probability of a defective tyres $= \frac{3}{20}$

Then q = probability of non defective tyre $= 1 - p = 1 - \frac{3}{20} = \frac{17}{20}$

Given n = 4

The probability that exactly one tyre will be defective

$$= p(r=1) = {}^4C_1 \left(\frac{3}{20}\right)^1 \left(\frac{17}{20}\right)^{4-1} = \frac{4 \times 3 \times 17^3}{(20)^4}$$

$$= 0 \cdot 3685$$

Example 6:

A fair coin is tossed six times. Find the porbability of getting four heads.

Solution:

p = probability of getting a head $= \frac{1}{2}$

q = probability of not getting head $= \frac{1}{2}$

and n = 6, r = 4.

We know that

$$p(4) = {}^6C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{6-4}$$

$$= \frac{6!}{4!\,2!}\left(\frac{1}{2}\right)^6$$

$$= \frac{6 \times 5}{2} \cdot \frac{1}{2^6}$$

$$P(4) = \frac{15}{64}$$

10.5.3 Poisson Distribution:

Poisson distribution due to french mathematician simeon denis poisson (1837) is a discrete probability distribution.

Definition: A random variable X is said to follow a poisson distribution if it assumes only non-negative valug and its probability mass function is given by

$$p(x, y) = P(X = x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & , \quad x = 0, 1, 2, \cdots \\ 0 & , \quad \text{otherwise} \end{cases}$$

Here $\lambda > 0$ is called the parameter of the distribution

Mean:

Mean of the poisson distribution

$$\text{Mean} = E(X) = \sum_{x=0}^{\infty} x \cdot p(x)$$

$$= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\,\lambda^{x}}{(x-1)!} = \lambda$$

$\therefore$ The parameter λ is the arithmetic mean of the poisson distribution.

Variance:

Variance of poisson distribution

$$V(X) = E\left(X^2\right) - [E(X)]^2$$

Skewness $\beta_1 = \frac{1}{\lambda} = \sum_{x=0}^{\infty} x^2\, p(x) - \lambda^2 = \lambda$

$\therefore$ Variance of the distribution = Mean of distribution = λ

$\therefore$ Standard deviation of poisson distribution $\sigma = \sqrt{\lambda}$

Example 7:

A manufacaturer knows that the censensers he makes contain on average 1% defectives. He packs them in boxes of 100. What is the probability that a box picket at random will contain 3 or more facility cendensers?

Solution:

P = probability of defective condensers = 1% = 0.01

n = Total number of condensers = 100

λ = mean = n p = 100 (0.01) = 1

From poisson distribution

$$P(X = x) = p(x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!} = \frac{e^{-1}\cdot 1^{x}}{x!} = \frac{e^{-1}}{x!}$$

$$p(x \geq 3) = 1 - p(x < 3) = 1 - \left[p(x=0) + p(x=1) + p(x=2)\right]$$

$$= 1 - \left[e^{-1} + e^{-1} + \frac{e^{-1}}{2}\right] = 1 - e^{-1} \times \frac{5}{2}$$

$$= 1 - 0\cdot 9197 = 0\cdot 0803$$

Example 8:

If a bank received on the average $\lambda = 6$ bad cheques per day what are the probabilities that it will receive 4 bad cheques on any given day.

Solution:

We have $P(X = x) = \dfrac{e^{-\lambda}\ \lambda^{x}}{x!}$

Here $\lambda = 6 \Rightarrow P(x = 4) = \dfrac{e^{-6}\ 6^{4}}{4!} = \dfrac{54}{e^{6}} = 0 \cdot 1339$

10.5.4 Pascal Distribution:

Suppose we are intersted in finding the p.m.f. of the number of trails (n) required to get 5 successes given the probability p, of success in any trail.

If n trails are required to get 5 successes then the last trail has to result in a success, while in the rest of the n - 1 trials, 4 successes have been obtained.

The p.m.f. of pascal distribution is

$$f(n) = \binom{n-1}{4}\ p^{4}\ q^{n-5} \cdot p$$

It is customary to write f(n) as $f\left(n/_{r,\,p}\right)$ as r and p are given here.

Mean: Mean of Pascal distribution is $\dfrac{r}{p}$

Variance: Variance of Pascal distribution is $\dfrac{rq}{p^{2}}$

Example 9:

The serews produced by a certain machine were checked by examining the samples of 12. The following table shows the distribution of 121 samples according to the number of defective items they contained.

Number of defectives in a sample of 12	Number of samples
0	7
1	6
2	19
3	35
4	30
5	23
6	7
7	1
Total	**128**

Fit a Binomial distribution and find the expected frequencies if the chance of a serew being defective is 1/2. Find the mean and variance of the fitted distribution.

The probability of a defective screw = 1/2.

$p = 1/2$

$q = 1/2$

n = 7

N = 128

Computation of probability and Frequencies

Number of defectives	Probabilities $P(r)\ {}^7C_r (1/2)^r (1/2)^{7-r}$	Expected Frequency $f(r) = 128 \times P(r)$	Observed Frequency
0	$P(0) = {}^7C_0 (1/2)^0 (1/2)^{7-0}$ $= 1/128$	1	7
1	$P(1) = {}^7C_1 (1/2)^1 (1/2)^6$ $= 7/128$	7	6
2	$P(2) = {}^7C_2 (1/2)^2 (1/2)^5$ $= 21/128$	21	19
3	$P(3) = {}^7C_3 (1/2)^3 (1/2)^4$ $= 35/128$	35	35
4	$P(4) = {}^7C_4 (1/2)^4 (1/2)^3$ $= 35/128$	35	30
5	$P(5) = {}^7C_6 (1/2)^6 (1/2)^2$ $= 21/128$	21	23
6	$P(6) = {}^7C_6 (1/2)^6 (1/2)^1$ $= 7/128$	7	7
7	$P(7) = {}^7C_7 (1/2)^7 (1/2)^0$ $= 35/128$	1	1

Mean = $np = 7 \times 1/2 = 3 \cdot 5$

Standard Deviation $= \sqrt{npq} = \sqrt{7 \times 1/2 \times 12}$

$= 1 \cdot 32$

Example 10:

Fit a poission's distribution to the given set of observations:

Death	0	1	2	3	4
Frequency	122	60	15	2	1

X	f	fX
0	122	0
1	60	60
2	15	30
3	2	6
4	1	4
	200	100

$$\lambda = \frac{\sum f}{N} = \frac{100}{200} = 0 \cdot 5$$

The probability of 0 deaths $P(0) = e^{\lambda} \frac{\lambda^0}{0!}$

From tables $e^5 = 0 \cdot 6065$, where $\frac{\lambda^0}{0!}$ is equal to 1

The expected number of 0 deaths in 200 cases

$= 0 \cdot 6065 \times 200 = 121 \cdot 30$

$P(1) = e^{\lambda} \frac{\lambda^1}{1!}$ or $P(0) \times \lambda = 121 \cdot 30 \times 0 \cdot 5 = 60 \cdot 65$

$P(2) = e^{\lambda} \frac{\lambda^2}{2!}$ or $P(1) \times \frac{\lambda}{2} = 60 \cdot 65 \times \frac{0 \cdot 5}{2} = 15 \cdot 16$

$$P(3) = e^{\lambda}\ \frac{\lambda^3}{3!} \text{ or } P(2) \times \frac{\lambda}{3} = 15 \cdot 16 \times \frac{0 \cdot 5}{3} = 2 \cdot 52$$

$$P(4) = e^{\lambda}\ \frac{\lambda^4}{4!} \text{ or } P(3) \times \frac{\lambda}{4} = 2 \cdot 52 \times \frac{0 \cdot 5}{4} = 0 \cdot 94$$

Expected Frequency Distribution

X	0	1	2	3	4	Total
f	121	61	15	2	1	200

Example 11:

One fifth percent of the blades produced by a blade manufacturing factory turns out to be defective. The blades are supplied in packets of 10. Use poission distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively, in the consignment of 1,00,000 packets $\left(e^{0 \cdot 02} = 0 \cdot 982\right)$

$P = 1/500\ ,\ n = 10$

$\lambda = n\,p = 1/500 \times 10 = 0 \cdot 02$

Probability of r defetive blades

$$= P(r) = e^{-0 \cdot 02} \times \left(\frac{0 \cdot 02}{r!}\right) \times N$$

Number of packets with no defective baldes

$= 1,00,000 \times e^{-0 \cdot 02}$

$= 10,000 \times 0 \cdot 9802$

$= 98 \cdot 020$

Number of packets with defective blades

$$P(1) = e^{-0 \cdot 02} \left(\frac{0 \cdot 02}{r!}\right)$$

$= 1,00,000 \times e^{-0 \cdot 02} \times 0 \cdot 02$

$= 91,020 \times 0 \cdot 02$

$= 1960 \cdot 40$

Number of packets with two defective blades

$$P(2) = 10,000 \times e^{-0\cdot02 \times \left(\frac{0\cdot02}{r!}\right)}$$

$$= 98,020 \times 0\cdot0002$$

$$= 19\cdot60$$

10.6 Summary:

We have introduced the base concept random variable and probability distribution. The discrete probability distributions, Binomial, and poisson distribution are given with examples.

10.7 Exercises:

1. Explain the concepts of random variable and probability distribution with examples.
2. What is a theoretical distribution, and why is it necessary to study such distributions?
3. The probability distribution of daily demand for a particular product is as follows:

Demand	50	75	100	115	125
Probability	0.26	0.16	0.37	0.16	0.05

Find the expected value of daily demand for the product.

4. Mr. Chand is planning to invest in Stock 1 and Stock 2. The rate of return on a single share of Stock 1 will be 20%, 24% or 28% with probabilities 0.30, 0.40, 0.30 respectively. On the other hand, the rate of return on a single share of Stock 2 will be 10%, 20%, 25%, 28% or 35% with respective probabilities of 0.15, 0.25, 0.30, 0.25 and 0.05.

   Which is the best investment option for Mr. Chand?

5. A consultant studying the safety of road traffic found from past records that accidents occur at an average rate of 5 per week.

   (i) What is the expected number of accidents per week?

   (ii) What is the probability of two or less number of accidents during the next month?

   (iii) What is the probbaility of at least 3 accidents during the next month?

6. 10% of screws produced in a factory turn out to be defective. Find the probability that in a sample of 10 screws chosen at random, exactly two will be defective.
7. If 10% of blots produced by a machine are defective, find the probability that out of 10 bolts chosen at random (i) none will be defective (ii) one will be defective, and (iii) at most two bolts will be defective.
8. Eight fair coins are thrown simultaneously. Find the probability of getting (i) exactly six heads (ii) at most six heads.

9. A box contains 100 transistors, 20 of which are defective and 10 are selected at random, find the probability that

   (i) All are defective

   (ii) At least one is defective

   (iii) All are good

   (iv) At most 3 are defective

10. In acertain city only 50% of the students are capable of doing college work actually go to college. Assuming that this claim is true, find the probability that among 18 such a capable students.

    (i) Exactly 10 will go to college

    (ii) At least 2 will go to college

    (iii) At most 17 will go to college

11. Fit a binomial distribution to the following data.

x	0	1	2	3	4	5
f	38	144	342	287	164	25

12. Fit a binomial distribution for the following:

x	0	1	2	3	4	5	6	7	8	9	10
f	6	20	28	12	8	6	0	0	0	0	0

13. 7 Coins are tossed at a time, 128 times. The number of heads observed at each throw is recorded and the results are given below:

No.of Heads	0	1	2	3	4	5	6	7
Frequency	7	6	19	35	30	23	7	1

    Fit a binomial distribution to the data assuming that the coins are (i) biased (ii) unbiased?

14. A manufacturer known that the razor blades he makes contain on a average 0.5% of defectives. He packs them in packets of 5. What is the probability that a packet picked at random will contain 3 or more faulty blades.

15. Suppose 3% of bolts made by a machine are defective the defects occurring at random during production. If bolts are packed 50 per box, find the probbaility that a given box will contain 5 defectives.

16. In a city, 6% of all drivers get at least one parking ticket per year. Determine the probabilities that among 80 drivers

    (i) 4 drivers will get at least one parking ticket in a year.

    (ii) At least 3 drivers will get at least one parking ticket in a year

17. An insurance company found that only 0.01% of the population is involved in a certain type of accident each year. If its 1000 policy holders were randomly selected from the population, what is the probability that not more than two of its clients are involved in such an accident next year?

18. Fit a poisson distribution for the following data and calculate the expected frequencies.

x	0	1	2	3	4	5	6	7	8
f	56	156	132	92	37	22	4	0	1

19. The distribution of typing mistakes committed by a typist is given below assuming the distribution to be poisson, find the expected frequencies.

x	0	1	2	3	4	5
f(x)	125	95	49	20	8	3

20. In 1000 sets of trials for an event of small probability the frequencies f of the numbers of x of successes are

x	0	1	2	3	4	5	6	7
f	305	365	210	80	28	9	2	1

10.8 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: K. CHANDAN

Lesson - 11

CONTINUOUS PROBABILITY DISTRIBUTION

Objectives:

- To study the characteristics of continous probability distribution.
- To analyze situations using normal distribution.
- To discuss the usefulness of standard normal tables.
- To discuss the normal approximation to binomial distribution.

Structure:

9.1 Introduction:

A random variable is said to be continuous if it can take any value over a range. The probability distribution associated with a continous random variable is termed as continuous probability distribution.

We consider some univariate continuous distributions in this chapter. The main continuous destributions like uniform distribution, normal distribution, gamma, beta, exponential, laplace, weilbul, logistic and cauchy distributions will discussed in detail in the subsequent sections.

11.2 Basic Concepts:

A random variable is said to be continuous. If it can take any value over a range. The probability distribution associated with a continuous random variable is termed as continous probability distribution.

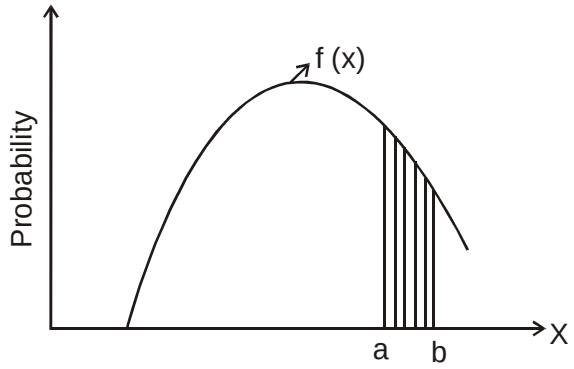


Fig - 1

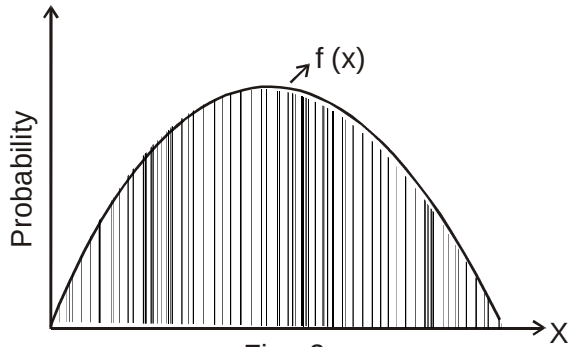


Fig - 2

We can see the graphical representation of discrete probability distributions (see fig 1 and 2). Since a discrate random variable cannot take on any value between two integers, we notice that the graph of the discrete probability distribution consists of vertical lines where the length of each vertical line represents the probability. On the otherhand, a continuous random variable can take on any value over a range and hence the graph of a continuous probability distribution will be a cure. The area between the X - axis and curve represents the probability, the total area being equal to 1.

Suppose "X" is a continuous random variable and $f(X)$ is the associated probability function. To find the probbaility that "X" takes on a value between two real numbers a and b. We have to use the concept of definite integral by integroting $f(X)$ between the values a and b, we get the required probbaility denoted by the area (see fig 1) between the X - axis and the curve representing $f(X)$.

If X is a random variable which can take on any value between $-\infty$ and $+\infty$, we get the total area which is equal to 1 (see fig - 2).

1.2.1 Random Variable: The concept of the random experiment leads to the notation of a sample space. The assignment and computaion of probabilities of events were studied in detail. In many experiments, we are interested not in knowing which of the outcomes has occured, but in the numbers associated with them.

For example, when n coins are tossed, one may be interested in knowing the number of heads. obtained when a pair of dice are tossed one may seek information about the sum of points.

Thus, we associated a real number with each out come of an experiment. In other words, we are considering a function whose domain is the set of possible outcomes, and whose range is a subset of the set of reals. Such a function is called a random variables.

Example 1:

We mean a real number X connected with the out come of a randome experiment E.

If E consists of two tosses the random variable which is the number of heads (0, 1 or 2).

Out Come:	H H	H T	T H	T T
Value of X:	2	1	1	0

11.2.2 Types of random variables:

Random variables are two types. They are

1. Discret random variable
2. Continuous random variable

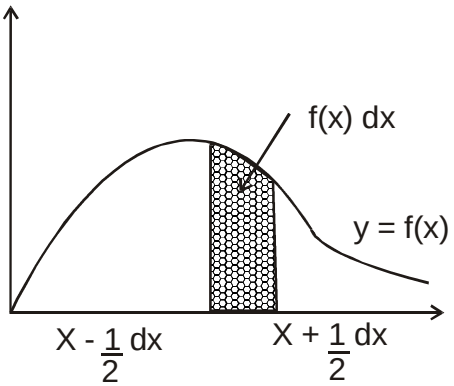
1. Continuous random variable: A random variable X is said to be continuous if it can take all possible values between certain limits. In other words, a random variable is said to be continuous when its different values cannot be put in 1 -1 correspondence with a set of positive integers.

A continuous random variable is a random variable that can be measured to any desired degree of accuracy.

Examples: Age, height, weight, etc.....

2. Probability Density Function: Consider the small interval (x, x + dx) of length dx round the point x. Let f(x) be any continuous function of x so that f(x) dx represents the probability that X falls in the infinitesimal interval (x, x + dx). Symbolically

$$P\left(x \le x \le x + dx\right) = fx\,(x)\,dx$$



In the figure $f(x)\,dx$ represents the area bounded by the curve $y = f(x)$, x - axis and ordinotes at the points x and $x + dx$. The function $f_x(x)$ so defined is known as probbaility density function or simply density function of random variable X and is usually obbreviated as p.d.f. the expression, $f(x)\,dx$, usually written as $d\,F(x)$, is known as the probbaility differential and the curve $y = f(x)$ is known as the probability density curve or simply probability curve.

The p.d.f. $f_x(x)$ of the r.v. "X" is defined as

$$f_x(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x + \delta x)}{\delta x}$$

The probability for a variate value to lie in the interval dx is $f(x)\,dx$ and hence the probability for a variate value to fall in the finite interval $[\alpha, \beta]$ is

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx$$

Which represents area between the curve $y = f(x), X-$ axis and the ordinates at $x = \alpha$ and $x = \beta$.

The probability density function of a r.v. "X", usually denoted by $f_x(x)$ or simply by $f(x)$ has the following obvious properties.

(i) $f(x) \ge 0$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(iii) The probability P(E) given by $P(E) = \int_{\in} f(x)\,dx$ is well defined for any event "$\in$".

11.3 Some Important Continuous Distributions:

1. Continuous Uniform or Rectangular Distribution:

Def: A random variable X is said to follow continuous uniform or rectangular distribution in an interval (a, b) if its density function is constant over the entire range of the variable X. Its functional form is,

$$f(x) = \begin{cases} 1/(b-a) & \text{if} \quad a \le x \le b \\ 0 & , \quad \text{otherwise} \end{cases}$$

It is denoted as $U(a, b)$. If $X \sim U(0, 1)$, $f(x) = 1$.

The distribution function of the variable $X \sim U(a, b)$ is

$$F(x) = \begin{cases} 0 & \text{if} \quad -\infty < x < a \\ \dfrac{x-a}{b-a} & \text{if} \quad a \le x \le b \\ 1 & \text{if} \quad b < x < \infty \end{cases}$$

Properties of rectangular distribution:

1. Its mean is equal to $\dfrac{b+a}{2}$
2. All the moments of odd order are zero
3. Its variance, $\mu_2 = \dfrac{(b-a)^2}{12}$
4. Its fourth central moment, $\mu_4 = \dfrac{(b-a)^4}{80}$
5. Median of rectangular distribution is $\dfrac{b+a}{2}$
6. Mean deviation about mean, $M \cdot D_{\bar{x}} = \dfrac{b-a}{4}$
7. Measure of skewness $\beta_1 = 0$ or $\gamma_1 = 0$
8. Measure of Kurtosis $\beta_2 = \dfrac{9}{5}$ or $\gamma_2 = \beta_2 - 3 = \dfrac{-6}{5}$

Pearsons coefficients β_1 and β_2 reveal that the rectangular distribution is symmetrical and platykurtic.

9. Moment generating function of rectangular distribution,

$$M_x(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

10. Characteristic function of rectangular distribution.

$$\phi_x(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$$

11. Mode does not exists as the probability at each point in the interval (a, b) remains the same.

12. If X and Y are independently and identically distributed rectangular or uniform $U(0,1)$ the distribution of the variates $(x+y), (x-y), xy$ and x/y are as follows.

$$f(x+y) = \begin{cases} x+y & , \quad 0 \le x+y \le 1 \\ 2-(x+y) & , \quad 1 \le x+y \le 2 \end{cases}$$

$$f(x-y) = \begin{cases} x-y+1 & , \quad -1 \le x-y \le 0 \\ 1-(x-y) & , \quad 0 \le x-y \le 1 \end{cases}$$

$$f(xy) = -\log(xy), \quad 0 < xy < 1$$

$$f(x/y) = \begin{cases} 1/2 & , \quad 0 \le x/y < 1 \\ y^2/2x^2 & , \quad 1 < x/y < \infty \end{cases}$$

13. If a variable $X \sim U(0,1)$ then the varibale $Y = -2 \log X$ is distributed as χ^2 with 2 d f.

14. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with distribution $U(0,1)$ then the variable $Y = -2 \sum_{i=1}^{n} \log x_i$ or $Y = 2 \log\left(1 \Big/ \prod_{i=1}^{n} X_i\right)$ is distributed as χ^2 with 2 n d.f.

2. Exponential Distribution:

Def: A continuous random variable "X" is said to follow exponential distribution if for any positive value λ, it has the probability density function,

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad \lambda > 0, x > 0$$

This is usually denoted as $E \times Po(\lambda)$. λ is know as the parameter of exponential distribution.

The distribution function of exponential variate "X" is

$$F(X) = \begin{cases} 1 - e^{-\lambda x} & , \quad \text{if } x > 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Properties of exponential distribution:

The well known properties of exponential distribution are:

1. Mean of exponential distribution is $1/\lambda$
2. Variance of exponential distribution is $1/\lambda^2$
3. Moments of all order exist. The first four central moments are:

$$\mu_1 = \frac{1}{\lambda}, \mu_2 = \frac{1}{\lambda^2}, \mu_3 = \frac{2}{\lambda^3}, \mu_4 = \frac{9}{\lambda^4}$$

If $\lambda > 1$, Mean > Variance. If $\lambda < 1$, Mean < Variance, if $\lambda = 1$, Mean = Variance.

4. The relationship between central moments and cumulants are:

$\mu_1 = k_1$, $\mu_2 = k_2$, $\mu_3 = k_3$ and $\mu_4 = k_1 + k_2^2$.

5. Pearson's measure of skewness $\beta_1 = 4$ or $\gamma_1 = \sqrt{\beta_1} = 2$
6. Pearson's measure of Kurtosic $\beta_2 = 9$ or $\gamma_2 = \beta_2 - 3 = 6$

The values of β_1 or γ_1 and β_2 and γ_2 clearly reveal that exponential distribution is positively skewed and is leptokurtic.

7. The median of exponential distribution is $1/\lambda$
8. Moment generating function

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

9. Characteristic function, $\phi_x(t) = \left(1 - \frac{it}{\lambda}\right)^{-1}$
10. It also possesses the memorylese property just like geometric distribution.

Example 2:

A highway petrol pump can surve on an average 15 cars per hour. What is the probbaility that for a particular car, the time taken will be less than 4 minutes?

Solution:

Exponnential distribution with $\lambda = 15$ (sensis rate) we are interested in finding the probbaility that t < 4 minutes $< \frac{4}{60}$ hrs.

From definition of c.d.f we went to find $F\left(\frac{4}{60}\right) = F\left(\frac{1}{75}\right)$ we have seen that $F(t) = 1 - e^{\lambda t}$

$$F\left(1/75\right) = 1 - e^{-15 < 1/25} = 1 - e^{-1} = 1 - 0 \cdot 368 = 0 \cdot 632$$

Example 3:

The distribution if the total time a light bulb will burn from the moment it is first put into service is known to be exponential with mean time between failure of the bulbs equal to 2000 hrs. What is the probability that a bulb will burn more than 2000 hrs.

Solution:

$$\lambda = \frac{1}{2000}, \; f(t) = \begin{cases} \frac{1}{2000} \cdot e^{-t/2000} & , \quad \text{for } t > 10 \\ 0 & , \quad \text{otherwise} \end{cases}$$

We are interested in finding the probability that t > 2000 hrs.

$$P(t > 2000) = 1 - P(t \le 2000) = 1 - F(2000)$$

$$F(2000) = 1 - e^{-2000/2000} = 1 - e^{-1}$$

$\therefore$ The required probability $= e^{-1} = 0 \cdot 368$

3. Normal Distribution:

Normal distribution was first discovered by De-Moinre in 1733 and was also known to Laplace in 1774. Latter it was derived by Kark Friedrich Gauss in 1809 and used it for the study of errors in astronomy.

A random variable X is said to follow a normal distribution with mean μ and variance σ^2, if its probability density function is

$$f_x\left(x : \mu,\ \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad \text{for } -\infty < x\ \infty,\ -\infty < \mu < \infty \text{ and } \sigma > 0$$

the variate X is said to be distributed normally with mean μ and variance σ^2 and is denoted as $X \sim N\left(\mu, \sigma^2\right)$.

If $\mu = 0$ and $\sigma = 1$ then $f_x(x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x^2}$

Here, X is said standard normal variate and is denoted as $X \sim N(0, 1)$. Also p.d.f. $f_x(x)$ is called standard normal distribution. If $X \sim N\left(\mu, \sigma^2\right)$ and we make a transformation $Z = \frac{X-\mu}{\sigma}$ the distribution as $Z \sim N(0, 1)$ irrespective of the values of μ and σ^2 in the distribution of X, Z is also called standard normal. We can write $f(Z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}Z^2}$.

Also $\int_{-\infty}^{\infty} f(Z)\, dZ = 1$ the area under the standard normal curve between the ordinates at $Z = 0$ and $Z_1 = \phi$ (Z) is given as,

Also $\phi(Z) = \int_{0}^{Z_1} f(Z)\, dZ$

$$\phi(Z) = \phi(-Z)$$

Characteristics of the normal distribution:

1. The normal distribution curve is bell - shaped and is symmetrical about the line $x = \mu$.
2. The mode of the normal curve lies at the point $x = \mu$
3. The area under the normal curve within its range $-\infty$ to ∞ is always unity, i.e.,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\ dx = 1$$

4. Mean = median = Mode

5. The normal curve is unimodal

6. All odd order moments of the normal distribution are zero.

7. The first rawmoment i.e., Mean = μ. Also $\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 3$, $\sigma^4 = 3\mu_2^2$.

8. The measure of skewness $\beta_1 = \frac{\mu_3^2}{\mu_2^2} = 0$ or $\gamma_1 = \sqrt{\beta_1} = 0$

9. Measure of kurtosis is $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$ or $\gamma_2 = \beta_2 - 3 = 0$

The value of β_1 and β_2 clearly reveal that the normal curve is symmetrical and mesokurtic.

10. Quartile deviation $Q \cdot D \cdot = Q_3 - Q_1/2 = \frac{2}{3}\sigma$

The importance and applications of the normal distribution:

Normal distribution plays a very important role in statistical theory because of the following reasons:

1. Most of the distributions that are encountered in proctice, for example, Binomial, Poisson, Hypergemotric etc. can be approximated by normal distribution.

2. Since the normal distribution is a limitting case of the binomial distribution for exceptionally large numbers it is applicable to many applied problems in kinetic theory of gases and fluctuations inthe magnitude of an electric current.

3. Event if a variable is normally distributed, it can some times be brought to normal from by simple transformation of the variable.

4. The proofs of all the tests of significance in sampling are based upon the fundamental assumption that the population from which the samples have been drawn is normal.

5. Normal distribution finds large applications in statistical quality control.

Example 4:

Suppose the weights of 800 Male students are normally distributed with mean $\mu = 140$ pounds and standard deviation 10 pounds. find the number of students whose weights are

(i) between 138 and 148 pounds

(ii) Morethan 152 pounds

Solution:

Let μ be the mean and σ be the standared deviation

Then $\mu = 140$ founds and $\sigma = 10$ pounds.

i) When $x = 138,\ Z = \dfrac{x - \mu}{\sigma} = \dfrac{138 - 140}{10} = -0 \cdot 2 = Z_1\ (\text{say})$

When $x = 148,\ Z = \dfrac{x - \mu}{\sigma} = \dfrac{148 - 140}{10} = 0 \cdot 8 = Z_2\ (\text{say})$

$\therefore P(138 \le x \le 148) = P(-0 \cdot 2 \le Z \le 0 \cdot 8)$

$= A(Z_2) + A(Z_1) = A(0 \cdot 8) + A(-0 \cdot 2)$

$= A(0 \cdot 8) + A(0 \cdot 2)$

$= 0 \cdot 2881 + 0 \cdot 0793$

$= 0 \cdot 3674$

Hence the number of students whose weights between 138 pounds and 148 pounds = 0.3674 x 800 = 294

ii) When x = 152, $Z = \dfrac{x - \mu}{\sigma} = \dfrac{152 - 140}{10} = 1 \cdot 2 = Z_1\ (\text{say})$

$\therefore P(x > 152) = P(Z > Z_1)$

$= 0 \cdot 5 - A(Z_1) = 0 \cdot 5 - A(1 \cdot 2)$

$= 0 \cdot 5 - 0 \cdot 3849 = 0 \cdot 1151$

$\therefore$ Number of students whose weights are more than 152 pounds

$= 800 \times 0 \cdot 1151 = 92$

Example 5:

A sales tax officer has reported that the average sales of the 500 business that he has to deal with during a year is Rs. 36,000 with a standard deviation of 10,000. Assuming that the sales in these business are normally distributed. find

(i) The number of business as the sales of which are Rs. 40,000

(ii) The percentage of business the sales of which are likely to range between Rs. 30,000 and Rs. 40,000

Solution:

Let μ be the mean and σ the standard deviation of the sales. Then we are given that $\mu = 36000$ and $\sigma = 10000$

Let the variable X denote the sales inthe business

when $X = 40000,\ Z = \dfrac{X - \mu}{\sigma} = \dfrac{40000 - 36000}{10000} = 0 \cdot 4$

when $X = 30000,\ Z = \dfrac{X - \mu}{\sigma} = \dfrac{30000 - 36000}{10000} = -0 \cdot 6$

(i) $P(X > 40000) = P(Z > 0 \cdot 4)$

$= 0 \cdot 5 - A(0 \cdot 4) = 0 \cdot 5 - 0 \cdot 1554 = 0 \cdot 3446$

Number of business as the sales of which are Rs. 40,000

$= 500 \times 0 \cdot 3446 = 172 \ (\text{approximately})$

(ii) $P(30,000 < X < 40,000) = P(-0 \cdot 6 < Z < 0 \cdot 4)$

$= A(0 \cdot 4) + A(-0 \cdot 6)$

$= A(0 \cdot 4) + A(0 \cdot 6)$

$= 0 \cdot 1554 + 0 \cdot 2257$

= 0.3811

$\therefore$ The requrired percentage of business = 38.11%

11.4 Normal Distribution Approximation to Binomial Distribution:

Another important application of normal distribution is that is approximates bionomial probabilities for a large number of Bernoulli trails. We know that when X follows binomial distribution with parameters n and p then,

Mean = n p and variance = n p (1 - p).

Then we have the following results

If X is a binomial random variable with parameters n and p then,

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches standard normal distribution for large n.

Normal approximation to binomial distribution is usually good when np and $np(1-p)$ are both equal to or greater than 5.

We may note that we are approximating a discrete distribution with a continuous distribution. There fore we use what is called a continuity correction factor to convert the discrete variable into contionuous variable. We subtract (or add) 0.5 to the value of binomial variable to find the required probability using normal tables.

Suppose X is a binomial variable with mean np and variance $np(1-p)$, then to find $p(x \le a)$ we proceed as follows:

1) Standardize X

i.e., $$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Then

$$p(x \le a) = p\left[Z \le \frac{a - np}{\sqrt{np(1-p)}}\right]$$

2. Using continuity correction factor, we write

$$P(x \le a) = P\left[Z \le \frac{(a+0\cdot5)-nP}{\sqrt{nP(1-p)}}\right]$$

which can be eaily obtained from tables.

Similarly, we can find $P(x \ge b)$ as

$$P(X \ge b) = P\left[Z \ge \frac{(b-0\cdot5)-nP}{\sqrt{nP(1-P)}}\right]$$

and $$P(x=k) = P\left[\frac{(K-0\cdot5)-nP}{\sqrt{nP(1-P)}} \le Z \le \frac{(K+0\cdot5)-nP}{\sqrt{nP(1-P)}}\right]$$

Example 6:

Find the probability that by Guess - Work a student can correctly answer 25 to 30 questions in a multiple - choic quiz consisting of 80 questions. Assume that in each question of 80 questions. Assume that in each question with four choices, only one choice is correct and student has no knowledge of the subject.

Solution:

Here $P = \frac{1}{4}$ so $q = 1 - P = 1 - \frac{1}{4} = \frac{3}{4}$

Mean, $\mu = nP = 80\left[\frac{1}{4}\right] = 20$ and $\sigma = \sqrt{npq} = \sqrt{20 \times \frac{3}{4}} = \sqrt{15}$

$x_1 = 25$ and $x_2 = 30$

Hence $Z_1 = \frac{\left(x_1 - \frac{1}{2}\right)\mu}{\sigma} = \frac{\left[25 - \frac{1}{2}\right] - 20}{3 \cdot 87} = \frac{4 \cdot 5}{3 \cdot 87} = 1 \cdot 16$

and $Z_2 = \frac{\left(x_2 + \frac{1}{2}\right) - \mu}{\sigma} = \frac{\left(30 + \frac{1}{2}\right) - 20}{3 \cdot 87} = \frac{10 \cdot 5}{3 \cdot 87} = 2 \cdot 71$

Hence the requeired probability

$= P\left(25 \le X \le 30\right)$

$= P\left(1 \cdot 16 \le Z \le 2 \cdot 71\right) = \left| A\left(2 \cdot 71\right) - A\left(1 \cdot 16\right) \right|$

$= 0 \cdot 496 - 0 \cdot 3770$

$= 0 \cdot 1196$

11.5 Summary:

In this lesson we have introduced probability distribution of a continuous random variable. We can understand the basic concepts and assumptions involved in the treatment of continuous probability distributions. Experimential and the normal distribution are explained with examples. This can be useful in decission making.

11.6 Exercises:

1. Briefly explain the characteristics of the extonential distribution and discuss some of its uses.
2. Describe the features of a normal curve.
3. Define the standard normal varibale why is it necessary to standardize a normal variable.
4. The daily consumption of milk in a excess of 20,000 gallans is approximately exponentially distributed with 0 = 3000. The city has a daily stock of 35,000 gallons, what is the probability that of two days selected at random the stock is insufficient for both the days.

5. 200 electric light bulbs were tested and the average life time of the bulbs was found to be 25 hours. Using the summary given below, test the hypothesis that the lifetime is exponentially distributed.

Life time in hours:	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
Number of bulbs :	104	56	24	12	4

You are given that an exponential distribution with parameter $\alpha > 0$ has the probability density function:

$$P(X) = \alpha\, e^{-\alpha x}, \quad (x \geq 0)$$

$$= 0, \quad (x < 0)$$

6. From the records of an insurance company is asserted that:

(i) There is a time log between the reporting of a cliam and its settlement,

(ii) The frequency density function for the number of claims getting settled, out of an initial number N of claims, at a point of time in the vicinity of time t (in years) is

$$N\, ke^{-kt}, \quad \geq 0, \; k > 0$$

(iii) The amount paid out as claim increases with the time lag between the reporing and settlement of a claim, the average amount paid out in respect of claims settled at time t being $e^{ct}, \; c > 0$

(a) Show that the proportion of claims outstanding at time t is e^{-kt}

(b) What is the average amount paid in respect of all the claims?

(c) Show that the average amount paid in respect of the clims settled within the first year of its being reported is...

$$\frac{k}{k-c} \; \frac{e^{k} - e^{c}}{e^{k} - 1}$$

Hoe does this compare with the result of (b)?

7. The mean height of 500 students is 151 cm. and the standard deviation is 15 cm. Assuming that the heights are normally distribted, find how many students heights lie between 120 and 155 cm.

8. The life of electronic tubes of a certain type may be assumed to be normally distributed with mean 155 hours and standard deviation 19 hours. Determine the probbaility that the life of a randomly choosen tube.

(i) is between 136 hours and 174 hours

(ii) Less than 117 hours (iii) Will be more than 395 hourns.

9. In a test on 2000 electric bulbs, it ws found that the life of a particular make, was normally distributed with an average life of 2040 hours and standard deviation of 60 hours. Find the number of bulbs likely to burn for (i) more than 2150 hours (ii) less than 1950 hours and (iii) more than 1920 hours and but les than 2160 hours.

10. In a referendum 60% of voters voted in favour. A random sample of 200 voters was selected. Find the probability by using normal distribution that in the sample

(i) More that 130 voted in favour ?

(ii) Between 105 and 130 inclusive voted in favour?

(iii) 120 voted in favour?

11.7 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics fir Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: K. CHANDAN

Lesson - 12

DECISION THEORY

Objectives:

After reading this lesson you should be able to

- Understanding of Decision Theory in real life problems.
- Marginal analysis can solve the uncertainty problems.
- Decision Tree approach is graphic reprexsentation of the decision process indicating decision alternatives.
- Importance of preference theory.

Structure:

12.1 Introduction:

Until a few years ago scientific aids to business management were used only when dealing with very specific problems such as inventory control. Recently, however, decision theory has been developed and has been shown to have rather wide application to business management problems. It is this wide applicability that has made decision theory so attractive to managers.

With due respect to its complexities, the basis of the managerial process is decision making. Decision theory is merely a description - written in mathematical terms of this aspect of the management process. As such, it is a valuable aid to scientific management. Instead, it simplifies application and understanding.

Decision Procedure: "Decision Theory" describes a process, which results in the selection of the proper managerial action from among well defined alternatives. By implication, "decision making" is the selection of the best alternative.

If decision theory is to be applicable to the managerial process, it must adhere to each of the following elements of decision making:

a) Definition of the the problem

b) Establishment of the appropriate decision criteria

c) Accurate determination of the environmental situation

d) Description of all alternative managerial actions

e) Development of the decision process

f) Solution of the problem

g) Making the decision

Each of the above elements will now be dealt with in more detail.

Definition of the Decision Making Problem:

All decision making problems can be characterized by,

a) The desire to attain an established goal.

b) The availability of many alternative managerial actions.

c) A particular environment, which exists with regard to the alternative actions, e.g. risk, certainty, uncertainty, conflict, ignorance.

The manager must analyze each problem in terms of each of the above characteristics. A solution is always more accesible when the manager has a more through understanding of the problem which he confronts.

Decision Makers:

Decision theory posits that the human decision maker brings to the resolution of a decision problem beliefs and preferences. Specifically, the theory presumes that the decision maker possesses a probability system that captures his or her (partial) beliefs about nature's selection of states, a belief system about the outcomes accruing to the performance of the acts in the vaious states of nature, and a preference structure over the outcomes. Thus, the decision maker is a triple <P, F, U> composed of a probability measure P, an outcome mapping F, and a utility function U.

The probability measure P is defined over the set of states of nature, and captures the decision maker's view of the process where by the states are selected by nature. The outcome mapping F is defined on the cartesian product of the states of nature and the acts, and presents the outcome resulting from performing each act in each of the states of nature. Thus, the outcome mapping F generates a new set O of outcomes. The utility function is defined over the set O of outcomes and represents the decision maker's preferences over the outcomes.

In what follows we will consider both decision making under uncertainty and decision making under risk. We will treat the former by simply ignoring the decision maker's probability system, and we will treat the latter by incorporating the decision maker's probability system. The focus of this course, however, is on decision making under risk.

Decision Making:

Decision making is composed of a two step process. First, the acts in A are ordered and second the "best" act is selected. Typically, the first step is completed by assigning a number to each act, and then using the complete ordering property of the real numbers to order the acts. If the outcomes are "goods", like gains, income, etc., then the best act is the act with the highest number. Contrariwise, if the outcomes are "bads", like losses, costs, etc., then the best act is the act with the smallest number.

Decision Rules:

Decision rules are composed of two commands. The first command tells the user how to assigns numbers to acts; the second command tells the user how to choose among the numbers assigned by the first command. Decision rules are designed to reflect some human attitude toward decision making. The rules considered here reflect two forms of pessimism and general rationality.

Characteristics of the solution:

(a) The problem situation was characterized a s a "risk" environment, and empirical probabilities were used. Just how valid were the data? Do they still apply?

(b) The decision criterion selection was based on a value judgement of long - run profit needs, short-run cash needs, firm objectives, etc.

(c) the decision process was merely a conceptualization of the appropriate relationship between environment and managerial actions. Is such a relationship valid?

(d) For the final decision to be optimum, the list of alternative managerial actions must be all inclusive. Was it?

If the manager can live with each of the above, then one can safely say that statistical decision theory proved to be of vital importance in determining the "best" answer.

Decision Making Environments:

Decisions are made under three types of environments:

1. **Decision making under conditions of certainty:** In this environment, only one state of nature exists i.e., there is complete certainty about the future. It is easy to analyse the situation and make good decisions.
2. **Decision making under conditions of uncertainty:** Here, more than one states of nature exist but the decision maker lacks sufficient knowledge to allow him assign probabilities to the various states of nature.
3. **Decision making under conditions of risk:** Here also, more than one states of nature exist but the decision maker has sufficient information to allow him assign probabilities to each of these states.

12.2 Decision Making Under Conditions of Certainty:

Since under this environment, only one state of nature exists, the decision maker simply picks up the best payoff in that one. Under conditions of certainty, the particular state of nature is

associated with probability 1. Through the state of nature is only, possible alternatives could be numerous.

Under conditions of uncertainty, the decision maker has a knowledge about the states of nature that happen but lacks the knowledge about the probabilities of their occurrence. Situations like lounching a new product fall under this category. The insufficient data lead to a more complex decision model and perhaps, a less satisfactory solution. How ever, one uses scientific methods to exploit the available data to the fullest extent.

In decision under uncertainty, situations exist in which two (or more) apponnetns with conflicting objectives try to make decisions with each trying to gain at the cost of the other(s). These situations are different since the decision maker is working against an intelligent apponent. The theory governing these types of decision problems is called the theory of games.

Maximax Criterin:

This criterion provides the decision maker with optimistic criterion. He finds the maximum possible pay off for each possible alternative and then chooses the alternative with maximum pay off with in this group. Table 12.1 to illustrate this method. The maximax payoff is Rs. 70,000 corresponding to the alternative construct.

Table 12.1

Alternatives	States of nature (product demand)				Maximum of row
	High (Rs.)	Moderate (Rs)	Low (Rs)	Nil (Rs)	Rs.
Expand	50,000	25,000	- 25,000	- 45,000	50,000
Construct	70,000	30,000	- 40,000	- 80,000	70,000 (Maximax)
Sub contract	30,000	15,000	- 1,000	- 10,000	30000

Maximin Criterion:

This criterion provides thedecisionmaker with pessimistic criterion. To use this criterion, the decision maker maximizes his minimum possible pay offs. He finds first the minimum possible pay off for each laternative and then chooses the alternative with maximum payoff within this group. the maximum payoff to the company as obtained from Table 12.2 is Rs. 10,000 corresponding to the alternative 'sub contract'.

Table 12.2

Alternatives	States of nature (product demand)				Minimum of row
	High (Rs.)	Moderate (Rs)	Low (Rs)	Nil (Rs)	Rs.
Expand	50,000	25,000	- 25,000	- 45,000	- 45,000
Construct	70,000	30,000	- 40,000	- 80,000	- 80,000
Sub contract	30,000	15,000	- 1,000	- 10,000	- 10,000 (Maximin)

Minimax Regret Criterion:

This decision criterion ws developed by L.J. Savage. He pointed out that the decision maker might experience regret after the decision has been made and the states of nature i.e., events have occurred. Thus the decision maker should attempt to minimize regret before actually selecting a particular alternative (strategy).

Table 12.3

Alternatives	States of nature (product demand)				Maximum of row
	High (Rs.)	Moderate (Rs)	Low (Rs)	Nil (Rs)	Rs.
Expand	20,000	5,000	24,000	35,000	35,000 (Minimax)
Construct	0	0	39,000	70,000	70,000
Sub contract	40,000	15,000	0	0	40,000

Amount of regrets are represented in table 12.3. This table shows that the company will minimize its regret to Rs. 35,000 by selecting alternative 'Expand'. It may be observed that which the other decision rules do not take into account the cost of opportunity last by making the wrong decision, the minimax regret criterion does so.

Criterion of Rationality:

This criterion is based upon what is known as the principle of insufficient reason. Since the probabilities associated with the occurrence of various events are unknown, there is not enough information to conclude that these probabilities will be different. This criterion assigns equal probabilities to all the events of each alternative decision and selects the alternative associated with the maximum expected pay off. Symbolically, if n denotes the number of events and P's denote the payoffs, then expected value for strategy, say S, is

$$\frac{1}{n}\left[P_1 + P_2 + \cdots + P_n\right]$$

Table 12.4

Alternatives	States of nature (product demand)				Expected Pay Off Rs.
	High Rs.	Moderate Rs.	Low Rs.	Nil	
Expand	50,000	25,000	- 25,000	- 45,000	$\frac{1000}{4}[50+25-25-45]=1,250$
Construct	70,000	30,000	- 40,000	- 80,000	$\frac{1000}{4}[70+30-40-80]=-5,000$
Subcontract	30,000	15,000	- 1,000	- 10,000	$\frac{1000}{4}[30+15-1-10]=8,500$

Thus the alternative 'sub contract' results in maximum average pay off of Rs. 8,500.

Example:

The following matrix gives the pay off of different strategies (alternatives) S_1, S_2, S_3 against condictions (evetns) N_1, N_2, N_3 and N_4 :

Table 12.5

	N_1 Rs.	N_2 Rs.	N_3 Rs.	N_4 Rs.
S_1	4,000	- 100	6,000	18,000
S_2	20,000	5,000	400	0
S_3	20,000	15,000	- 2,000	1,000

Indicate the decision taken under the following approach:

(i) Pessimistic (ii) Optimistic (iii) Regret and (iv) Equal probability.

Solution: For the given payoff matrix, the values corresponding to pessimistic, optimistic and equal probability criteria are given below:

Table 12.6

	Pessimistic (maximin) Value	**Optimistic (maximax) Value**	**Equal probability value** $= 1/n\,(P_1 + P_2 + \cdots + P_n)$
S_1	- Rs. 100	Rs. 18,000	$Rs \cdot 1/4(4,000-100+6,000+18,000) = Rs \cdot 6,975$
S_2	Rs. 0	Rs. 20,000	$Rs \cdot 1/4(20,000+5,000+400+0) = Rs \cdot 6,350$
S_3	- Rs. 2,000	Rs. 20,000	$Rs \cdot 1/4(20,000+15,000-2,000+1,000) = Rs \cdot 8,500$

Thus under pessimistic approach, S_2 is the optimal decision, under optimistic appraoch, S_2 or S_3 are the decision alternatives and under equal probability approach, S_3 is the alternative to be selected.

Table 12.7 represents the regret for every event and for each alternative calculated by the expression.

$$i^{th} \text{ regret} = (\text{maximum pay off} - i^{th} \text{ pay off}) \text{ for the } j^{th} \text{ event.}$$

Table 12.7

	N_1 Regret (Rs.)	N_2 Regret (Rs.)	N_3 Regret (Rs.)	N_4 Regret (Rs.)	Maximum Regret (Rs.)
S_1	16,000	15,100	0	0	16,000
S_2	0	10,000	5,600	18,000	18,000
S_3	0	0	8,000	17,000	17,000

The decision alternative S_1 would be chosen since it corresponds to the minimal of the maximum possible regrets.

12.3 Decision Making Under Conditions of Risk:

Most business decisions may have to be made under conditions of risk. Here more than one states of nature exists and the decision maker has sufficient information to assign probabilities to each of these states. These probabilities could be obtained from the past records or from simply the subjective judgement of the decision maker. Under conditions of risk, a number of decision criteria are available which could be of help to the decision maker.

Expected Value Criterion:

This criterion requires the calculation fo the expected value of each decision alternative which is the sum of the weighted payoffs for that alternative, where the weights are the probabilities assigned to the states of nature that can happen. Also known as expected monetary value (EMV) criterion, it consists of the following steps:

1. Construct a payoff table listing the alternative decisions and the various states of nature. Enter the conditional profit for each decision event combination along with the associated probabilities.
2. Calculate the EMV for each decision alternative by multiplying the conditional profits by assigned probabilites and adding the resulting conditional values.
3. Select the alternative that yields the highest EMV.

Example:

A newspaper boy has the following probabilities of selling a magazine.

No. of Copies Sold	Probability
10	0.10
11	0.15
12	0.20
13	0.25
14	0.30
	1.00

Cost of copy is 250 paise and sale price is 300 paise. He cannot return unsold copies. How many copies should be order?

Solution:

The no.of copies for purchases and for sales which have meaning to the newsboy are 10, 11, 12, 13 or 14. These are his sales magnitudes. There is no reason for him to buy less than 10 or more than 14 copies. Table 12.8, the conditional profit table, shows the profit resulting from any possible combination of supply and demand. Stocking of 10 copies each day will always result in a profit of 500 paise irrespective of the demand. For instance, even if the demand on some day is 13 copies, he can sell only 10 and hence his conditional profit is 500 paise.

When he stocks 11 copies, his profit will be 550 paise on days when buyers request 11, 12, 13 or 14 copies. But on days when he has 11 copies on stock and buyers buy only 10 copies, his profit decreases to 470 paise. The profit of 500 paise on the 10 copies sold must be reduced by 30 paise, the cost of one copy left unsold. The same will be true when he stocks 12, 13 or 14 copies. Thus the conditional profit in paise is given by

$$\text{Pay off} = 50 \times \text{copies sold} - 30 \times \text{copies unsold}.$$

Table 12.8

Conditional Profit Table (Paise)						
Possible Demand (no.of copies)	**Probability**	**Possible stock action**				
		10 copies	**11 copies**	**12 copies**	**13 copies**	**14 copies**
10	0.10	500	470	440	410	380
11	0.15	500	552	520	490	460
12	0.20	500	550	600	570	540
13	0.25	500	550	600	650	620
14	0.30	500	552	600	650	700

Next the expected value of each decision alternative is obtained by multiplying its conditional profit by the associated probability and adding the resulting values. This is shown in table 12.9

Table 12.9

Expected Profit Table						
Possible Demand	**Probability**	**Possible stock action**				
		10 copies	**11 copies**	**12 copies**	**13 copies**	**14 copies**
10	0.10	50	47	44	41	38
11	0.15	75	82.5	78	73.5	69
12	0.20	100	110	120	114	108
13	0.25	125	137.5	150	162.5	155
14	0.30	150	165	180	195	210
Total Expected Profit (paise)		**500**	**542**	**572**	**586**	**580**

The newsboy must, therefore, order 13 copies to earn the highest possible average daily profit of 586 paise. This stocking will maximize the total profits over a period of time. Of course there is no guarantee that he will make a profit of 586 paise tomorrow. How ever, if he stocks 13 copies each day under the condition given, he will have average profit of 586 paise per day. This is the best he can do because the choice of any one of the other four possible stock actions will result in a lower daily profit.

12.4 Marginal Analysis:

In any decision making problems, the use of conditional profit and expected profit tables would be quite cumbersome because of the large number of computations required. For 100 values of demand levels and stock actions, the calculations involved would be tremendous. This excessive computational work can be avoided by incremental or marginal approach according to this approach; any additional unit purchased will either be sold or remain un sold. If P represents the probability of selling one additional unit, then (1 - P) must be the probability of not selling it. If the additional unit is sold, the conditional profit will increase as a result of the profit earned from this unit. This is termed as marginal profit, MP . If the additional unit is not sold, the conditional profit reduces and the amount of reduction is called the marginal loss, ML resulting from stocking of an item that is not sold.

Additional units should be stocked so long as the expected marginal profit from stocking each of them is more than the expected marginal loss from stocking each. The expected marginal profit from stocking and selling an additional unit is the marginal profit of the unit multiplied by the probability that the unit would be sold i.e., P(MP). The expected marginal loss from stocking and not selling an additional unit is the marginal loss incurred if the unit remains unsold multiplied by the probability that the unit would not be sold i.e., $(1-P)(ML)$. Thus the units should be stocked upto the point where

$$P(MP) = (1-P)(ML)$$

$$(\text{or}) \qquad P = \frac{ML}{MP + ML}$$

The latter P represents the minimum required probability of selling at least on additional unit to justify the stocking of that additional unit. Additional units should be stocked so long as the probability of selling at least and additional unit is greater than P.

Example:

A milkman buys milks at Rs. 2 per litre and sells for Rs. 2.50 per litre. Unsold milk has to be thrown away. The daily demand has the following probability distribution:

Demand (litres):	46	48	50	52	54	56	58	60	62	64
Probability:	0.01	0.03	0.06	0.10	0.20	0.25	0.15	0.10	0.05	0.05

If each day's demand is independent of previous day's demand, how many litres should be ordered every day?

Solution:

Here M P = Rs. (2.50 - 2.00) = Rs. 0.50,

and M L = Rs. 2.00

The milk man should stock additional litres of milk so long as the probability of selling at least an additional litre of milk is greater than P, where

$$P = \frac{ML}{MP + ML} = \frac{2 \cdot 00}{2 \cdot 00 + 0 \cdot 50} = 0 \cdot 8$$

The value of 0.8 for P implies that in order to justify the stocking of an additional unit, there must be at least 0.8 cumulative probability of selling that unit. The cumulative probabilities of sales are computed in table 12.10. Additional units should be stocked so long as the probbaility of selling at least an additional unit is greater than P. The optimum number of litres of milk to be stocked is 54. If the number is increased to 56, the cumulative probability will become 0.60, which is less than the required value of 0.8

Table 12.10

Cumulative Probabilities of Sales

Sales (Litres of milk)	Probability of this scales level	Cumulative probability that Sales will be at this level or higher
46	0.01	1.00
48	0.03	0.99
50	0.06	0.96
52	0.10	0.90
54	0.20	0.80
56	0.25	0.60
58	0.15	0.35
60	0.10	0.20
62	0.05	0.10
64	0.05	0.05

Expected marginal profit = For P = 0.8, = P(M P) = RS. 0.8 x 0.50 = Rs. 0.40

Expected marginal loss = (1 - P) (M L) = Rs. 0.2 x 2.00 = Rs. 0.40

For 56 litres of stock level expected marginal loss will be more than expected marginal gain.

The use of marginal analysis yields the same solution as provided by the EMV and EOL approaches. How, ever, the computational effort required in this approach is much less.

12.5 Decision Tree:

A decision tree is a graphic representation of the decision process indicating decision alternatives, states of nature, probabilties attached to the states of nature and conditional benifits and losses. It consists of a network of nodes and branches. Two types of nodes are used: decision node represented by a asquare and state of nature (chance or event) node represented by a circle. Alternative courses of action (strategies) originate from the decision node as main branches (decision branches). At the end of each decision branch, there is a state of nature node from which emanate chance events in the form of sub - branches (chance branches). The respective pay offs and the probabilities associated with alternative courses and the chance events are shown along side these branches. At the terminal of the chance branches are shown the expected values of the out come.

The general appraoch used in decision tree analysis is to work backward through the tree from right to left, computing the expected value of each chance node. We then choose the particular branch leaving a decision node which leads to the chance node with the highest expected value. This is known as roll back or fold back process.

Ilustration: Suppose we have the decision making problem represented by the following table:

Table 12.11

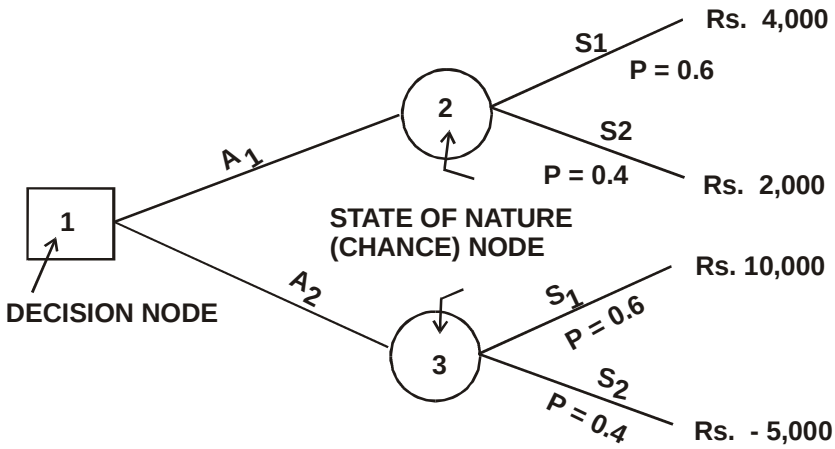
Conditional Profits			
States of Nature	**Probability**	**Alternative actions**	
		(A_1) **Produce 25 units**	(A_2) **Produce 75 units**
S_1 (High demand)	0.6	4,000	10,000
S_2 (Low demand)	0.4	2,000	- 5,000

The decision tree for the above problem is shown in Fig 12.1. For a decision alternative (strategy) the EMV is calculated by summing the products of payoff of each state and its probability.

For example, EMV of decision alternative A, (or node 2) is

= Rs. (4,000 x 0.6 + 2,000 x 0.4) = Rs. 3,200

Fig 12.1



Decision trees are useful for representing the inter - related, sequential and multi - dimensional aspects of a decision making problem. By drawing a decision tree, one is in a position to visualaise the entire complexity of the decision problem in all its dimensions as also the actual processes and stages for arriving at the final decision.

Steps in Decision Tree Analysis:

1. Identify the decision points and the alternative courses of action at each decision point systematically.
2. At each decision point determine the probability and the payoff associated with each course of action.
3. Commencing from the extreme right end, compute the expected payoffs (EMV) for each course of action.
4. Choose the course of action that yield the best payoff for each of the decisions.
5. Proceed backwards to the next stage of decision points.
6. Repeat above steps till the first decision point is reached.
7. Finally, identify the courses of action to be adopted from the beginning to the end under different possible outcomes for the situation as a whole.

Advantages and Limitations of Decision Tree Approach:

Advantages of the Decision Tree Approach:

1. It structures the decision proces and helps making in an orderly, systematic and sequential manner.
2. It requires the decision maker to examine all possible out comes, whether desirable or undesirable.
3. It communicates the decision making process to others in an easy and clear manner, illustrating each assumption about the future.
4. It displays the logical relationship between the parts of a complex decision and identifies the time sequence in which various actions and subsequent events would occur.
5. It is especially useful in situations where in the initial decision and its out come affects the subsequent decisions. It can be applied in various fields such as introduction of a new product, marketing, make or buy decisions, investment decisions, etc.

Limitations of Decision Tree Approach:

1. Decision tree diagrams become more complicated as the number of decision alternatives increases and more variables are introduced.
2. It becomes highly complicated when interdependent alternatives and dependent varibales are present in the problem.
3. It assumes that utility of money is linear with money.

4. It analyses the problem in terms of expected values and thus yields an average valued solution.

5. There is often inconsistency is assigning probabilities for different events.

Example:

A client asks on estate agent to sell three properties A, B and C for him and agrees to pay him 5% commission on each sale. He specifies certain conditions. The estate agent must sell property A first, and this he must do within 60 days. If and when A is sold the agent receives his 5% commission on that sale. He can then either back out at this stage or nominate and try to sell one of the remaining two properties within 60 days. If he does not succeed in selling the nominated property in that period, he is not given the opportunity to sell the other. If he does sell it in that period, he is given the apportunity to sell the third property on the same conditions. The following table summarises the prices, selling costs (incurred by the estate agent when ever a sale is made) and the estate agents estimated probability of making a sale.

Table 12.12

Property	Price of Property Rs.	Selling Costs Rs.	Probbaility of Sale
A	12,000	400	0.7
B	25,000	225	0.6
C	50,000	450	0.5

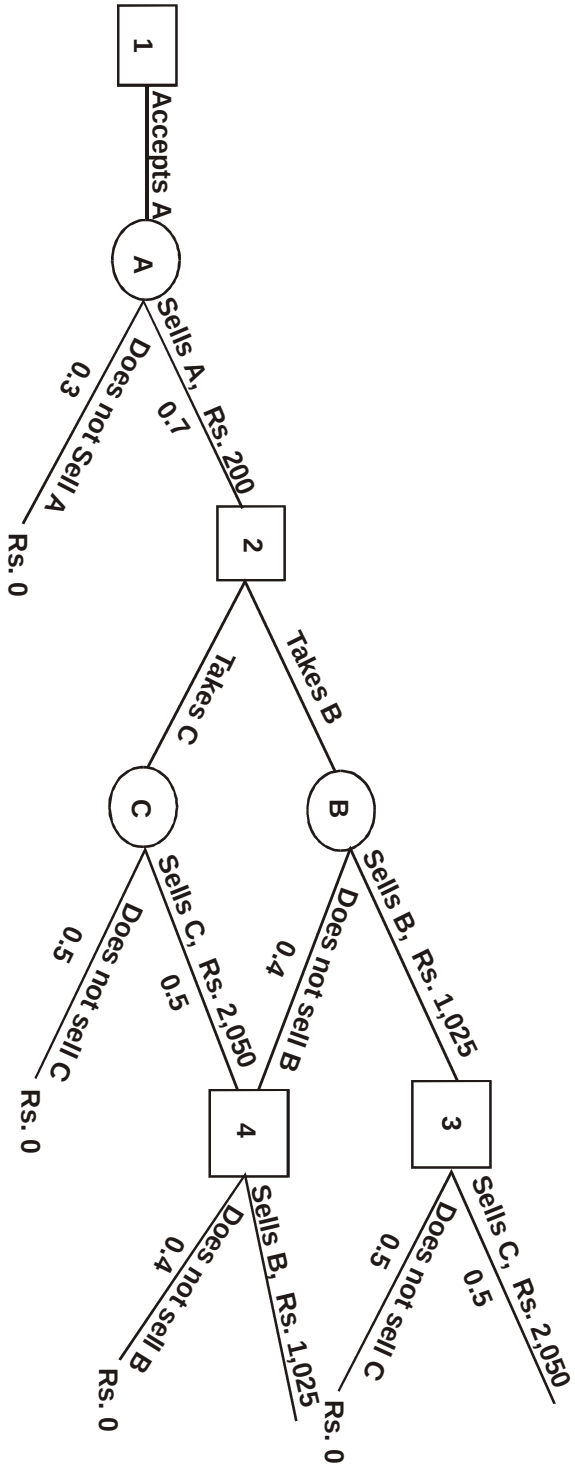
(i) Draw up an appropriate decision tree for the estate agent.

(ii) What is the estate agent's best strategy under EMV approach?

Solution:

The estate agent gets 5% commission if he sells the properties and satisfies the specified conditions. The amount he receives as commission on sale of properties A, B and C will be Rs. 600, Rs. 1,250 and Rs. 2,500 respectively. Since the selling costs incurred by him are Rs. 400, Rs. 225 and Rs. 450 his conditional profits from sale of properties A, B and C are RS. 200, Rs. 1,025 and Rs. 2,050 respectively. The decision tree for the problem is show in fig 12.2

Fig 12.2


1
Accepts A
A
Sells A, Rs. 200
0.7
Does not Sell A
0.3
Rs. 0
2
Takes B
Takes C
B
C
Sells B, Rs. 1,025
Does not sell B
0.4
Sells C, Rs. 2,050
0.5
Does not sell C
0.5
Rs. 0
3
4
Sells C, Rs. 2,050
0.5
Does not sell C
0.5
Rs. 0
Sells B, Rs. 1,025
Does not sell B
0.4
Rs. 0

EMV of node 3 = $\text{Rs.}(0\cdot5\times2,050+0\cdot5\times0) = \text{Rs. } 1,025$

EMV of node 3 = $\text{Rs.}(0\cdot6\times1,025+0\cdot4\times0) = \text{Rs. } 615$

EMV of node B = $\text{Rs.}\left[0\cdot6(1,025\,1,025)+0\cdot4\times0\right] = \text{Rs. } 1,230$

EMV of mode C = Rs. $\left[0\cdot5(2,050+615)+0\cdot5\times0\right] = \text{Rs. } 1,332\cdot50$

$\therefore$ EMV of node 2 = $\text{Rs. } 1,332\cdot50$ (Higher of the EMV of B and C)

$\therefore$ EMV of node A = $\text{Rs.}\left[0\cdot7(200+1,332\cdot50)+0\cdot3\times0\right] = \text{Rs}\cdot1,072\cdot75$

$\therefore$ EMV of node 1 = Rs. 1,072.75

The optimal strategy path is drawn in bold lines. Thus the aptimum strategy for the estate agent is to sell A; if he sells A then try to sell C and if he sells C; then try to sell B to get an optimum exptected amount of RS. 1,072.75

12.6 Preference Theory:

Observation that, all else being equal, people prefer to hold on to cash (liquidity) and that they will demand a premium for investing in non - liquid assets such as bonds, stocks and real estate. The theory suggests that the premium demanded for parting with cash increases as the period (term) for getting the cash back increases. The rate in the increase of this premium, however, slows down with the increase in term. In the language of finanical trading, this theory is expressed as "forward rates should exceed the future spot rates".

Revealed Preferece Theory:

Pioneered by American economist Paul Samuelson, is a method by which it is possible to discern the best possible option on the basis of consumer behavior essentially, this means that the preferences of consumers can be revealed by their purchasing habits. Revealed preference theory came about because the theories of consumer demand were based on a diminishing marginal rate of substitution (MRS). This diminishing MRS is based on the assumption that consumers make consumption decisions based on their intent to maximize their utility. While utility maximization was not a controversial assumption, the underlying utility functions could not be measured with great certainty. Revealed preference theory was a means to reconcile demand theory by creating a means to define utility functions by observing behavior.

Example:

If a person chooses a certain bundle of goods (ex 2 apples, 3 bananas) while another bundle of goods is affordable (ex 3 apples, 2 bananas), then we say that the first bundle is revealed preferred to the second. It is then assumed that the first bundle of goods is always preferred to the second. This means that if the consumer ever purchases the second bundle of goods then it is assumed that the first bundle is unaffordable. This implies that preferences are transitive. In other words if we have bundles A, B, C, ..., Z and A is revealed preferred to B which is revealed

prferred to C and so on then it is concluded that A is revealed preferred to C through Z. With this theory economists can chart indifference curves which adhere to already developed models of consumer theory.

The weak Axiom of Revealed Preference:

The Weak Axiom of Revealed Preference (WARP) is a characteristic on the choice behavior of an economic agent . For example, if an individual chooses A and never B when faced with a choice of both alternatives, they should never choose B when faced with a choice of A, B and some additional options. More formally, if A is ever choosen when B is available, then there can be no optimal set containing both alternatives for which B is chosen and A is not.

12.7 Summary:

The major points are:

(a) Statistical decision theory is applicable to decision making problems within an environment of risk or uncertainty.

(b) The decision theory process leads the manager to a "best" solution only if many complex value judgments have been properly made. These value judgments must be made in light of a through understanding of the problem.

(c) Statistical decision theory complements managerial experience and knowledge of business practices it does not and cannot replace the basic function of a manager decision making.

12.8 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: K. CHANDAN

Lesson - 13

SAMPLING METHOD

Objectives:

On successful completion of this lesson you should be able to:

- Significance of sampling
- Various sampling methods with their advantages and limitations
- Distinguish between probability and non - probability sampling
- Uses of stratified sampling
- Choice of approriate sampling technique

Structure:

13.1 Introduction:

SAmpling is a part of our day - to - day life. A quality controller takes a few items and decides whether the lot is in occardance with the desired specifications or not. A pathologist takes

a few drops of blood and tests for any change in blood of the whole body than normal. In all these situationssampling is invitable and gives satisfactory results.

Sampling is inevitable in the following situations:

(i) When population is infinite

(ii) When the item or unit is destroyed under investigation

(iii) When the result are required in a short time

(iv) When resources for survey are limited particularly in respect of money and trained persons.

(v) When area of survey is wide

Even in those cases where complete enumeration is possible it is not preferred due to the facts that it is much more time consuming and expensive, requires more skilled and techinical personnel, more errors are caused due to greater volume of work measurement errors etc... Complete enumeration is used only far various censures or in case of small population.

Population is agroup of items. Units or subjects which is under reference of study population may consist of finite infite real and hypothetical. Population is termed is universe by a number of statisticions and scientists.

13.2 Sampling:

Sample is a part of fraction of a population selected on some basis. Sample consists of a few items of a population. In principle a sample chould be such that it is a true representative of the population usually a random sample is selected. If the population is reasanably homogeneous, a simple random sample is most preferred one. But the moment one storts identifying sampling units on the basis of their characteristics, it gives rise to different sampling methods.

(1) Sampling Method: By sampling method we mean the manner or scheme through which the required number of units are selected in a sample from a population.

(2) Objective of sampling: The foremost purpose of sampling is togather maximum information aboutthe population under consideration at minimum cost, time and human power. This is best achieved when the sample contains all the properties of the population.

(3) Sampling unit and its Examples: The constituents of a population which are the individuals to be sampled from the population and cannot be further sub divided for the purpose of sampling at a time are called sampling units. For instance, to know the average income per family, the head of the family is a sampling unit. To know the average yield of wheat, each farm owner's field of wheat is a sampling unit.

(4) Sampling Frame: For adopting sampling procedure it is essential to have a list or a map identifying each sampling unit by a number. Such a list or map is called sampling frame.

A list of voters, a list of house holds, a list of technical persons, areas in a map marked by numbers for soil surveys, a list of villages in a district, a list of farmers fields etc... are a few examples of sampling frame.

(5) Distinguish between complete enumeration and sampling study: In complete enumeration each and every unit of the population is studied and results are based on all units of the population. Whereas, in sampling study only a selected number of units are studied and results based ont he data of these units are supposed to yield information about the whole population.

13.3 Methods of Sampling:

When it is decided to take sample from the population, it is necessary to choose some methods of sampling. There are many methods of choosing a sample from the population. The choice of the sampling method depends upon the nature of data and the purpose of the enquiry. The various methods of sampling also called sampling design can be grouped into two broad heads (1) Random sampling (2) Non - random sampling.

(1) Random Sampling: It is otherwise called as probbaility sampling. In probbaility sampling all the items in the population have a chance of being choosen in the sample. However random sampling does not mean haphazard selection or the term random sample is not used to describe the data in the sample but the process employed to selectthe sample. Randomness is thus a property of sampling procedure instead of individual sample.

Advantages of Random Sampling: The following are the basic advantages of random sampling.

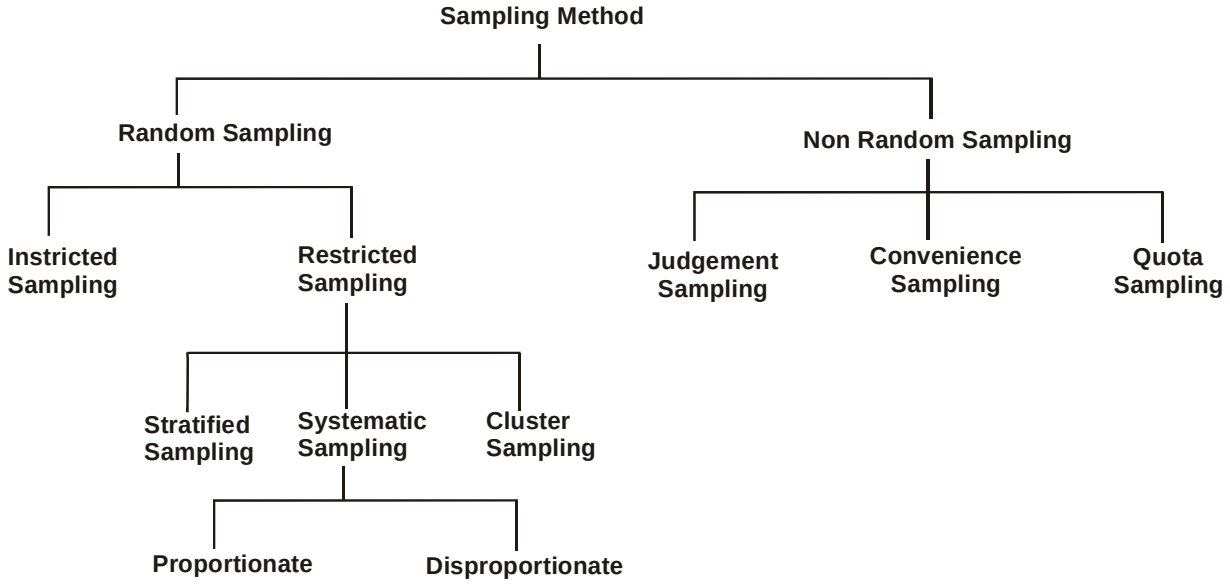
1) Random sampling provides estimates which are not biased
2) Existence of detailed information about universe is not required
3) Helps in evaluating the relative effciency of various sample designs

Limitations of Random Sampling: In spite of the great advantages of random or probbaility sampling it suffers from some limitations because of which non probability sampling or non random sampling techiniques are used. The main limitaions are:

1. It requires very high level of skill and experience for its use.
2. It takes a lot of time to plan and execute a probbaility sample
3. The cost involved are generally high as compared to non random sampling

(2) Non - Random Sampling: Non - random sampling is a process of sampling without use of randomisation. A non random sample is selected on the basis of some one's experience about the population or convenience. That is why it is otherwise called as Purposive sampling where the basis of selection is other than probability consideration. Purposive sampling means selecting the items of the sample in accordance with some purposive principle. In this method the criterion for selection is laid down first and items and selected in accordance with it. Thus the probability of inclusion of some units are very high, while the probability of inclusion of others very low. In this sampling procedure personal element has a greater chance of entering into selection of sample. Sometimes the investigator may select a sample to yield favourable results - in results bias on sampling technique. In small enquires this sampling design may be helpful with its relative advantage of time and cost but samples so selected do not possess statistical characteristics to make inference about the population.

The sampling methods can be classified further as follows:



13.4 Simple random sampling:

(i) Unrestricted random sampling: In this type of sampling each and every unit of the population has equal chance of being included int he sample. Simple random sampling is an exaple of unrestricted sampling.

(ii) Restricted Sampling: If aninvestigator has any idea aboutthe heterogenity of sampling units, the population is divided into homogeneous groups and sample is drawn independently from each group. Such a process of sampling, multistage sampling, etc... are covered underthe category of restricted sampling.

Difference between sample random sampling with replacement and without replacement. If the units are selected or drawn one by one in such a way that a unit drawn at a time is replaced back to the population before the subsequent draw, it is known as simple random sampling with replacement (Srswr). In this type of sampling from a population of sizze N, the probability of a selection of a unit at each draw remains 1/N. In Srswor, a unit can be included more than once in a sample. Therefore, if the required sample size is n, the affective sample size is sometimes lessthan n due to the inclusion of one or more units more than once.

With the idea that effective sample size should be adhered to the simple random sampling without replacement (srswor) is adopted in this method a unit selected once is not included in the population at any subsequent draw. Hence, the probability of drawing a unit from a population of N units at $_{r}^{th}$ draw is $1/(n-r+1)$.

In simple random sampling the probbaility of selection of any sample of size n from a population consistingof N units remains the same i.e., $1/\begin{bmatrix} N \\ n \end{bmatrix}$ here $\begin{bmatrix} N \\ n \end{bmatrix}$ is the number of all possible sample.

Factors responsible for the size of a sample:

The size of a simple depends upon the following factors:

1. The purpose for which the sample is drawn.
2. The heterogeneity of the sampling units in the population. More is the neterogeneity, larger is the size of the sample.
3. Resources available for the study in terms of time and money
4. Number of technicak persons and / or equipment available.
5. Precision of estimates required is an important factor in determining the size of a sample greater isthe precision required, usually a large sample is preferred.

Sample Mean:

If n sample observation are $x_1, x_2, \cdots, x_n$ the formula for sample mean is,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

Simple Varience:

If $x_1, x_2, \cdots, x_n$ are the observations on n sample units and $\bar{x}$ is its mean, the formula for varience is,

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

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \right]$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n} \right\}$$

To ensure randomness in the selection one may adopt either (1) lottery method or (2) consult table of random numbers.

(1) Lottery Method: In this method all the items in the population are numbered. The numbers are written in small chits of papers or on cards which are homogeneous in all respects. These numbered chits of paper must be identical in size, Colour and shape. All the numbered slips are then folded and placed in a drum or container and well shuffled. A blind folded

selection is then made of the number of slips required to constitute the desired size of the sample. The numbers corresponding to the slip drawn will constitue the sample. The selection of item depends entirely on chance. However the possibility of personal prejudice or bias cannot be ruled out if the slips are not of uniform size or shape. This method works well with small population. There is the added problem too of not being certain that the chits are not mixed properly.

2. Table of Random Numbers: The lottery method becomes quire cumbersome to use with a large population. Personal bias also cannot be excluded altogether. At the same time randomness in the drum is not as simple as it appears to be because slips may stick together or to the sides. If the slips are too much they do not get mixed thoroughly. When the population is too large it takes too much time to number individual members of the population. Since lottery method is not always practicable an alternative method of selection fo random sample is employed. It is that of using the table of random number. Severral standard random number tables are available of which any one can be used.

They are:

(i) Tippet's table of random numbers

(ii) Fisher and Yates table of random numbers

(iii) Kendall and Babington Smith numbers

Among these tables Tipper's table of random numbers is most popular and widely employed for selecting random sample. The table consists 41600 random digits grouped into 10400 sets of four digit random numbers. These numbers have been put to various tests from time to time and their randomness have been proved. The digits in each cloumn and in each row are in random number. It makes no difference from where one starts and in which direction one proceeds. The column arrangement is considered most convenient and the number of columns to be used depends on the size of the sample.

The method of drawing random sample comprises of the following:

1. Suppose we had a population of 600 units, identify all the units with a serial number ranging from 000 to 599.

2. The select at random any page of random number table and pickup nymbers given in it in any row, column or diagonal at random and also specify the sequence of selection of numbers.

3. Suppose we have specified to select first column and planning to move down to pickup further numbers then choose first three digits of the first number in random table. Verify whether it falls within our required range of 000 to 599. If it falls we accept the same as a number for our random sample, otherwise we discard it and proceed further.

4. Listing this procedure we selct the numbers to the extent of required sample size.

Use of Random Number Table:

Suppose we want 10 sample units out of 600 population size, using random numbers given in Appendix, if we follow 2st column and downward the following numbers would result, 126, 547, 371, 519, 171, 105, 421, 347, 041, 536.

Thus it may be easy to draw good random simple from random number table but it is more suitable when the population is finite and it is possible to number them.

Example:

Merits and Limitations:

Merits:

1. As the selection of items in sample depend entirely on chance sampling is not affected by personal judgement or bias.
2. The investigator can assess the accuracy of his results because the sampling error is inversely proportional to square root of number items in the sample.
3. AS comparred to judgement sampling a random sample represents the universe in a better way. As the sample size increase it become increasingly representative of population.
4. It is relatively inexpensive and needs much less time and energy.

Limitations:

This method suffers from limitations like:

1. The use of simple random sampling necessitates numbering of the whole population. In case of very large population numbering the entire population is extremely costly and time consuming. In many cases, it is often impossible to number each member of the population.
2. The size of sample required to ensure statistical reliability is usually larger under random sampling.
3. In cases of hetcrogeneous population, the random sample will fall to depict the true characteristics of a population as some of the groups may not be represented at all in the sample.

Example:

A some what stilted, if accurate, definition. Let's see if we can make it a little more real. How do we select a simple random sample? Let's assume that we are doing some research with a small service agency that wishes to assess client's views of quality of service over the past year. First, we have to get the sampling frame organized. To accomplish this, we'll go through agency records to identify every client over the past 12 months. If we're lucky, the agency has good accurate computerized records and can quickly produce such a list. The, we have to actually draw the sample. Decide on the number of clients you would like to have in the final sample. For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is $\mathrm{f} = \mathrm{n}/\mathrm{N} = 100/1000 = \cdot 10 \text{ or } 10\%$. Now to actually draw the sample, you have several options. You could print off the list of 1000

clients, tear then into separate strips, put the strips in a box mix them up real good, close your eyes and pull out the first 100. But this mechanical procedure would be tedious and the quality of the sample would depend on how throughly you mixed them up and how randomly you reached in. Perhaps a better procedure would beto use the kind of ball machone that is popular with many of the state lotteries. You would need three sets of balls numbered 0 to 9. One set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list of names from 1 to 1000 and then use the ball machione to select the three digits that selects each person. The obvious disadvantage here is that you need to get the ball machines. (Where do they make those things, anywar? Is there a ball machine industry?)

Neither of these mechanical procedures is very feasible and with the development of inexpensive computers there is a much easier way. Here's a simple procedure that's especially useful if you have the names of the clients already on the computer. Many computer programs can generate a series of random numbers. Let's assume you can copy and paste the list of client names into a column in an EXCEL spreadsheet. Then, in the column right next to it paste the function = RAND () which is EXCEL's way of putting a random number between 0 and 1 in thecells. Then, sort both columns - the list of names and the random number - by the random numbers. This rearranges the list in random order from the lowest to the highest random number. Then, all you have to do is take the first hundred names inthis sorted list. Pretty simple. You could probably accomplish the whole thing in under a minute.

13.5 Stratified Sampling:

Stratified sampling comes under the category of restricted sampling. In this type of sampling method first the whole population is divided into homogenous groups under certain criterion. These groups are termed as strata. The sample is drawn randomly from each stratum independently. the estimates are calculated from the data obtained from all the strata.

Information about each individual sampling unit is rarely available. Hence, the strata are formed on some broad basis such as localities ina city, districts in a state etc.

Advantages of stratified sampling:

Different advantages of stratified sampling can be summarised as follows:

1. If the admissible error is given small sample is needed which results into a cut of expenditure.
2. In case the cost of survey isfixed, there is reduction in error due to stratification.
3. Through stratification it is possible to gather the information or obtain the estimates for each stratum seperately and also an estimate for the whole population.
4. Stratified sampling is very convenient from organisational point of view.
5. If need be different sampling schemes can be used to draw samples from different strata. But it creates many complications and hence it is mostly avoided.

Two - way stratification:

Sometimes the population is stratified according to two factors, e.g., the persons are stratified according to their qualifications and monthly income. In this way we have a two way table and units belong to different cells or substrata. Samples are drawn from each substratum (cell) independently.

Such a sampling procedure is known as two - way stratified sampling. Two way stratification is also termed as deep as deep stratification. Further two - way stratification is generolly more efficient than one way stratification still it is eldom used. Make use of varying probabilities of selection in stratified sampling. Yes, often the units within a stratum are selcted with replacement and with varying probabilities of selection.

Proportionate and Disproportionate Stratified Random Sample:

A stratified sample may either be proportionate or disproportionate in a proportional stratified sampling plan an equal proportion of units are drawn from each stratum for example if the population consists of 5 groups their proportions are 5, 15, 20, 25 and 35 percent of the population and a sample of 1000 is drawn the desired proportional sample may be obtained in the following manner.

From stratum one	1000 x 0.05	=	50
From stratum two	1000 x 0.15	=	150
From stratum three	1000 x 0.20	=	200
From stratum four	1000 x 0.25	=	250
From stratum five	1000 x 0.35	=	350
Size of entire sample			1000

In disproportionate stratified sample an equal number of cases is drawn from each stratum regardless of how the stratum is represented. Thus in the above sample 200 from each stratum may be drawn. In case of proportional plan the total number of samples would properly represent all the strata. This eliminates the difference between the strata and reduces sampling error. Such a method adds to the precisions of the sample estimate when within strata variablitiy is least.

Example:

Limitations:

1. It needs complete and upto data list of finite population. Such lists are not generally available.
2. Utmost care must be excercised in dividing the population into various strata. Each stratum must contain as far as possible homogeneous items, as otherwise results may not be reliable. In the absence of proper stratification the sample may have effect of blas.
3. It is tedious and time consuming to stratify the population also in many cases information needed to set up groups may not be available.
4. The items from each stratum should be selected at random which needs skilled sampling supervision.

For Example:

Let's say that the population of clients for our agency can be divided into three groups. Caucasian, African - American and Hispanic - American Furthermore, let's assumethat both the African - Americans and Hispanic Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of n = 100 with a sampling fraction of 10% we would expect by chance alone that we would only get 10 and 5 persons from each of our

two smaller groups. And by chance we could get fewer than that ! If we stratify we can do better. First let's determine howmany people we want to have in each group Let's say we still want to take a sample of 100 from the population of 1000 clients over the past year. But we think that in order to say anything about subgroups we will need at last 25 cases in each group.

N = 100	1	26	51	76
	2	27	52	77
	3	28	53	78
	4	29	54	79
	5	30	55	80
	6	31	56	81
want n = 20	7	32	57	82
	8	33	58	83
	9	34	59	84
	10	35	60	85
	11	36	61	86
N/n = 5	12	37	62	87
	13	38	63	88
	14	39	64	89
	15	40	65	90
select a random number from 1 - 5 : chose 4	16	41	66	91
	17	42	67	92
	18	43	68	93
	19	44	69	94
	20	45	70	95
	21	46	71	96
start with # 4 and take every 5th unit	22	47	72	97
	23	48	73	98
	24	49	74	99
	25	50	75	100

So, let's sample 50 CAucasians, 25 AFrican - Americans and 25 Hispanic - Americans. We know that 10% of the population, or 100 clients are African - American. If we randomly sample 25 of these, we have a within - stratum sampling fraction of 25/100 = 25%. Similarly, we know that 5% or 50 clients are Hispanic - American. So our within - stratum sampling fraction will be 25/50 = 50%. Finally, by subtraction we know that there are 850 Caucasian clients. Our within - stratum sampling fraction for them is 50/850 = about 5.88%. Because the groups are more homogeneous within group than across the population as a whole, we can expect greater statistical precision (less variance). And because we stratified, we know we will have enough cases frome ach group to make meaningful subgroiup inferences.

13.6 Systematic Sampling:

When the population units occur in a deck or sequence or line and a sample of size n is to be drawn, the population is divided into n sequential groups and one unit us drawn from each group situated at equal distances.

The selection procedure is such that one unit is drawn randomly fromt he first group say j^{th} unit is selected. Then select $\left(j^{th}\right), (j+2k), \ldots, \left(j+\overline{n-1}k\right)^{th}$ units from the subsequent groups. Such a selection procedure is known as linear systematic sampling this procedure fails if the population size N is not a multiple of n.

Advantages of systematic sampling:

Some of the principle advantages of systematic sampling are given below:

1. The method of selection is very simple.
2. The method of selection is cheap in terms of time and money.
3. The sample is distributed over the whole population and hence all contiguous parts of the population are well represented in the sample.
4. It is easy to locate selected units and is very convenient from organisational point of view.

Disadvantages of Systematic Sampling:

1. If the variation in the units is periodic, the units at regular intervalsare cirrelated. In this situation the sample becomes highly lop - sided and hence the estimates are biased.
2. No single reliable formula is available for estimating the standard error of sample mean. A formula is good enough if the population is of the type it has been expected to otherwise not this is a great drawback of systematic sampling.

Situations Systematic Sampling is preferred over other sampling procedures:

Systematic sampling is preferably used when the information is to be collected from cards, trees in a forest, houses in blocks, entries in a register which are in a serial order etc...

Circular Systematic Sampling:

Linear systematic sampling fails if $N \neq nk$. Circular systematic sampling was first used by D.B. Lahri in 1952. In circular systematic sampling take $\frac{N}{n}$ as rounded to the nearest integer. Select a random number from 1 to N. Suppose the selected number is m. Now select every $(m + jk)^{th}$ unit when $m + jk < N$ and every $(m + jk - N)^{th}$ unit when $m + jk > N$ putting j = 1, 2, ..., tilln units are selected. Such a procedure of selection is known as circular systematic sampling.

Compare systematic sampling with stratified sampling:

(i) Systematic Sampling resembles stratified sampling in the sense that groups of k units look like strata but no criterion has been considered in the formation of groups that ensures homogeneity.

(ii) No independent samples are drawn from each group.

(iii) In systematic sampling we have only one sample from the whole population. The above three points clearly reveal that systematic sampling is quite different from stratified sampling.

Systematic sampling superior to simple random sampling and stratified sampling comment:

Nothing definite can be said about the superiority of systematic sampling over simple random sampling and vice - versa as it depends more on the structure of the population one may be better than the other in particular cases. The same statement holds good while comparing systematic sampling with stratified random sampling.

13.7 Cluster Sampling:

In many situations, the sampling frame for elementary units of the population is not available, moreover it is not easy to prepare it. But the information is available for groups of elemetns so called clasters. For instance the list of houses may be available but not the persons residing in them, lsit of individual forms may not be available but the list of villages is generally available. Hence in these situations houses or villages is generally available. Hence, in these situations houses or villages are known as clusters and selection has to be made of houses or villages in the sample. Such a sampling procedure is known as cluster sampling.

When the entire area cotnaining the population is divided into smaller area or segments, these small areas or segments are taken as sampling units. This procedure is known as area sampling. Clusters or area segments are also known as primary units.

In cluster sampling, preaution should be taken that a unit should never belong to mroe than one cluster. Also each elementary unit of the population should definitely belong to one primary unit.

Situtations the Cluster Sampling be Preferred:

The cluster sampling is used when:

(i) The sampling frame is not availbele and it is too expensive and time consuming to prepare it.

(ii) The sampling units are situated distant apart. In thsi situation selection of elementary units makes the survey very cumbersome. For instance, selction of farmers in a state.

(iii) The elementary units may not be easily identifiable and locatable. For example, the animals of certain species, the migratory populations etc.

Sampling design is used to select clusters from a population:

Usually simple random sampling without replacement is used to select clusters or area segmetns from a population. But any other design can also be used.

Efficiency of cluster sampling as compared to simple random sampling without replacement.

The efficiency of cluster sampling as compared to simple random sampling is,

$$E = \frac{S^2}{M S_b^2}$$

Where, $S^2 =$ The varience of cluster

$M =$ The cluster size for each cluster

$S_b^2 =$ The difference of the total varience and with in clusters varience

From the formula it is apparent that the efficiency of the cluster sampling increases.

(i) If the cluster size decreases

(ii) If the clusters are so formed that the variation within clusters is as large as possibleand between clusters is as small as possible.

Main differences between cluster sampling and stratified sampling are:

(i) In stratified sample, a sample is drawn frome ach stratum (cluster) whereas in cluster sampling, a cluster (stratum) is selected as such.

(ii) A reterogeneous cluster is more prefereable where as a homogeneous stratum is always desirable.

13.8 Double Sampling:

Many a time, some initial information necessitated to draw a sample is available. In that situation such information can be gathered by taking a large sample provided it is not very expensive and time consuming. Then take a sub sample to estimate the main characters as per the objectives of the survey. Such a sampling process is known as double sampling. Double sampling is also called two prases sampling. For example an investigator want to draw a sample of agricultural holdings with probability proportional to size. But the areas of holding is collected on alarge number of former's holdings and then a sample is selected with pps for the main survey.

Double sampling as also helpful in stratified sampling to determine the strata sizes if not known. Double sampling is also very helpful in ratio and regression method of estimation.

13.9 Multi Sampling:

In simple - stage cluster sampling it is costly to include every elementary unit of the selected clusters in the survey. Moreover, it appears superfluous when the clusters are homogeneous. Hence, it is better to select as sample from each selected cluster rather than surveying the clusters as a whole. Selection fo a sample from each selected cluster is known as sub sampling. Under such a sampling procedure a sample is drawn in two stages i.e., in the first stage a sample of clusters is selected, and inthe second stage a sample of elementary units is drawn from each selected cluster. This kind of sampling procedure is known as two stage sampling. For example, if a survey is conducted to have an estimate of crop production, one may prefere to use two - stage sampling. Select villages as first stage units and farms in the villages as second stage units.

The selection procedure can be extended to any number of stages. Hence, in general it is known as multi - stage sampling.

13.10 Two stage sampling design:

It has been found that two stage sampling design is generally less efficient than single stage sampling except when the correlation between elements in the same (first) stage units is negative.

Inverse Sampling:

When the character under study rarely exists and the proportion P of units possessing he character isvery small, a simple random sample without replacement does not yield satisfactory results. Hence, for getting good estimate of P. Haldane, 1946 and Fimmey 1949 suggested the method of inverse sampling. In inverse sampling, the size of the sample 'n' is not fixed but selected process continues unitl a prdecided number of units posessing the rate character or attributes has been selected in the sample.

Suppose N is the number of units in the population and P, the proportion of units possessing the rare character or attribute under study. The number of units passessing the rare character is

NP. Now to estimate P, draw a sample with srswor until the sample contains m units having rare character. Let n be the total number of units selected containing m units of intrest. in this type of selection n is a random variable and follows hypergemetric distribution. An unbaised estimate of P

is, $P = \frac{m-1}{n-1}$

An unbiased estimate of the varience of P is

$$S_p^2 = \frac{P(P-1)}{(n-2)}\left[1 - \frac{n-1}{N}\right]$$

If $\frac{n-1}{N}$ is negligible,

$$S_P^2 = \frac{P(P-1)}{n-2}$$

13.11 Non Probability Sampling:

The difference between non propability sampling is that nonprobability sampling does not involve random selection and propability sampling does. Does that mean that non probability samples aren't representative of the population? Not necessarily. But it does ,eam tjat mpm propability samples cannot depend upon the rationale of probability theory. AT least with a probabilistic sample,w e know the odds or probability that we have represented the piopulation well. We are able to estimate confidence intervals for the statistic. With non probability samples, we may or may not represent the population well ad it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over nonprobabilistic ones and consider them to be more accurate anmd rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here we consider a wide range of non probabilistic alternatives.

We can divide nonprobability sampling methods into two broad types accidnetal or purposive. Most sampling methods are purposive in nature because we usually approach the sampling problem with a specific plan in mind. The most important distinctions among these types of sampling methods are the ones between the different types of purposive sampling approaches.

Accidnetal, Haphazard or Convenience Sampling:

One of the most common methods of sampling goes under the various titles listed here. I would include in this category the traditional man on the street (of course now it's probably the person on the street) interviews conducted frequently by television news programs to get a quick (although nonrepresentative) reading of public opinion. I would also argue that the typical use of college students in much psychological research is primarily a matter of convenience. (You don't really believe that psychologists use college students because they believe they;re representative of the population at large do you?). In clinical practice we might use clients who are available to us as our sample. In many research contets, we sample simply by asking for volunteers. Clearly, the problem with all of these types of samples is that we have no eidence that they are representative of the populations we're interested in generalizing to and in many cases we would clearly suspect that they are not.

Purposive Sampling:

In purposive sampling, we sample with a purpose in mind. We usually would have one or more specific predefined groups we are seeking. Fro instance have you ever run into people in a mall or on the street who are carrying a clipboard and who are stopping various people and asking if they could interview them? Most likely they are conducting a purposive sample (and most likely they are engaged in market research). They might be looking for Caucasian females between 30 - 40 years old. They size up the people passing by and anyone who looks to be in that category they stop to ask if they will participate. One of the first things they're likely to do is verify that the respondent does in fact meet the criteria for being in the sample. Purposive sampling can bevery useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern. With a purposive sample, you are likely to get the opinions of your target population, but you are also likely to overweight subgroups in your population that are more readily accessible.

All of the methods that follow can be considered subcategories of purposive sampling methods. WE might sample for specific groups or types of people as in modal instance, expert or quota sampling. We might sample for diversity as in heterogeneity sampling. Or, we might capitablize on informal social networks to identify specific respondents who are hard to locate otherwise as in snowball sampling. In all of these methods we know what we want we are sampling with a purpose.

Modal Instance Sampling:

In statistics, the mode is the most frequently occuring value in a distribution. In sampling, when we do a modal instance sample, we are sampling the most frequent case, or the "typical" case. In a lot of informal public opinion polls, for instance, they interview a "typical" voter. There are a number of problems with this sampling approach. First, how do we know what the "typical" or "modal" case is? We could say that the modal voter is a person who is of average age, educational level and income in the population. But it's not clear that using the averages of these is the fairest (consider the skewed distribution of income for instance). And how do you know that those three variables age education income are the only or even the most relevant for classifying the typical voter? What if religion or ethnicity is an importnat discriminator? Clearly, modal instance sampling is only sensible for informal sampling contexts.

Expert Sampling:

Expert sampling involves the assembling of a sample of persons with known or demonstrable experience and expertise in some area. Often we convene such a sample under the auspices of a "panel of expers." There are actually two reasons you might do expert sampling. first, because it would be the best way to elicit the views of persons who have specific expertise. In this case, expert sampling is essentially just a specific subcase of purposive sampling. But the other reason you might use expert sampling is to provide evidence for the validity of another sampling approach you've chosen. For instance let's say you do modal instance sampling and are concerned thatt he criteria you used for defining the modal instance are subject to criticism. You might convene an expert panel consisting of persons with acknowledged experience and insight into that field or topic and ask them to examine your modal definitions and comment on their appropriateness and validity. The advatnage of doing this is that you aren't out on our own trying to defend your decisions you

have some acknowledged experts to back you. The disadvatnage is taht even the experts can be, and often are wrong.

Quota Sampling:

In quota sampling, you select people nonrandomly according to some fixed quota. There are two types of quota sampling proportional and non proportional. In proportional quota sampling you want to represent the major characterisitcs of the population by sampling a [roportional amount of each. For instance if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So if you've already got the 40 women for your sample, but not the sixty men,. you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota". The problem here (As in much purpoisive sampling) is that you have to decide the specific characteristics on which you will base the quota. will it be by gender age education race religion etc...?

Non proportional quota sampling is a bit less restrictive. Inthis method you specify the minimum number of sampled units you want in each category. Here you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population. This method is the non probabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in your sample.

Heterogeneity Sampling:

We sample for heterogeneity when we want to include all opinions or views and we aren't concerned about representing these views proportionately. Another terms for this is sampling for diversity. In many brainstorming or nominal group processes (including concept mappong) we would use some form of heterogeneity sampling because our primary interest is in getting broad spectrum of ideas not identifying the average or modal instance ones. In effect what we would like to be sampling is not peiople, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population not the population of people who have the ideas. Clearly in order to get all of the ideas and especially the outlier or unusual ones. We have to include a broad and diverse range of participants. Heterogeneity sampling is in this sense, almost the opposite of modal instance sampling.

Snowball Sampling:

In snowball sampling you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who tehy may know who also meet the criteria. although this method would hardly lead to representative samples there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find. For instance if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However if you go to that area and identify one or two you may find that they know very well who the other homeless people in their vicinity are and how you can find them.

Judgement Sampling:

In judgement sampling as the name means, selection of items to be included in the sample depends on the judgment discretion of the investigaor,. In other words the investigator exercises

his judgment to choose the sample most typical of the universe with regard to the characteristics. In this sampling technique that character or qualitites of the universe about which information is to be collected forms the basis of the judgment in slectiong the sample. For example if investigation has to be done about the expenditure of students in a hostel. Then under this system the investigator will pickup such students who are neither miserly nor extravagant.

Choice of Appropriate Sampling Technique:

It is very difficult to say that any one of the technique discussed above would always be the best. AS each method has got its speciality no one method is regarded the best in all circumstances. FActors likes size of the sample, size of universe availability of finance time nature of the universe etc... would influence the selection of a particular method of sampling.

Merits and Limitation of Sampling:

Merits:

(1) Less Time Consuming: Since sample is a study of a part of universe considerable time and efforts are saved not only in collecting data but also in its processing.

(2) Economical: Although the amount of effort and expenses involved in collecting informations is always greater per unit of sample than the complete census the total financial burden is less than that of census.

(3) More Reliable Result: Inaccuracy of inforamtions. Incompleteness of returns are likely to be more serious in census method than sampling method as more effective precautions can be exercised in sample survey to ensure that information is accurate and complete. AT the same time service of experts to impart through training to investigators in asample survey can reduce the possibility of errors.

(4) More Detailed Information: Since sampling technique saves time and money it is possible to collect more detailed information in sample survey.

(5) Non Applicability of Census Method In All the Studies: There are some cases in which census method is inapplicable. If the population is infinite or the investigation destroys the population unit. Sampling method only can be adopted.

Limitations:

Despite of several advantages, ampling is not altogether free from limitaions. The followings are the main limitaions of sampling:

1. Unless the sample survey is properly planned the results obtained would be inaccurate and missleading.
2. In the absence of qualified experienced and unbiased persons the results, information from sample survey cannot be relied upon.
3. If the information is required for each and every unit in a population we need to depend upon census than sample survey.
4. There is every likehood of sampling and non sampling errors.
5. There is possibility of bias on the part of investigator regarding inclusion or exclusion of items in the sample.

13.12 Summary:

In this lesson we have looked various sampling methods available when one wants to make some interenvees about apopulation without essumerating it completcly. Looking at some situations where sampling was being done and then found that in many situations sampling may be the only feasible way of knowing some thing about the population. In some caases complete enumeration is possible it is not preferred. We noted that there are two basic two baisc methods if samnpling probabilioty sampling anon probabilty samnpling. among the probbaility sampling methods simple random sampling method is the best. We have also discussed some of the non probbaility sampling methods.

13.13 Exercises:

1. What is the need of sampling as compared to complete enumeration?
2. In what situations ampling is in eviatable?
3. What is a sample? What is meant by sampling method?
4. Give different types of sampling schemes and describe them in brief?
5. What do you understand by simple random sampling?
6. What do you understand by stratified random sampling?
7. What do you understand by systematic sampling?
8. What is cluster sampling?
9. Discuss double sampling?
10. Discuss multi sampling?
11. Discuss two stage sampling?
12. What are non probbaility sampling? and explain briefly.

13.14 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.
2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.
3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. K. MADHU BABU

Lesson - 14

SAMPLING DISTRIBUTION

Objectives:

After reading this lesson you should be able to

- to understand the meaning of sampling distribution and standard error
- to develop confidence intervals of the population mean and population proportion

Structure:

14.1 Introduction:

One resorts to sampling for a variety of reasons. To mention some, the sample bestows economy in collecting data, timeliness in deciding promptly, However, it is of common interest to draw generalisations about population with sample information in most cases. The procedure involves complex concepts which need clarification and hence the present lesson is devoted to illustrate the sampling distribution for different statistics. The importance of such distributions, their parameters and their contribution in estimating the population characteristics are presented here.

14.2 Sampling Distribution:

A sampling distribution is a theoretical probability distribution of any sample statistic that would result from drawing all possible samples from a population.

However it would not be practical possible to obtain the sampling distribution empirically, and hence theoretically coneived. There would be a sampling distribution for every characteristic of the sample mean, variance, proportion etc...

Sampling distribution and statistical inference:

Sampling distribution is nothing but a theoretical probability distribution of a sample statistic - would be of much help to find answers to the issues of how far the mean of a sample $\left(\bar{x}\right)$ is exactly equal to the mean of the population $\left(\mu\right)$ from which it is drawn? It provides the basis for considering nearerness of sample means to population mean.

Example of sampling distributions

Population	Sample	Sample Statistic	Sampling Distribution
University Graduates	30 Graduates from each affiliated college	Average Entrance exam., score	Sampling distribution of means
All colour picute tubes produced by Mfg. firm	50 picture tubes from each production lot	Proportion of defectives	Sampling distribution of proportions
Inter - university Volley - ball team	Groups of 5 players	Median height	Sampling distribution of Median
Electric bulbs Manufactured	Groups of 100 bulbs from each lot	Variance in the average life of burning time	Sampling distribution of standard deviation

Any probability distribution can be described to some extent by its first two moments, mean and standard deviation.

Different possible ways of drawing samples:

Every enumerator assigned a sample size of 100 students to measure the height of students. The sample is drawn from good number of universities. suppose one enumerator is planning to draw 100 students from out of 1000 students. He can drawn it in ${}^{1000}C_{100}$ ways. Hence, each sample drawn fairly includes statistical random number of sample units.

Characteristics:

Mean $= \mu_{\bar{x}}$; Mean of sample - $\bar{x}$

S.D. $= \sigma_{\bar{x}}$; S.D. of sample - S

Let us analyse the observations of the above survey work. Suppose, some - how we are able to conduct a census survey, we may be able to get population disribution. The characteristic of such distribution of heights of university students can be denoted as μ (mu) and σ (sigma) for its mean and standard deviation respectively. Suppose if the enumerators are able to draw all posible samples with 100 students each from the total population, the mean and standard deviation of each one of such sampels can be denoted as $\bar{x}$ (x bar) and s (small s). Obviously individual sample means may not be nearer to population mean. They may tend to be nearer to population mean but rarely equals to it.

Let us construct a distribution with all sample means derived from each sample. Such distribution is called sampling distribution of means. It would have its own mean μ_x (mu x - bar) and its own standard deviation σ_x (sigma $\bar{x}$ - bar).

Sampling distribution and its characteristics

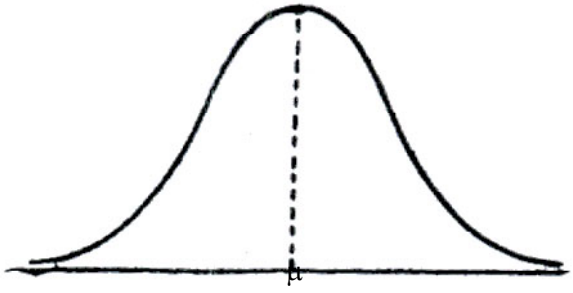
Sampling Distribution	Meaning of its characteristics
Sampling distribution of Mean	The mean and standard deviation of this sampling distribution explains the distribution of sample means
Sampling distribution of proportions	The mean and standard deviation of this sampling distribution explains the distribution of sample proportions

The differences in population sample and sampling distribution can be shown as in figure 14.1(a):

(a) Population Distribution:

This distribution is the distribution of all 1000 students. It has $\mu =$ the mean $\sigma =$ standard deviation.

Fig 14.1 (a)



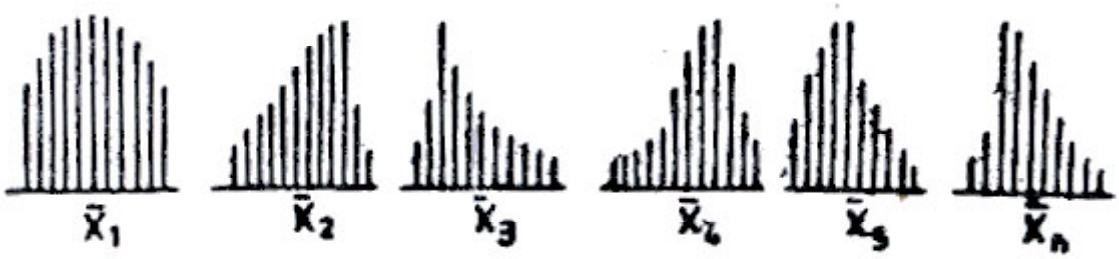
(b) Sample Distributions:

${}^{n}C_{r}$ number of samples could bedrawn from the population, each with a fixed sample size. Each sample distribution is discrete. Then

$\bar{x}$ = mean of each sample

s = sample standard deviation

Fig 14.1 (b)



(c) Sampling Distribution:

Some how if we can draw all possible samples from the population and the means of all sample distributions, when tabulated, we arrive at sampling distribution. It has:

μ_x = Mean of the sampling distribution

σ_x = Standard deviation of sampling distribution

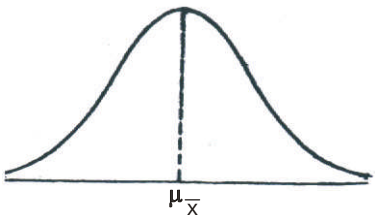


Fig 14.1 (c)

Construction of sampling distribution from means

Mean of sampling distribution and population distribution is one and the same:

It is belived that the sampling distribution in a way helps us to understand the population parameters as it includes all possible samples from such population. This advantage has made this concept very popular in statistical inferences.

14.3 Concept of Standard Error:

Standard deviation of sampling distribution:

The standard deviation of a sampling distribution is popularly called as "Standard Error". In order to avoid confusion or to indicate something else statisticians have coined this term. More specifically, the standard deviation of sampling distribution specially constructed using different sample statistics is denoted as standard error of that specific statistic. For example:

(i)	Standard deviation of the sampling distribution of Means	is called as	Standard error of mean Mean
(ii)	Standard deviation of the sampling distribution of sample proportions	is called as	Standard error of proportion
(iii)	Standard deviation of sampling distribution of sample variances.	is called as	Standard error of variances

Standard error and sampling error:

Suppose we have collected different samples from the same population, it is highly unlikely that all sample means would be the same. By any chance, if we know the population mean. the sample mean being exactly equal to population mean is also unlikely. The inter - sample variations in sample statistic resulting from 'sampling error', is purely due to chance. Such variaton results in standard deviation of sampling distribution of sample means. The standard error, thus, measures the extent to which the means from different samples vary because of chance factor in sampling process.

Statistical formulae to estimate standard error:

The sampling distribution is a theoretical distribution. The size, cost and time associated with census survey of population prohibit the analysis to consider all possible samples from a population. Hence, statisticians have developed formulae for estimating the characteristics of these theoretical distribution from a single sample and its statistics. The formulae differ on the basis of statistic considered in constructing the sampling distribution. For example, some of the formulae developed are given below:

Formulae for standard Error:

Formula for standard error for selected distributions

	Type of sampling distribution	**Formula for standard error**
(i)	Sampling distribution of sample means	$\frac{\sigma_p}{\sqrt{n}}$
	$\left(\begin{matrix}\sigma_p = \text{Population standard deviation } \mu \\ n = \text{sample size}\end{matrix}\right)$	
(ii)	Sampling distribution of sample medians	$\cdot 25331 \frac{\sigma_p}{\sqrt{n}}$

(iii) Sampling distribution of sample standard deviations $\frac{\sigma_p}{\sqrt{2n}}$

(iv) Sampling distribution of sample coefficient of correlation $\frac{1-r^2}{\sqrt{n}}$

(v) Sampling distribution of sample regression coefficient $\frac{\sigma_y\sqrt{1-r^2}}{\sigma_x\sqrt{n}}$

14.4 Sampling Distribution of Means:

The discussion on the concepts of population mean, sample mean and the variance in general. Sampling distribution of means are described when the population distribution is normal state the central limit, theorms and the relationship between sample size and standard error are discussed.

Example 1:

Let us consider the case of an investment planner of a mutual fund company who is interested in investing in corporate securities. The unit holders of the mutual fund company, naturally concerned with the rate of return being carned for the invested rupee. The greater the return, larger the attrac tion to the unit holders to further invest and derive the benefits. Keeping this objective, investment planner is interested in investing equally in the following stock.

Stocks	A	B	C	D	E
Return	12%	17%	- 8%	21%	15%

With only five stocks avarilable to invest, we can calculate the mean returns of the population.

Calculation of population mean:

The population mean $= \mu = \frac{\sum X_i}{n}$

where $X_i =$ individual return for the stock

$n =$ population size

Then $\mu = \frac{12\% + 17\% + (-8\%) + 21\% + 15\%}{5}$

$= 11\cdot 4\%$

Calculation of population standard deviation:

To understand the reliability of the average ler us calculate standard deviation as well

$$\sigma = \sqrt{\frac{\sum \left(X_i - \mu\right)^2}{N}}$$

Then,

$$\sigma = \sqrt{\frac{(120 - 11\cdot 4)^2 + (17 - 11\cdot 4)^2 + (-8 - 11\cdot 4)^2 + (21 - 11\cdot 4)^2 + (15 - 11\cdot 4)^2}{5}}$$

$$= 10\cdot 13\%$$

Suppose due to non - availability of information the investment planner is left with the data regarding only 3 scripts at a given point of timethen he has to design his investment plan on the basis of a sample of just 3 out of 5 at any time. Although the investment planner selects only one sample, he has a wide choice to consider any three scripts. He can choose ${}^{n}C_{r}$ or ${}^{5}C_{3}$ ways in selecting samples.

$${}^{n}C_{r} = {}^{5}C_{3} = \frac{5!}{3!(5-3)!} = 10 \text{ possible samples}$$

Possible ways of drawing a sample:

The mean and standard deviation of all 10 possible samples are worked out and given in the following table:

Possible samples and average returns

Sample Stocks	Returns	$\bar{X}$
Sample 1 ABC	12% , 17%, - 8%	7%
Sample 2 ABD	12%, 17%, 21%	16.67%
Sample 3 ABE	12%, 17%, 15%	14.67%
Sample 4 ACD	12%, - 8%, 21%	8.33%
Sample 5 ACE	12%, -8 %, 15%	6.33%
Sample 6 ADE	12%, 21%, 15%	16.0%
Sample 7 BCD	17%, - 8%, 21%	10.0%
Sample 8 BCE	17%, - 8%, 15%	8.0%
Sample 9 BDE	17%, 21%, 15%	17.67%
Sample 10 CDE	- 8%, 21%, 15%	9.33%

The mean of individual sample means can be found to range from 6.33% to 17.67%. If by chance, the investment planner chooses the third sample, the sampling error works out to be fairly smaller and the sample 5 or sample 9 are on the fartherest side on either way.

14.4.1 Sampling distribution of mean:

Estimation of sampling distribution:

We can now consider the mean of the sampling means to verify how far it is nearer to population mean. It can be worked out as average of sample means.

$$\mu_{\bar{x}} = \sum_{i=1}^{K} \bar{x}_i / K$$

where $\overline{X_i} = i^{th}$ sample mean

K = number of possible samples

Then $$\mu_x = \left[\begin{matrix} 7\% + 16 \cdot 67\% + 14 \cdot 67\% + 8 \cdot 33\% + 6 \cdot 33\% + \\ 16 \cdot 0\% + 10 \cdot 0\% + 8 \cdot 0\% + 17 \cdot 67\% + 9 \cdot 33\% \end{matrix} \right] \div 10$$

$$= 11 \cdot 4\%$$

It is exactly similar to population mean. This is always true and $\overline{X}$ of a sample is an unbiased estimator of population mean.

Calculation of Standard Error:

On the other hand, the relationship between population standard deviation and standard error of sampling distribution do not match so much. The distribution of sample means is less variable than the population from which samples are drawn. And hence the standard error would be smaller than population standard deviation. Let us calculate it:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{K} \left(\overline{X_i} - \mu_{\bar{x}} \right)^2}{K}}$$

$$= \sqrt{\frac{\left(7 - 11 \cdot 4^2 \right) + \left(16 \cdot 6^{-} - 11 \cdot 4 \right)^2 + \cdot \cdot \cdot + \left(9 \cdot 33 - 11 \cdot 4 \right)^2}{10}}$$

$$= 4 \cdot 13$$

The standard error indicates the spread in the distribution of all possible sample means due to sampling error.

14.4.2 Sampling From Normal Distribution:

If population is normal, sampling distribution is also normal:

Statisticians have proved that if a population is normally distributed with mean μ, and standard deviation σ, the sampling distribution of sample means is also normally distributed with a mean μ_x and standard deviation of $\sigma/\sqrt{n}$. It indicates that the mean of the sampling distribution is equal to population mean and the standard error equals to population standard deviation divided by samples size.

Example 2:

A large sized co-operative Departmental Stores calculates that the consumers account balances at the end of every month which are distributed with a mean of Rs. 600 and a standard deviation of Rs. 300 If a sales analyst takes a random sample of 100 accounts and finds the sampling distribution of mean values as normal with

$$\mu_x = \mu = \text{Rs}\cdot\ 600/- \quad \text{and} \quad \sigma_x = \text{S}\cdot\text{E}\cdot = \frac{\sigma}{\sqrt{x}} = \frac{300}{\sqrt{100}} = 30$$

Observe that the standard deviation of sampling distribution is far less than the standard deviation of population. If the sample size further increases, the standard error still decreases.

If the sales analyst is interested in finding the proportion of accounts with balance due larger than Rs. 650/- or above we can determine such probability by using standard normal probability distribution.

$$\boxed{Z = \frac{X - \mu}{\sigma_x}}$$

where X = Sample Value

μ = Population Mean

$\sigma_{\bar{x}}$ = Standard error of the mean = $\sigma/\sqrt{x}$

For X = 650

$$Z = \frac{\text{Rs. } 650 - 600}{\text{Rs. } 30}$$

$$= \frac{50}{30} = 1\cdot 67$$

Standard deviation from the mean of S.N.V. distribution.

Area under normal curve tables (See Appendix) give us an area of 0.4525 corresponding to Z value of 1.67.

Considering the upper side of the normal curve to represent the account balances of Rs. 650 and above, the area under the normal curve can be observed as 0.0475. It means, that the proportion of account balances that are above Rs. 650 works out to 4.75% of total accounts balances.

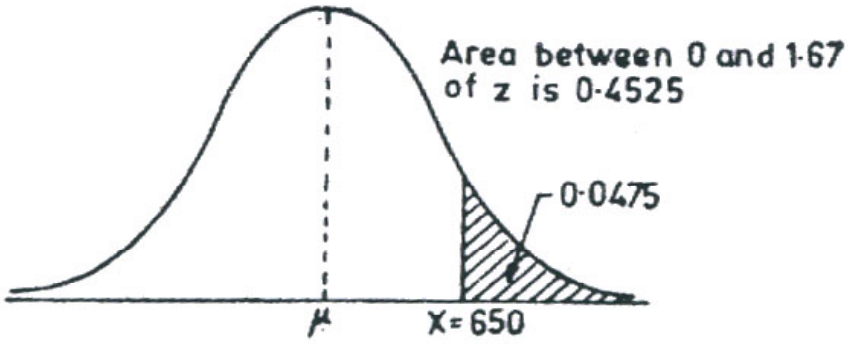


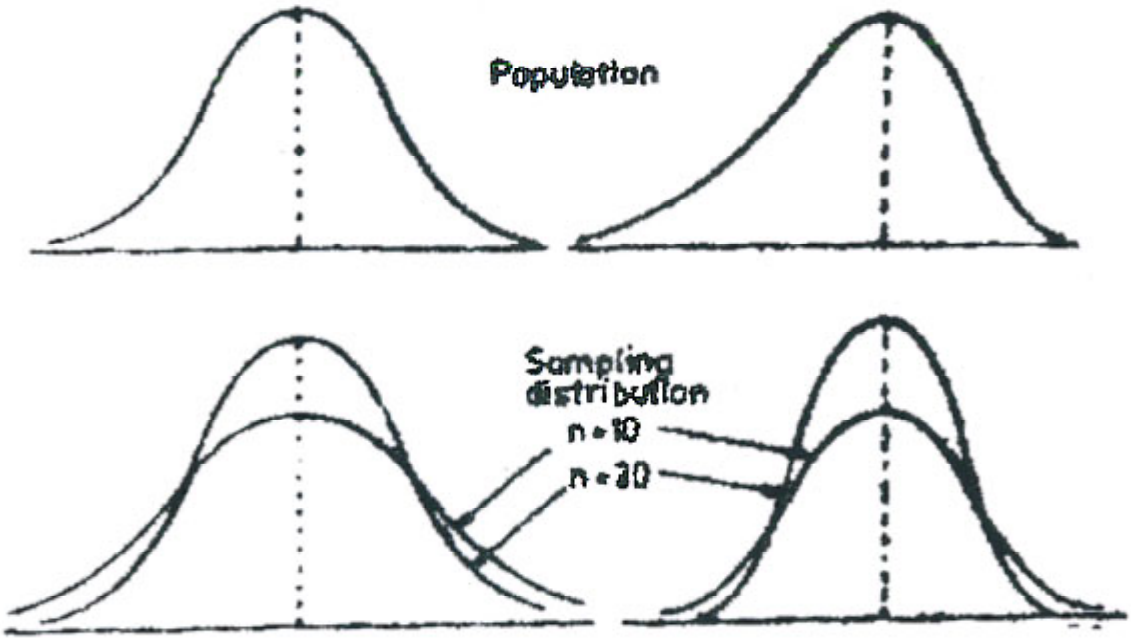
Fig 14.2

Probbaility of sample above Rs. 650

14.4.3 Sampling From Non - Normal Distributions:

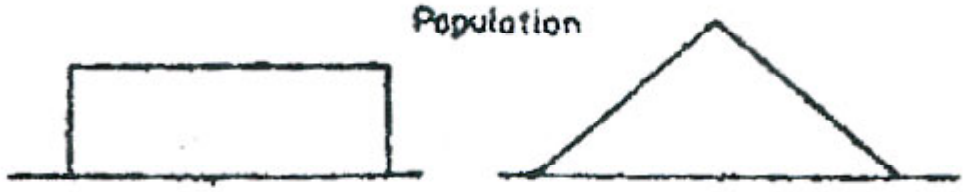
Central Limit Theorem:

When the population distribution is not normal, the distribution of sample mean is not normal. But one of the major findings of statisticians is that in most situations the distribution of sample means approximately works out to normal as 'n' gets large. A popular central limit theorem says



(a) Population is symmetric and normal.

(b) Population skewed.



(c) Uniform distribution.

(d) Triangle distribution.

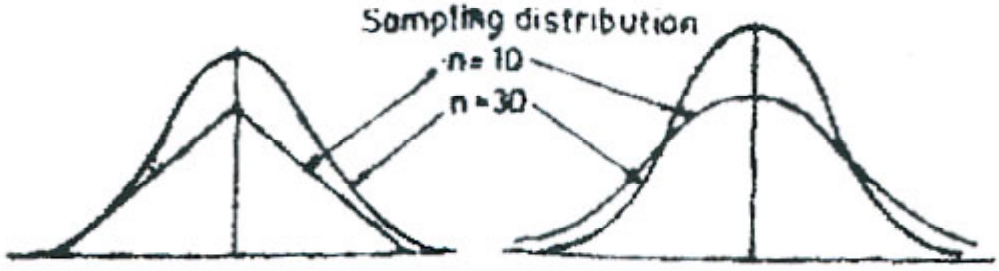


Fig 14.3 Few examples of non - normal populations and sampling distribution

14.5 Sampling distribution of the mean when population σ is unknown:

Substitute 'σ' with sample 'S':

The entire analysis described so far assumes that the analyst knows the population parameter, especially the population standard deviation (σ). But in reality the estimation of such a parameter, when N is large, is difficult and unimaginable. However, the statisticians observe that when 'n' is large this does not pose any problem even when σ is unknown, and it is reasonable, in such cases to substitute it with the sample standard deviation.

t-distribution when n is less than 30:

However, when it comes to the question of small sized sample of 'n', the sampling distribution is assumed to follow a new distribution called 't - distribution'. It also holds good only when the samples are drawn from a normal population.

t - distribution:

The variance of 't' distribution is 'υ':

The shape of t - distribution is similar to normal distribution. It is also symmetrical and bell shaped. Like standard normal distribution, the t distribution has the mean 'Zero', but its variance is not 1. Statisticians claim that its variance depends on a new parameter $\upsilon(\mathrm{nu})$ called degrees of freedom, which approaches to 1 when n is large. The 't' distribution is relatively flatter than normal distribution and the tails are longer than in normal distribution. Different 't' distributions emerge at different degrees of freedom. The degree of freedom is equal to $(n-1)$.

Different 't' distributions for different $(n\text{-}1)$ df:

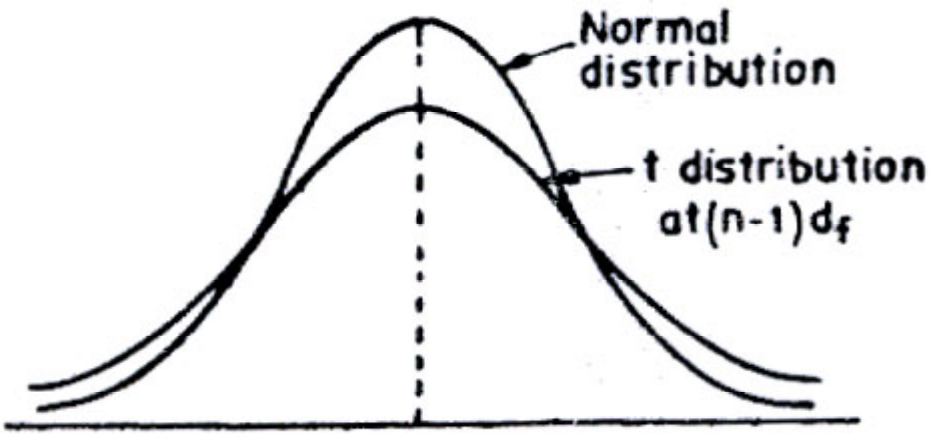


Fig 14.4 Normal and t distributions

If $\overline{X}$ is the mean of a random sample with size 'n' taken from a population which follows normal distribution, then

$$= \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

is the value of a random variable having the t - distribution with parameter $\upsilon = n-1$.

Example 3:

A manufacturer of small electric motors claims that, on an average the motors last 450 hours before their first, "break - down". A sample of 25 of these motors showed an average of 443.5 hours of use before a break - down and with a standard deviation of 24 hours. In order to verify calculate 't' statistic.

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$= \frac{443 \cdot 5 - 450}{24/\sqrt{25}} = -1 \cdot 35$$

It is a value having 't' distribution with $\upsilon = 25 - 1 = 24$ degrees of freedom. From tables, we can find for $\upsilon = 24$ the probability that 't' will exceed at 0.05 level is - 1.711 and hence we can accept the claim of the manufacturer here.

14.6 Sampling distribution of variance:

Squared value will always be positive:

So far we have seen the construction of sampling distribution for sample means. The sampling distributions can also be constructed for other statistics of the sample. Such a construction would be of help for making inferences about population. Let us see the characteristics of sampling distribution for variances of random sample from normal population. Since the variance is nothing but the square of standard deviation (S^2) and it cannot be a negative number, and the sampling distribution of such a variable can not be a normal curve. In fact it must be related to 'Gamma' distribution, also called chi-square distribution. As in the case of 't' distribution the chi - squre distributions also different at different degrees of freedom. The following figure shows the shape of chi - square distributions also different at different degrees of freedom. The following figure shows the shape of statistic distribution at different degrees of freedom.

If S^2 is the variance of a randomly drawn sample with size 'n' from a population normally distributed having a variance of σ^2 then

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

Chi-Square Distribution

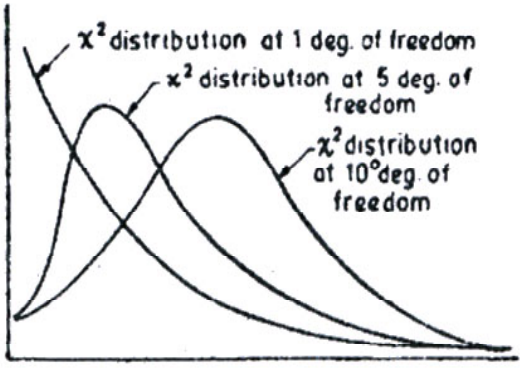


Fig - 14.5 χ^2 - Distribution at different degrees of freedom

is the value of a random variable having chi - square distribution with the parameter. Generally $\upsilon = n-1$.

F distribution:

Another issue relating to the distribution of variance is the distribution of ratio of variances in between two samples. This type of problem crops up when we are interested in determining whether the two samples have come from populations with equal variances. To determine whether the ratio of two sample variances is too small or too large, we refer to a variance ratio distribution known as F - distribution. In this distribution, now that two samples are involved, two degrees of freedom V_1 and V_2, one belonging to the numerator, and other belonging to the denominator exists. Graphically it can be marked as shown below.

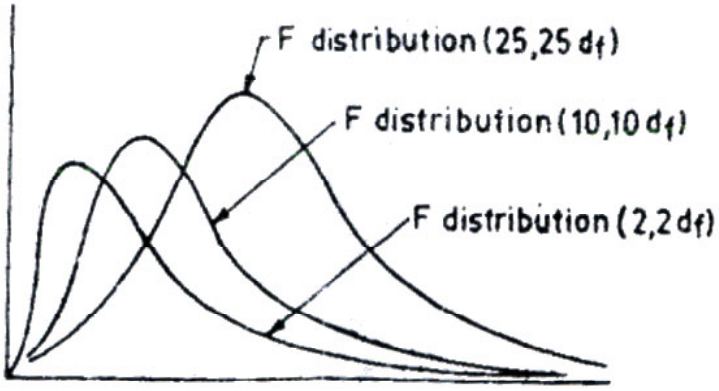


Fig - 14.6 F distribution with different degrees of freedom

14.7 Sampling distribution of proportions:

Sometimes a proportion of population possesses a particular attribute that is of interest to decision makers. A marketing executive may be interested is knowing which proportion of target group is interested in the company's product compared to a competing brand. Let us randomly select a customer. The probability that the customer prefers out brand compared to others say p.

Otherwises, q. In a large population, let 'r' be the individuals prefering our product out of 'n' sample subjects. The probability of 'r' success in such a case is assumed to follow binomial probability distribution viz.,

$$P(r) = {}^{n}C_{r}\ p^{r}\ a^{n-r}$$

14.8 Estimation:

Statistical Inferences:

The purpose of most statistical investigations is to generalise the findings obtained from random samples to the population from which the samples were obtained. Modern business decisions require us to make lot of estimations. For example, the market research heavily depends on statistical estimation. A market researcher selects a sample of customers, from that he estimates the proportion of customers in the entire market who prefer the company's product. How to use sample statistics to estimate population parameters?

Types of Estimates:

There are two types of estimates made about population: a point estimate and an interval estimate.

A point estimate is a single number that is used to estimate an unknown population parameter. An internal estimate is a range of values used to estimate a population parameter.

Example4:

To illustrate these concepts, let us consider case of the raw leather ordering mechanism of Ponds (India) Limited, leather division. Whenever the export order arrives for particular variety of shoe - uppoer the production executive makes a guess of the requirement of raw leather. While making such a guess, he considers the average size of shoe (although in reality it ranges from sizes of 5 to 11) and the number of 'pairs of order' to make. Here he considers one single average size of shoe as point estimate for the true average size of the entire lot of production schedule. He arrives at single average size of the shoe either by considering a sample of ordered variety or by experience. Similar point estimates become the bases for a cost accountant in ascertaining the average cost for different processes.

Interval estimate:

But most of the times, the estimates made as per the above procedure may not be equal to the population value. One probable limitation in a sample is the 'sampling error'. Hence point estimate may prove to be wrong. To overcome this problem, the most common procedure is to calculate an interval which is likely to contain the true population parameter. This estimation is also called Range Estimation.

Standard Error's Role in Estimation:

The standard error masures the dispersion of sample means in a sample distribution. It heklps in estimating the range. If we are clear about the confidence interval within which population parameter falls, it can be captured by arranging interval coefficient. An interval coefficient is the number of standard errors on either side of the population mean necessary to include a percentage of possible sample means equal to the confidence level. The interval estimation would be equal to:

Point estimate $\pm$ (interval coefficient) (Standard Error)

Confidence Level:

If the sample size is large, we can presume normal distribution and for different confidence levels and the interval estimation is:

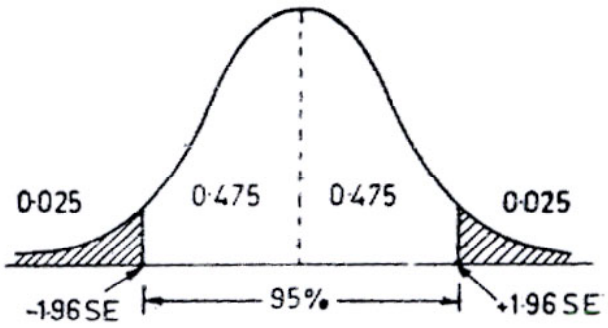
@ 95% confidence level $= \overline{X} \pm 1{\cdot}96\ \sigma \Big/ \sqrt{n}$

@ 99% confidence level $= \overline{X} \pm 2{\cdot}57 \Big/ \sqrt{n}$

Therefore, in motor cycle tyres example, the interval estimates at 95 percent confidence level are:

$\overline{X} \pm 1{\cdot}96\sigma_x$ or $\overline{X} \pm 1{\cdot}96\ \sigma \Big/ \sqrt{n}$

Suppose the advertising manager knows by experience that population standard deviation as 10 months then



Point Estimate $\pm\ 1{\cdot}96$ standard error

$$\overline{X} \pm 1{\cdot}96 \frac{10}{\sqrt{200}}$$

$$\overline{X} \pm 1{\cdot}39$$

Range 34.61 months to 37.38 months.

It is believed that the true population parameter is likely to fall within the interval.

Interval construction: Probability of population mean, falling within the interval:

The properties of standard error of the mean is of great importance in interval construction. Suppose we have considered s 2 SE level to construct an interval. It indicates a probability of 0.955 that mean of a sample will fall in $\pm\ 2$ standard errors of the population mean. In other words, 95.5 percent of all sample means are within plus or minus 2 standard errors from μ and in turn μ is within plus or minus 2 standard errors of 95.5 percent of all sample means. Suppose we have taken 100 samples from a population, 95.5 percent samples would contain population mean. The concept is presented graphically here under:

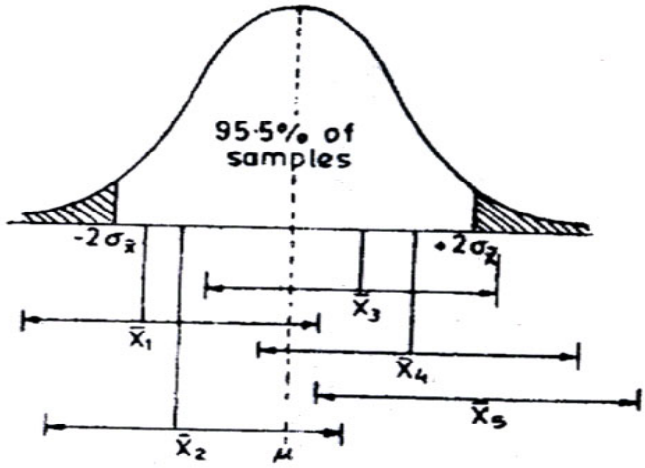


Fig 14.7 Intervals of sample means

In the above figure, one can observe that most sample intervals include population mean. The only sample X_5 does not contain it. It can be observed such samples are falling in either of the tails of the sampling distribution.

14.8.1 Estimation of Intervals For Large Sample Means:

Large Sample Follow 'Z' Distribution:

The advertising manager of Bajaj Auto Ltd. has selected a random sample of 500 people who have purchased their 'Chetak' model 150 cc scooter six months back. He is interested in estimating the average mileage per litre of petrol for six month-old vehicles. His sample showed a mean of 45 km pl. and a standard deviation of 15 km pl. Let us determine his estimation about the population with 95 percent confidence.

Sample size = 500

Sample Mean = 45 km pl.

Sample standard deviation = 15 km pl.

In most business applications, the population mean and standard deviations are very difficult to calculate or to determine. When σ of population is unknown, it has to be estimated. It is generally estimated to be equal to sample standard deviation 's'. When sample size is large, the confidence interval estimate of the population mean is

$$\bar{X} \pm 1{\cdot}96\, s \Big/ \sqrt{n}$$

In the above illustration, $\bar{X} = 45$ km pl. s = 15 km. pl at 95 percent confidence, the interval estimate is

$$45 \pm 1{\cdot}96 \left(15 \Big/ \sqrt{500}\right)$$

$$45 \pm 1{\cdot}31$$

Range $43 \cdot 685 - 46 \cdot 31$ km pl

For constructing his advertisement solgan, the advertising manager can confidently take the average mileage of Chetak model scooter to be between 43.685 km pl. to 46.31 km pl.

EStimation of Interval for Proportion:

In the earlier part of this lesson, we have observed that the sampling proportions follow binomial distribution of

$$P(r) = {}^{n}C_{r}\ p^{r}\ q^{n-r}$$

where n = Identical trials, each one results in two possible outcomes.

p = probability of success.

q = probability of otherwise

The average of such a binomial distribution $\mu = np$ and the standard deviation $\sigma = \sqrt{npq}$. The standard error of population is given as

$$\sigma_p = \sqrt{pq/n}$$

When population proportion is not known, and sample size is large, the sampling distribution of proportions can be approximated by normal distribution. Then the sample proportions can be substituted to population. Then the sample proportions can be substituted to population proportion and the standard error can be approximated to

$$S_p = \sqrt{\hat{p}\hat{q}/n}$$

where $\hat{p}$ is sample proportion. The confidence interval for population proportion will be:

$$\hat{p} \pm Zs_p$$

or	Point Estimate of the sample	±	Internal coeffcient	×	Standard Error of the sampling distribution

Example 5:

One of the factories of Ponds Leather Division is producing 50,000 pairs of shoe - uppers daily. From a sample of 500 pairs, it is observed by the quality control executive that 2% were of substandard quality. The executive is interested in knowing the number of pairs that could be reasonably expected to have spoiled in the daily production assigning 95 percent confidence level.

Let $\hat{p}$ (be) the proportion of substandard quality shoes in the sample.

$$\hat{p} = 2/100 = 0 \cdot 02$$

Then standard error $= \sqrt{\frac{pq}{n}} = \sqrt{\frac{0 \cdot 02 \times 0 \cdot 98}{500}}$

$$= 0 \cdot 0063$$

At 95% confidence level, interval of population proportion would be

$$p \pm 1 \cdot 96 \sqrt{pq/n}$$

$$0 \cdot 02 \pm 1 \cdot 96 \times 0 \cdot 0063$$

Range 0.0077 - 0.0323 proportion.

The number of pairs expected to be spoiled in a daily production would be $0 \cdot 0077 \times 50,000$ and $0 \cdot 0323 \times 50,000$ i.e., between 385 to 1615 pairs of shoe - upper.

14.8.2 Estimation of intervals for small samples:

Whenever the sample size is less than 30 and population standard deviation is unknown, it is customary to use 't' distribution. However, in using 't' distribution we assume that the population is normal or approximately normal.

Small samples follow t - distribution:

For illustration, a Marketing Manager is interested in knowing the annual demand for the brand of Talcom Powder Tins being produced by their entrerprise. Towards this purpose, he has drawn a random sample of 16 house - wives and enquired about the number of Tins of Talcom Powder that their families use in an year. The answers are as follows:

5	3	7	11	12	6	3	2
10	10	8	3	5	9	6	4

Let us calculate the mean and standard deviation of the sample

$$\bar{X} = 6 \cdot 5 \text{ tins} \qquad s = 3 \cdot 08 \text{ tins}$$

The intervals can be estimated as follows:

At 90% confidence level $t_{0 \cdot 05}$ for 15 d.f. = 1.753

At 95% confidence level $t_{0 \cdot 025}$ for 15 d.f. = 2.131

At 90% confidence level : $\bar{x} \pm t_{0 \cdot 05}$ for $15 \text{ d} \cdot \text{f}. \left(s / \sqrt{n}\right)$

$$: \quad 6\cdot5 \pm 1\cdot753\left(3\cdot08/\sqrt{16}\right)$$

$$: \quad 6\cdot5 \pm 1\cdot34981$$

Range: 5.15 tins - 7.85 tins

At 95% confidence level : $\bar{x} \pm t_{0\cdot025}$ for 15 d.f. $\left(S/\sqrt{n}\right)$

$$: \quad 6\cdot5 \pm 2\cdot131\left(3\cdot08/\sqrt{16}\right)$$

$$: \quad 6\cdot5 \pm 1\cdot64087$$

Range : 4.86 tins 8.15 tins

Estimation of interval for the difference between two populations:

Managers are often interested in ascertaining the differences in alternative courses of action. A training director is interested in knowing the differences in productivity of employees under different training packages. A mrketing manager is interested in finding the differences in two different brands being produced by his enterprise and that of his competitors. Electric bulbs being manufactured by two companies is as follows:

	X Co.	Y Co.
No. of Bulbs Sampled	100	100
Mean life in hours	1250	1200
Standard deviation of life in hours	100	75

Let us calculate the confidence range of differences between two competitors.

Statistic:

The point estimate for the difference between population means $\left(\mu_{x1} - \mu_{x2}\right)$ is $\left(\overline{x_1} - \overline{x_2}\right)$ is the difference in sample means. The sampling distribution of the difference $\left(\overline{x_1} - \overline{x_2}\right)$ will be normally distributed with mean.

$$\mu_{\overline{x_1} - \overline{x_2}} = \mu_{x_1} - \mu_{x_2}$$

The standard error of the difference between two sample means is

$$\sigma_{x_1 - x_2} = \sqrt{\frac{\sigma_{x_1}^2}{n_1} + \frac{\sigma_{x_2}^2}{n_2}}$$

In the given problem, the confidence interval would be:

point estimate $\pm$ (interval coefficient) (standard error of the estimate)

$$\left(\overline{x_1} - \overline{x_2}\right) \pm z\left(\sqrt{\frac{\sigma_{x_1}^2}{n_1} + \frac{\sigma_{x_2}^2}{n_2}}\right)$$

where, z = interval coefficient

@ 95% confidence = $\pm 1{\cdot}96$

@ 99% confidence = $\pm 2{\cdot}57$

$$\left[1250 - 1200\right] \pm 1{\cdot}96\sqrt{\frac{100^2}{100} + \frac{75^2}{100}}$$

$$50 \pm 24{\cdot}5$$

Range: Lower confidence limit = 50 - 24.5 = 25.5 hrs.

Upper confidence limit = 50 + 24.5 = 74.5 hrs.

Therefore the range of difference between the competing products is 25.5 hours to 74.5 hours in terms of burning life of the bulbs being produced by them

14.8.3 Difference between two population proportions:

Example 6:

The marketing executives, generally, when they want to assess their market share, are interested in estimating the difference between two population proportions of those who use their brand and others. For example, the Nirma Industries Ltd. has carried out an intensive advertising campaign to encourge housewives belonging to middle income groups to use their new detergent powder. Subsequently a sample investigation is carried out in two towns of A and B to determine the proportion of house wives who now use their detergent poweder. In town A 500 housewives are interviewed and found 225 are using the new detergent poweder. In town B 150 out of 300 families interviewed are found using new detergent poweder. If town A and town B really differ in respect of population distribution, how far the sample study would be helpful to the marketing manager to go about similar estimates of differences in similar types of towns.

For this intervals of difference in population proportion $\left(p_1 - p_2\right)$ has to be estimated. The difference, as a variable, is expected to generate a 'sampling distribution of difference in proportions with a standard deviation of

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

where p, q are population proportions.

when the sizes of samples are sufficiently large, the standard error of sampling distribution can be estimated with substituted values, as follows:

$$S\hat{p}_1 - \hat{p}_2 = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{p}\,\hat{q}$ are sample proportions

Then confidence interval would be

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z\sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}$$

(or) $\text{Interval Estimate} \pm \begin{bmatrix}\text{Confidence}\\ \text{Coefficient}\end{bmatrix} \quad \begin{bmatrix}\text{Standard}\\ \text{Error}\end{bmatrix}$

For Nirma products problem, the interval at 95% level of confidence is

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm 1{\cdot}96\sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\left(0{\cdot}45 - 0{\cdot}50\right) \pm 1{\cdot}96\sqrt{\frac{0{\cdot}45 \times 0{\cdot}55}{500} + \frac{0{\cdot}50 \times 0{\cdot}50}{300}}$$

Range: Upper confidence level: 0.2143 or 2.14%

Lower confidence level : - 0.1214 or - 12.14%

Hence the marketing executive can be confident about the difference in proportions of population using their brand between towns certainly lie between + 2.14% to - 12.14%. This interval certainly contains the population difference between towns.

14.9 Summary:

We have discussed the sampling distribution of some commonly used statistics. A sampling distribution of a sample statistics has been introduced as the probability distribution the probability density function of the sample statistics. We have presented central limit theorem and its important. We discussed chi - Square distribution, t distribution and F - distribution. We also explained the point and interval estimations.

14.10 Exercises:

1. Explain the sampling distribution of a statistic error and standard error.
2. What do you mean by 'estimation'. How does point estimation differ from interval estimation?
3. How do you test the significance of the difference between the means of two samples.

4. Explain how chi - square and F distributions would be helpful in estimating the sampling distribution of sample variances.

5. A.P. co - optex extends rebate sales and sales on credit to gazetted and non - gazetted employees in their state during festive occassions. The credit extended to customers would be collected in the form of monthly instalments directly. The sales executive who is interested in the outstanding balances of accounts receivable has checked 100 accounts out of 1000 accounts. He has observed that the average balance in customer's accounts is Rs. 450, with a standard deviation of Rs. 150. He wants to estimate the average balance for the entire population with 95 percent confidence limit.

6. Bharat Electronics Limited wishes to manufacture RAM chips for computers by replacing currently used Batch processing to Automatic on - line processing. However, the production engineer has made a time estimate between two processings.

	Mean	Standard Deviation
Batch processing (no. of chips produced in a week)	200	40
Semi - Automatic assembly line	250	120

Find, whether it is advisable to shift the assembly line of processing considering 95% confidence interval.

7. A consumer appliance store decides to launch heavy advertisement by making one - real film to be presented in all local cinema theatres by replacing their presently used slide eprojections. The sales for each of the 10 days after the cinema projections and started are compared with sales of 10 days before the start of cinema commercial. Test the improvement in sales during the lattter period at 5% level of significance.

Sales before the commercial		Sales after the commercial	
Rs. 6,000	20,000	Rs. 8,000	10,000
" 8,000	14,000	" 8,000	16,000
" 4,000	5,000	" 5,000	14,000
" 10,000	8,000	" 10,000	12,000
" 8,000	12,000	" 12,000	15,000

14.11 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics for Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. K. MADHU BABU

Lesson - 15

TESTING OF HYPOTHESIS

Objectives:

After reading this lesson you should be able to

- To understand the concept of Hypothesis and learn the procedure associated with its testing.
- To appreciate the importance of confidence intervals and significance levels.
- To perform tests concerning the population mean, proportion and difference between two population.

Structure:

15.1 Introduction:

Generally a hypothesis is established beforehand. By hypothesis we mean to give postulated or stipulated value(s) of a parameter Also, instead of giving values, some relationship between parameters is postulated in the case of two or more populations. On the basis of observational data, a text is performed to decide whether the postalated hypothesis be accepted or not. This involves certain amount of risk. Theis amount of risk is teremed as a level of significance. when the hypothesis is accepted, we consider it a non-significant result and if the reverse situation occurs, it is called a significant results. The test, which are dealt with in this chapter pertain to parametric tests. A test is defined as, "A statistical test in a procedure governed by certain rules,

which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample values.

Statistical tests of hypothesis play an important role in industry biological science, behavioral science and economics, etc... The use of tests has been made clear through a number of practical problems.

1. A feed manufacturer announces that in feel contaxt 40% protien. Now to make sure whether his clain in correct or not, one has to take a random sample of the product and by chemical analysis. Find the protien percentages in the samples. From these observed values, the would decide about the manufacturer's clain for this product. This is deve by performing a test of significant.

2. There is a provided A which produces contain items. It is considered that a new process B is better than process A. Both the processes are put under operation and then the items produced by them are sampled and observations and taken as them. A statistical test is performed based on these observations which enables v to decide whether process B is better than A or not.

3. Often we are inteverted to know what is the best does of a chemical treatment? Two or more doses of the chemical are applied or administered on a numsev of subjects and response is onserved. Now it is tested statistically whether the doses differ significantly or not.

There is not end to such type of practical problems where statistical tests can be applied. These are only a few examples. Have, one important point is to be noted. What ever conclusion are drawn about the population, they are always subjected to some error. Hence there is always some risk involved in these decissions. This, a level of significance is always associated with these decisions. Now we will discuss various terms involved in testing of hypothesis in an exoct way before describing the statistical tests.

Two - action decision problem: A test of a statistical hypothesis is a two - action decision problem after the experimental sample values have been obtained, the two - action being the acceptance or rejection of the hypothesis under consideration.

What is hypothesis: A hypothesis is an assertion or conjeture about the parameter(s) of population distribution(s).

Some basic concepts:

1. Null Hypothesis: Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

2. Type - I and Type - II Errors: After applying a test, a decision is taken about the acception or rejection of null hypothesis vis - a - vis the alternative hypothesis. There is always some possibility of committing an error in taking a decision about the hypothesis. There errors can be two types.

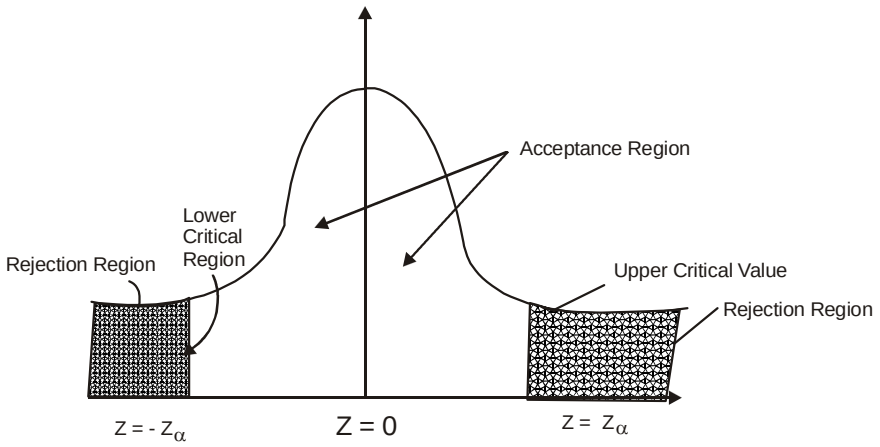
Type I Error: Reject null hypothesis (H_0) when it is true.

Type II Error: Accept null hypothesis (H_0) when it is false.

These two types of error can be better understood with an example where a patient is given a inedicine to curve some disease and his condition is scrutinised for some time. It is just possible that the medicine has a positive effect but it considered that it has no effect or adverse effect. Thuse, it is the first kind of error or type I error. On the contrary, if the medicine has an adverse effect but in considered to have a positive effect, it is called the second kind of error or type - II error.

Level of Significance: It is the quantity of risk of type - I error which we are ready to tolerate in making a decision about H_0. In other words, it is the probability of type - I error which is tolerable. The level of significance is denoted by α and is conventionally chosen as 0.05 or 0.01. Level $\alpha = 0 \cdot 01$ is used for high precision and $\alpha = 0 \cdot 05$ for moderate precision.

Two - tailed test at level of significance $'\alpha'$:

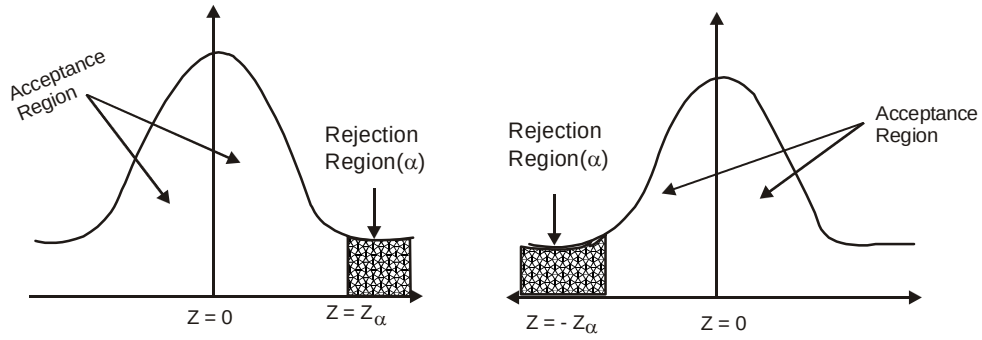


In the case of one-tailed alternative,

$P(Z > Z_\alpha) = \alpha$ if it is one-tailed (right)

$P(Z < -Z_\alpha) = \alpha$ if it is one-tailed (left)

For level of significance 'α'



Right tailed test **Left tailed test**

Power of the test: The size of a best is the probability of rejecting the null hypothesis when it is true, and is usually denoted by α. The level of significance and size are synonymous in a practical sesen. Therefore

$$P\left(\text{ reject } H_0 / H_0 \right) = \alpha$$

The power of a test is defined as a probability of rejecting the null hypothesis when it is actually false, i.e., when H_1 is true. In short

$$\begin{aligned} \text{Power} &= P\left(\text{reject } H_0 / H_1 \right) \\ &= 1 - P\left(\text{accept } H_0 / H_1 \right) \\ &= 1 - \left(\text{Prob. of type II error} \right) \\ &= 1 - \beta \end{aligned}$$

where β is the probability of type II error. Among a class of tests, the best test is the one which has the maximum power for the same size.

P - value of the test: Another approach is to find out the p - value at which H_0 is significant, i.e. to find the smallest level α at which H_0 is rejected. In this situation, it is not inferred whether H_0 is accepted or reject at level 0.05 or 0.01 or any other level. But the statistician only gives the smallest level α at

Hypothesis Testing Procedure:

1. Null hypothesis H_0 is defined

2. Alternative hypothesis H_1 is also defined after a careful study of the problem and also the nature of the test (whether one - tailed or two - tailed) is decided.

3. Level of signification α is fixed or taken from the problem if specified and Z_α is noted.

4. The test - statistics $Z = \frac{t - E(t)}{S \cdot E \cdot (t)}$ is computed

5. Comparison is mode between $|Z|$ and Z_α. If $|Z| < Z_\alpha$, H_0 is accepted or H_1 is rejected, i.e. it is concluded that the difference between t and $E(t)$ is not significant at α % los.

   On the otherhand, If $|Z| > Z_\alpha$, H_0 is rejected or H_1 is accepted, i.e. it is conclude that the diffference between t and $E(t)$ is significant at α% los.

15.4 Test of significance for single mean:

We have proved that if $x_i, i = 1, 2, \cdots, n$ is a random sample of size n from a normal pop with mean μ variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n i.e., $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$ However, this results holds i.e., $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ even in random sampling from non - normal population provided the sample size n is large.

Thus for large sample, the standard normal variate corresponding to $\bar{x}$ is:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

under the null hypothesis, H_0 that the sample has been drawn from a population with mean μ and variance σ^2 i.e., there is no significant difference between the sample mean $(\bar{x})$ and population mean (μ) the test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

If the population S.D. σ is unknown then we use its estimate provided by the sample variance is given by

$$\hat{\sigma}^2 = S^2 \Rightarrow \hat{\sigma} = S \text{ (for large sample)}$$

$$Z = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

If $|Z| < Z_{\alpha}\ 5\%$ we accept the null hypothesis H_0 otherwise we reject H_0 at 5% los.

Example 1:

A sample of 900 members has a mean 3.4 cms. and S.d 2.61 cms. Is the sample from a large population of mean 3.25 cms and s.d 2.61 cms?

Solution:

Null hypothesis H_0 - The sample has been drawn from the population with mean $\mu = 3 \cdot 25$ cms and $S \cdot D \cdot \sigma = 2 \cdot 61$ cms, alternative hypothesis $H_1 : \mu \neq 3 \cdot 25$ (two - tailed)

Test statistic under H_0, the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Here we are given that

$\bar{x} = 3 \cdot 4$ cms, $n = 900$ cms, $\mu = 3 \cdot 25$ cms and $\sigma = 2 \cdot 61$ cms

$$Z = \frac{3 \cdot 40 - 3 \cdot 25}{2 \cdot 61/\sqrt{900}}$$

$$Z = 1 \cdot 73$$

Since $|Z| < 1 \cdot 96$ we conclud that data don't provide vs and avidence against the null hypothesis (H_0) which may there be accepted at 5% los.

Which H_0 is rejected. This facilitates an individual to decide for himself as to how much significant the data are. The approach avoids the imposition of a fixed level of significance. About the acceptation or rejection of H_0, the experimenter can himself decide the level α by comparing it with the - p value. The criterian for this that if the P - value is less than or equal to α, reject H_0 otherwise accept H_0.

15.5 Test of significance for difference of mean:

1. When population variance is known: Let $\overline{x_1}$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\overline{x_2}$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and varience σ_2^2. Then, since sample sizes are large

$$\overline{x_1} \sim N\left(\mu_1, \sigma_1^2/n_1\right) \quad \text{and} \quad \overline{x_2} \sim N\left(\mu_2, \sigma_2^2/n_2\right)$$

Also $\overline{x_1} - \overline{x_2}$ being the difference of two independent normal variates is also normal variate. The $Z\ (S \cdot N \cdot V \cdot)$ corresponding to $\overline{x_1} - \overline{x_2}$ is given by

$$Z = \frac{\left(\overline{x_1} - \overline{x_2}\right) - E\left(\overline{x_1} - \overline{x_2}\right)}{S \cdot E \cdot \left(\overline{x_1} - \overline{x_2}\right)} \sim N(0, 1)$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. there is no significant difference between the sample means, we get

$$E\left(\overline{x_1} - \overline{x_2}\right) = E\left(\overline{x_1}\right) - E\left(\overline{x_2}\right) = \mu_1 - \mu_2 = 0$$

$$V\left(\overline{x_1} - \overline{x_2}\right) = V\left(\overline{x_1}\right) + V\left(\overline{x_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

the covariance term vanishes, since the sample means $\overline{x_1}$ and $\overline{x_2}$ are independent.

Thus under $H_0 : \mu_1 = \mu_2$ the test statistic becomes (for large samples)

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e., if the sample have been drawn from the population with common S.D. σ, then under $H_0 : \mu_1 = \mu_2$

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

If $|Z| \leq Z_{\alpha}$ 5%, we accept H_0, otherwise we reject H_0 at 5% level of significance.

Example 2:

The mean of two single large samples of 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches? (TEst at 5% level of significance)

Solution:

We are given

$n_1 = 1000,\ n_2 = 2000,\ \overline{x_1} = 67 \cdot 5$ inches, $\overline{x_2} = 68 \cdot 0$ inches Null hypothesis, $H_0 : \mu_1 = \mu_2$ and $\sigma = 2 \cdot 5$ inches i.e. the samples have been drawn from the same POP. of S.D. = 2.5 inches.

Alternative Hypothesis : $H_1 : \mu_1 \neq \mu_2$ (Two tailed)

Test statistic under H_0 the test statistic

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

$$Z = \frac{67 \cdot 5 - 68 \cdot 0}{\sqrt{(2 \cdot 5)^2\left(\frac{1}{1000} + \frac{1}{2000}\right)}}$$

$$Z = -5 \cdot 1$$

$$|Z| = 5 \cdot 1$$

Conclusion:

Since $|Z| > 3$, the value is high significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with S.D. 2.5

15.6 The population variance is unknown:

If $\sigma_1^2 \neq \sigma_2^2$ and σ_1^2 and σ_2^2 are not known, then they are estimated from sample values. This result in some error, which is practically immaterial, if samples are large. These estimates for large samples are given by

$$\hat{\sigma}_1^2 = S_1^2 \simeq s^2$$

$$\hat{\sigma}_2^2 = S_2^2 \simeq s_2^2$$

Then the test statistic be comes

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim N(0, 1)$$

Using normal test, we compute the value of Z.

Example 3:

The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average wage of a sample of 200 workers in plant B was Rs. 2.87 with a standard deviation of Rs. 1.28 can an applicant safely assume that the hourly wage paid by plant B are higher than those paid by plant A?

Solution:

Let X_1 and X_2 denote the hourly wages (in Rs.) of workers in plant A and plant B respectively.

Then we are given

$$n_1 = 150, \ \overline{x_1} = 2 \cdot 56, \ s_1 = 1 \cdot 08 \cdot$$

$$n_2 = 200, \ \overline{x_2} = 2 \cdot 87, \ s_2 = 1 \cdot 28 \cdot$$

Null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. there is no significant difference between the mean level of wages of workers in plant A and plant B.

Test statistic under H_0 the test statistic is:

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$Z = \frac{2 \cdot 56 - 2 \cdot 87}{\sqrt{\frac{(1 \cdot 08)^2}{150} + \frac{(1 \cdot 28)^2}{200}}}$$

$$Z = -2 \cdot 46$$

$$|Z| = 2 \cdot 46$$

Conclusion:

Since calculated value of Z is less than critical value, it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by palnt B are certainly higher than those paid by plant 'A'.

1. One - tailed and two - tailed tests:

If θ_0 is a population parameter and θ is the corresponding statistic and if we set up the null hypothesis $H_0 : \theta = \theta_0$, then the alternative hypothesis which is complementary to H_0 can be any one of the following:

(i) $H_1 : \theta \neq \theta_0$, $i \cdot e \cdot \theta > \theta_0$ or $\theta < \theta_0$

(ii) $H_1 : \theta > \theta_0$

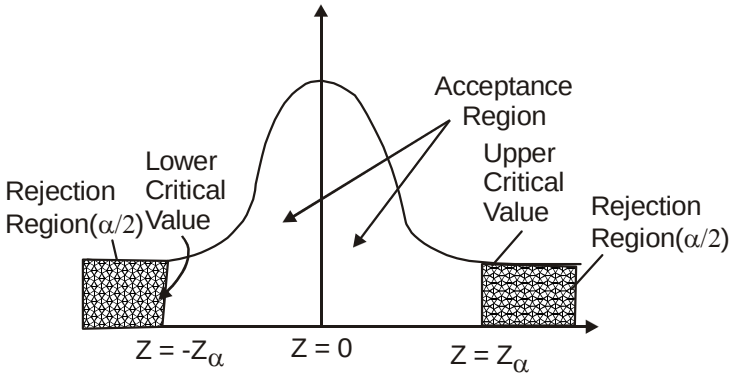
(iii) $H_1 : \theta < \theta_0$

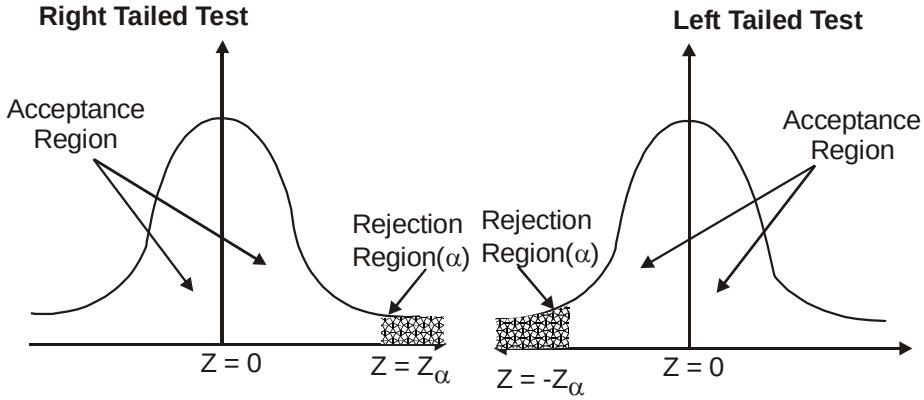
H_1 given in (i) is called a two - tailed alternative hypothesis, where as H_1 given in (ii) is called right tailed alternative hypothesis and H_1 given in (iii) is called a left tailed alternative hypothesis.

When H_0 is tested while H_1 is a one tailed alternative (right or left), the test of hypothesis is called a one - tailed test.

When H_0 is tested while H_1 is a two - tailed alternative, the test of hypothesis is called a two - tailed test.

The application of one - tailed or two - tailed test depends upon the nature of the alternative hypothesis. The choice of the appropriate alternative hypothesis depends on the situation and the nature of the problem concerned.





15.7 Testing of population proportions:

1. Test for single proportion:

If x is the number of success in n independent trails with constant probability P of success for each trail

$$E(x) = nP \quad , V(P) = nPQ$$

where $Q = 1 - P$ is the probability of failure. It has been proved that for large the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$ i.e.,

$$Z = \frac{x - E(x)}{\sqrt{V(x)}} = \frac{x - nP}{\sqrt{nPQ}} \sim N(0,1)$$

and we can apply the normal test.

Remarks:

In a sampel of size n, let x be the number of persons possessing a given attribute then observed proportion of success $= x/n = P$.

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(x) = \frac{1}{\cancel{n}} \cdot \cancel{n} P = P$$

$$\therefore E(p) = P, \quad V(p) = V\left(\frac{x}{n}\right) = \frac{1}{n^2} V(x) = \frac{1}{n^x} n P Q = \frac{PQ}{n}$$

Since x and consequently x/n is asymptotically normal for large n, the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{S \cdot E(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0, 1)$$

If $|Z| < 1 \cdot 96$ we do not reject H_0 otherwise we reject H_0 at 5% los.

Example 4:

A dice n thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the dice can not be regarded as an unbaised one.

Solution:

n = 9,000 ; x = number of successes = 3,240

Under the null hypothesis $n = H_0$: that the die is unbiased one

P = probability of success = 3 or 4 $= \frac{1}{6} + \frac{1}{6} = \frac{1}{\cancel{3}}$

Alternative hypothesis $H_1 : P \neq \frac{1}{\cancel{3}}$

We have $Z = \frac{X - nP}{\sqrt{n P Q}} \sim N(0, 1)$

$$Z = \frac{3240 - 9000 \times \frac{1}{3}}{\sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}}} = 5 \cdot 36$$

$$|Z| = 5 \cdot 36$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the dice is almost certainly biased.

Example 5:

In a sample of 1000 people in Maharastra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% los.

Solution:

x = number of rice eaters = 540

p = sample proportion of rice eaters $= \frac{x}{n} = \frac{540}{1000} = 0 \cdot 54$

Null hypothesis H_0 both rice and wheat are equally popular in the state so that

p = population proportion of rice eaters in Maharastra = 0.5

Q = 1 - P = 0.5

Test statistic under H_0, the test statistic $Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$

Now $Z = \frac{p - P}{\sqrt{0 \cdot 5\, x^{0 \cdot 5} / 1000}} = 2 \cdot 532$

Conclusion:

Conclude that rice and wheat are equally popular in maharastra state. The significant values of Z at 1% los for two - tailed test is 2.58. Since computed Z = 2.532 is less than 2.58 it is not significance at 1% los hence the null hypothesis is accepted.

15.8 Test of significance for difference of proportions:

Suppose we want to compare two district populations with respect to the prevalence of cetain attribute say A, among their memebers. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by

$$p_1 = \frac{x_1}{n_1} \text{ and } p_2 = \frac{x_2}{n_2}$$

If P_1 and P_2 are the population proportions, then

$$E(p_1) = P_1 \text{ , } E(p_2) = P_2 \cdot$$

and $v(p_1) = \frac{P_1 Q_1}{n_1}$ and $v(p_2) = \frac{P_2 Q_2}{n_2}$

since for large samples p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the null hypothesis $H_0 : P_1 = P_2$ i.e., there is no significance difference between the sample propertion, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0$$

Also $V(p_1 - p_2) = V(p_1) + V(p_2)$

the covariance term $\text{cov}(p_1, p_2)$ varishes, since sample proportions are independent

$$\therefore V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Since under $H_0 : P_1 = P_2 = P(\text{say})$

Hence under $H_0 : P_1 = P_2$ the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

In general, we do not have any information as to the proportion of A's in the proportions from which the samples have been taken. Under $H_0 : p_1 = p_2 = p(\text{say})$ an unbiased estimate of the population proportion P, based on both the samples is given by

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{Q} = 1 - \hat{P}$$

In this case the test statistic becomes

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

If $|Z| < 1{\cdot}96$, we donot reject H_0 otherwise we reject H_0 at α % los.

Example 6:

A random sample of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not at 5% level.

Solution:

Null hypothesis $H_0 : P_1 = P_2 = P$ (say) i.e., there is no significant difference between the opinion of men and women as far as proposal of flyover is concerned we are given that

$n_1 = 400$; $x_1 =$ number of men favouring the proposal = 200

$n_2 = 600$, $x_2 =$ number of women favoring the proposal = 325

$p_1 =$ proportion of women favoring the proposal in the sample $= \frac{x_1}{n_1} = \frac{200}{400} = 0{\cdot}5$

$p_2 =$ proportion of women favoring the proposal in the sample $= \frac{x_2}{n_2} = \frac{325}{600} = 0{\cdot}541$

Test statistic since samples are large , the test statistic under the null hypothesis H_0 is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = 0{\cdot}525$

$$\hat{Q} = 1 - \hat{P} = 1 - 0 \cdot 525$$

$$\hat{Q} = 0 \cdot 475$$

$$\therefore \; Z = \frac{0 \cdot 500 - 0 \cdot 541}{\sqrt{0 \cdot 525 \times 0 \cdot 475 \left(\frac{1}{400} + \frac{1}{600} \right)}}$$

$$Z = -1 \cdot 269$$

$$|Z| = 1 \cdot 269$$

Conclusion:

Since $|Z| = 1 \cdot 269$ which is less than 1.96 it is not significant at 5% los. Hence H_0 may be accepted at 5% los, and we may concluded that men and women do not differ significantly as regards proposal of flyover is concerned.

15.9 Summary:

The difference between hypothesised population parameter and the sample statistic is neither too small to accept ignoring the difference or so large to reject automatically. This type problem are can understand and learn how to proceed objectively to accept or reject an assumption on the basis of sample information we developed a test procedures for different situations.

15.10 Exercises:

1. 20 people were attacked by a disease and only 18 servived will you rejected the hypothesis that survived rate attacked by this disease is 85% is favour of hypothesis that it is more at 5% los.
2. 160 heads are obtained in tossing a coin 400 times thus their appear to be an unbaised one.
3. A machine puts out 10 imperfect articals in a sample of 200. After machine is overhauled its putsout 4 imperfact articals in a sample of 100. As the machine have been included?
4. A sample of 100 rods is drawn from a large population. If the sample mean is 3.2" and S.D. 0.6". can it be infered that the sample has couse from a population with mean 3".
5. In A.U. it was found that standard dividation height if the students was 3 inches. when a sample of 100 students was taken the sample mean is found to be 68 inches. On the bais is if this information can you decide that the average height of the student 70 inches.
6. In a random sample of 400 persons from a large population, 120 are females. Can it be said that males andfemales are int he ratio 5 : 3 in the population? use 1% level of significance.
7. A coin is tossed 900 times and heads appear 490 times. Dies this result support the hypothesis that the coin is unbiased.

8. A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler.

9. In a sample of 500 people in Tamil Nadu 280 are tea drinkers and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in this state at 1% level of significance.

10. A manufacturer claimed that atleast 98% of thesteel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test his claim at a significance level 5%.

11. The machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

12. In a sample of 600 students of a certain college 400 are found to use dot pens. In another college, from a sample of 900 students 450 were found to use dot pens. Test whether the two colleges are significantly different with respect to the habit of using dot pens.

13. In a certain district A, 450 persons were considered regular consumers of tea out of a sample of 1000 persons. In another district B, 400 were regular consumers of ten out of a sample of 800 persons. Do these facts reveal a significant difference between the two districts as far as tea drinking habit is concerned?

14. A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved.

15. A random sample of 100 articles selected from a batch of 2,000 articles shows that the average diameter of the articles = 0.354 with a S.D. = 0.048. Find 95% confidence interval for the avaerage of this batch of 2,000 articles.

16. A sample of 100 iron bars is said to be drawn from a alrge number of bars whose lengths are normally distributed with mean 4 feet and S.D. 0.6 feet. If the sample mean is 4.2 feet, can the sample be regarded as a truly random sample.

17. The mean life of a sample of 100 electric bulbs produced by a company is found to be 1570 hrs with a S.D. of 120 hrs. If μ is the mean life time of all the bulbs produced by the company, test the hypotheis $\mu = 1600$ hrs against the alternative hypothesis $\mu \neq 1600$ hrs at 5% level of significance.

18. A sample of 400 male students is found to have a mean height of 171.38 cm. Can it be reasonably regarded as a sample from alarge population with mean height of 171.17cm and S.D. 3.30 cm.

19. A sample of 100 workers in a large plant gave a mean assembly time of 294 seconds with a S.D. of 12 seconds in a time and motion study. Find a 95% confidence interval for the mean assembly time for all the workers in the plant.

20. The mean breaking strength ofthe cables supplied by a manufacturer is 1800 with a S.D. 100. By a new technique in the manufacturing process, it is claimed that the breaking strength ofthe cables have increased. In order to test this claim, a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance.

21. An investigation fo the relative merits of two kinds of flash light batteries showed that a random sample of 100 batteries of brand A tested on the average 36.5 hrs with a S.D. of 1.8 hrs while a random sample of 80 batteries of brand B tested on the average 36.8 hrs with a S.D. of 1.5 hrs. USe a level of significance of 0.05 to test whether the observed difference between the average life times is significant.

22. Given the following information relating to two places A and B. Test whether there is any significant difference between their mean wages:

	A	**B**
Mean wages (Rs.)	47	49
S.D. (Rs.)	28	40
Number of workers	1000	1500

23. The mean consumption of food grains among 400 sampled middle class consumers is 380 gms per day per person with a S.D. of 120 gms. A similar sample survey of 600 working class consumers gave a mean of 410 gms with a S.D. of 80 gms. Are we justified in saying that the two classes consume the same quantity of food grains. Use 5% level of significance.

24. In a city 250 men out of 750 were found to be smokers. Does this information support the conlusion that he majority of men in this city are smokers.

15.11 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. K. MADHUBABU

Lesson - 16

CHI-SQUARE TESTS

Objectives:

To understand the role of Chi-square

- To understand the role of Chi-square distribution in testing of Hypothesis.
- Using Chi-square statistics in developing and conducting tests of goodness of fit and other tests.
- To educate F - ratio to perform tests concerning the equality of variances of

Structure:

16.1 Introduction:

Sampling distribution describes the manner in which a statistic or a function of statistics which is/are a function(s) of the random sample variate values $x_1, x_2, \cdots, x_n$ will vary from one sample to another of the same size. Some popular and useful sampling distributions are x^2, t, z and f.

16.2 CHI-SQUARE DISTRIBUTION :

The Chi-square distribution was first discovered by 'Helmert' in '1876' and later independently by 'Karl Pearson' in '1900'.

The square of a standard normal variate is known as a chi-square variate i.e., If x is N(0, 1) variate, then x^2 is known as the Chi-square variate of $x \sim N(\mu, \sigma^2)$ then the standard normal deviate $Z = \left(\frac{x-\mu}{\sigma}\right) \sim N(0,1)$ and Z^2 is distributed a Chi-square (χ^2) with 1 d.f. If $x_1, x_2, \cdots, x_n$

are n independent variates distributed as $N(\mu_i, \sigma_i^2)$ then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - u_i}{\sigma_i} \right)^2$$ is distributed as Chi-square with n degrees of freedom.

The probability density function of the Chi-square distribution is

$$f(\chi^2) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} \; e^{-\chi^2/2} \; (\chi^2)^{\frac{n}{2}-1} \quad 0 \le \chi^2 \le \infty$$

The Chi-square can be expressed in terms of sample variance (S^2). If $x_1, x_2, \cdots, x_n$ is a random sample of size n from a normal population the quantity $\frac{KS^2}{\sigma^2}$ is distributed as Chi-square with K d.f. where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ and $K = n - 1$. σ^2 is the population variance of the population from which the sample has been drawn.

To get a confidence interval σ^2 we can locate two points on the χ^2 distribution. χ^2_{UCL} cuts off 0.025 of the area in the upper tail of the distribution, and χ^2_{LCL} cut of 0.025 in lower tail for 5% level of significance following the χ^2 tables.

Limits can be estimated

$$\sigma^2_{LCL} = \frac{Ks^2}{\chi^2_{LCL}}$$ - for lower confidence limit

$$\sigma^2_{UCL} = \frac{Ks^2}{\chi^2_{UCL}}$$ - for upper confidecne limit

16.2 χ^2 WITH N DF : (PROPERTIES)

- Mean of the χ^2 distribution is n.
- Variance of the χ^2 distribution is 2n.
- Mode of the χ^2 distribution is $(1-2t)^{-n/2}$

- C.F. of the χ^2 distribution is $(1-2it)^{-\frac{n}{2}}$
- Measure of Skewness is $\beta_1 = \frac{8}{n}$
- Measure of Kurtosis is $\beta_2 = 3 + \frac{12}{n}$
- If $x_1, x_2, \cdots, x_k$ are k independently distributed Chi-square variates such that $x_i \sim \chi^2_{ni}$, $\sum_{i=1}^{k} x_i$ is also a chi-square variate with Σn_i d.f. This is additive or reproductive property of chi-square.
- If $x \sim \chi^2_{n_1}$, $x + y \sim \chi^2_{n_1+n_2}$, then $y \sim \chi^2_{n_2}$
- If n is large, then $\lim_{n\to\infty} \chi^2_n \to N(n, 2n)$. This is known as the limiting property of the Chi-square.
- If $x \sim \chi^2_{n_1}$ and $y \sim \chi^2_{n_2}$ are two independent Chi-squae variates, the distribution of the quotient (x/y) is beta distribution of type II i.e., $x/y \sim \beta_{11}\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$
- If all $x_i \sim N(0,1)$ for $i = 1, 2, \cdots n, \Sigma x_i^2 \sim \chi^2_n$.

  But if x_i^2's are distributed normally with unit variance and non-zero means i.e., $x_i \sim N(\mu_i, 1)$. The distribution of Σx_i^2 is known as non-central chi-square with non-centrality parameter λ where $\lambda = \frac{1}{2}\Sigma u_i^2$.

16.2.2 APPLICATIONS

χ^2 distribution has a large number of applications in statistics.

- To test if the hypothetical value of the population variance is $\sigma^2 = \sigma_0^2$
- To test the goodness of fit

- To test the independence of attributes
- To test the homogeneity of independent estimates of the population variance.
- To combine various probabilities obtained from independent experiments to give a single test of significance.
- To test the homogeneity of independent estimates of the population correlation coefficient.

16.2.3 CONDITIONS FOR VALIDITY OF χ^2 – TEST

χ^2 test is an approximate test for large values of n.

- The sample observations should be independent.
- Constraints on the cell frequencies, if any, should be linear

  e.g. $\Sigma n_i = \Sigma \lambda_i$ or $\Sigma O_i = \Sigma E_i$
- N, the total frequency should be reasonably large say, greater than 50.
- No theoretical cell frequency should be less than 5. If any theoretical cell frequency is less than 5, then for the application of χ^2- test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

It may be noted that the χ^2- test depends only on the set of observed and expected frequencies and on d.f. It does not make any assumptions regarding the parent population from which the observations are taken. Since χ^2 defined in does not involve any population parameters, it is termed as a statistic and the test is known as non-parametric test or Distribution-Free Test.

16.3 TESTING OF POPULATION VARIANCES :

Suppose we want to test if a random ample $x_i\,(i = 1,2,...,n)$ has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$ (say).

Under the null hypothesis that the population variance is $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})^2}{\sigma_0^2}\right]$$

$$= \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\Sigma x_i)^2}{n}\right]$$

$$= \frac{ns^2}{\sigma_0^2}$$

follows χ^2 distribution with (n - 1) df.

By comparing the calculated value with the tabulated value of χ^2 for (n - 1) d.f. at certain level of significance (5%) we may retain or reject the null hypothesis.

NOTE :

1) The above test can be applied only if the population from which the sample is drawn is normal.

2) If the sample size is large (> 30), then we can use 'Fisher's' approximation

$$\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right)$$

i.e., $Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$ and apply Normal Test.

Examples 1 : (for one tailed test) (as measured by var)

- It is believed that the precision of an instrument is no more than 0.16 carryout the test at 1% level given 11 measurements of the same subject on the instrument :

  2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6,2.7, 2.5

Solution :

Null Hypothesis

$H_0 : \sigma^2 = 0.16$

$H_1 : \sigma^2 > 0.16$

x	$x - \bar{x}$	$(x - \bar{x})^2$
2.5	-0.01	0.0001
2.3	-0.21	0.0441
2.4	-0.11	0.0121
2.3	-0.21	0.0441

2.5	-0.01	0.0001
2.7	0.19	0.0361
2.5	-0.01	0.0001
2.6	0.09	0.0081
2.6	0.09	0.0081
2.7	0.19	0.0361
2.5	-0.01	0.0001

$\bar{x} = \frac{27.6}{11} = 2.51$ $\Sigma(x - \bar{x})^2 = 0.1891$

Under null hypothesis, the test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma(x - \bar{x})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182$$

which follows χ^2 - distribution with d.f. n - 1 = 10

Since the calculated value of χ^2 is less than the tabulated value 23.2 of χ^2 for 10 d.f. at 1% P.O.S., it is not significant.

Hence H_0 accepted and we conclude that the data are consistent with the hypothesis that precision of instrument is 0.16.

16.4 GOODNESS OF FIT TEST :

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "Chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If $f_i\,(i = 1, 2, \cdots, n)$ is a set of observed frequencies and $e_i\,(i = 1, 2, \cdots, n)$ is the corresponding set of expected frequencies then Karl Pearson's Chi-square given by

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - e_i)^2}{e_i}\right]$$

Accept H_0 if $\chi^2 = \chi^2_{\alpha(n-1)}$ and reject H_0 if $\chi^2 > \chi^2_{\alpha(n-1)}$, where χ^2 is calculated value of Chi-square obtained on using given data and $\chi^2_{\alpha(n-1)}$ is the tabulated value of Chi-square for (n-1)df at $\alpha\%$ I.O.S.

Example 2 : The demand for a particular spare part in a factory was found to vary from day-to-day. On a sample study the following information was obtained.

Days :	Mon	Tue	Wed	Thu	Fri	Sat
No. of parts : Demanded	1124	1125	1110	1120	1126	1115

Test the hypothesis that the no. of parts demanded does not depend on the day of the week.

Solution : Null hypothesis, H_0 that the no. of parts demanded does not depend on the day of week.

Under H_0 , the expected frequencies of the spare part demanded on each of the six days.

$$\frac{1}{6}(1124+1125+1110+1120+1126+1115) = 1120$$

Days	Frequency		$(f_i - e_i)^2$	$(f_i - e_i)^2 / f_i$
	Observed	Expected		
Mon	1124	1120	16	0.014
Tue	1125	1120	25	0.022
Wed	1110	1120	100	0.089
Thu	1120	1120	0	0
Fri	1126	1120	36	0.032
Sat	1115	1120	25	0.022
Total	6720	6720		0.179

$$\chi^2 = \Sigma\left(\frac{f_i - e_i}{f_i}\right)^2$$

$$\chi^2 = 0.179$$

The tabulated $\chi^2_{0.05}$ for 5 d.f. = 11.07.

Since calculated value of $\chi^2 <$ tabulated χ^2 it is not significant and H_0 accepted at 5% I.O.S.

$\therefore$ the no. of parts demanded are same over the 6 day period.

16.4.1 A TWO-TAILED TEST OF A VARIANCE :

A Management professor just gave an examination to his class of 31 freshmen and sophomores. The mean score was 72.7 and the sample standard deviation was 15.9. Does the exam meet his goodness criterion. We can summarize the data.

$\sigma H_0 = 13 \leftarrow$ hypothesized value of s.d.

$S = 15.9 \leftarrow$ Sample S.d.

$n = 31 \leftarrow$ sample size.

If the professor uses a significance level of 0.10 in testing his hypothesis, we can symbolically state the problem.

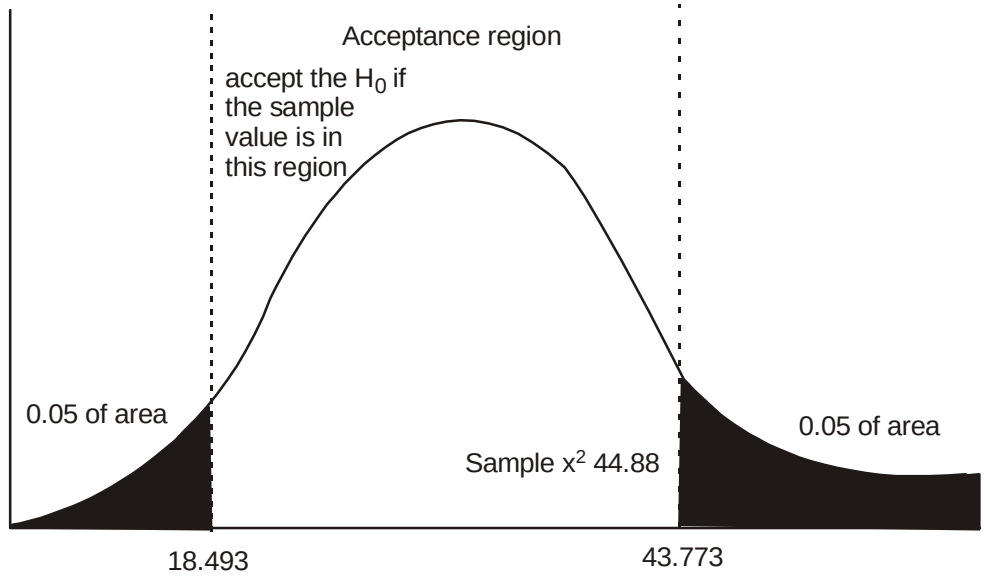
$H_0 : \sigma = 13 \leftarrow H_0 :$ the true standard deviation is 13 points.

$H_1 : \sigma \neq 13 \leftarrow H_1 :$ the true s.d. is not 13 points

$\alpha = 0.10 \leftarrow$ L.O.S. for testing hypothesis.

χ^2 statistic is $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2} = \dfrac{30(15.9)^2}{(13)^2}$ = 44.88

tabulated χ^2 value for 0.05 IOS each tail of the curve are 18.493 & 43.773.



The sample value of χ^2 is not in the acceptance region. So the professor should reject the H_0 .

16.4.2 A ONE-TAILED TEST OF A VARIANCE :

Precision Analytics manufactures a wide line of precision instruments and has a fine reputation in the field for quality of its instruments. It will not release ananalytic balance for sale for example, unless that balance shows a variability significantly below one microgram when weighing quantities of about 500 grams. A new balance has just been delivered to the quality control division from the production line.

The new balance is tested by using it to weigh the same 500 g standard weight 30 different times. The sample s.d turns out to be 0.73 micro grams. Should this balance be sold? We summarize the data

σ_{H_0} : 1 hypothesized value of the population

S = 0.73 sd sample sd

n = 30 sample size and

$H_0 : \sigma = 1$ H_0 : the true s.d. is 1 microgram

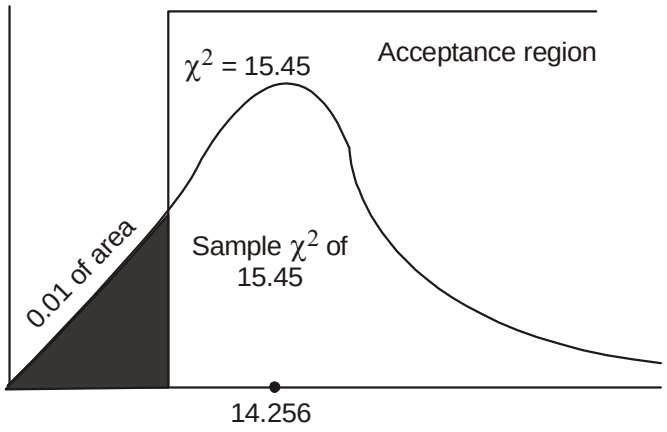
$H_0 : \sigma < 1$ H_1 : The true s.d. is less than 1 microgram

$\alpha = 0.01$ I.O.S. for testing hypothesis

Calculate χ^2 statistic

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

$$= \frac{29(0.73)^2}{(1)^2}$$



We will reject the H_0 and release the balance for sale if the statistic is sufficiently small.

16.4.3 TESTING THE EQUALITY OF TWO POPULATION VARIANCES (F TEST) :

Suppose we want to test

i) Whether two independent samples $x_{ij}\,(i=1,2,\cdots,n_1)$ and $y_j,(j=1,2,\cdots,n_2)$ have been drawn from the normal populations with the same variance σ^2.

ii) Whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis that

i) $\sigma_x^2=\sigma_y^2=\sigma^2$ i.e., the population variances are equal

ii) Two independent estimates of the population variance are homogeneous, the static F given by

$$F=\frac{S_x^2}{S_y^2}$$

where $S_x^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2$

$$S_y^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_j-\bar{y})^2$$

are unbiased estimates of the common population σ^2 obtained from two independent samples and it follows Snedecors F - distribution with (υ_1,υ_2) d.f. where $\upsilon_1=n_1-1$ and $\upsilon_2=n_2-1$

By comparing the calculated of F obtained by using for the two given samples, with the tabulated value of F for (n_1, n_2) d.f. at L.O.S. (5% or 1%) H_0 is either reject or accepted.

16.4.4 A ONE-TAILED TEST OF TWO VARIANCES :

A purchase manager wanted to test if the variance of prices of unbranded bolts was higher than the variance of prices of branded bolts. He needed strong evidence before he could conclude that the variance of prices of unbranded bolts was higher than the variance of prices of a branded bolts. He obtained price quotations from various stores and found that hte sample variance of prices of unbranded bolts from 13 stores was 27.5. Similarly the sample variance of prices of a certain brand of bolts from 9 stores was 11.2 what can the purchase manager conclude at a significance level of 0.05?

SOLUTION : Null Hypothesis : $H_0:\sigma_1^2\le\sigma_2^2$

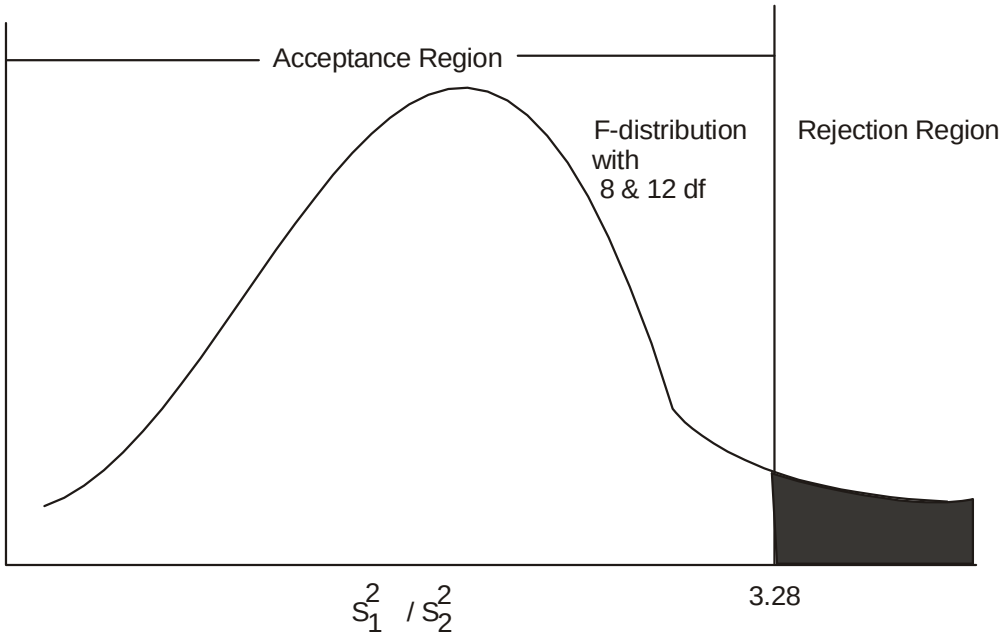
Alternative Hypothesis : $H_1 : \sigma_1^2 > \sigma_2^2$

Testing the equality two population means we had used the difference in sample means as the test statistic. The distribution of $\left(S_1^2 - S_2^2\right)$ is not known. If $\sigma_1^2 = \sigma_2^2$ then $\frac{S_1^2}{S_2^2}$ will have an F-distribution with $\left(n_1 - 1\right)$ and $\left(n_2 - 1\right)$ degrees of freedom.

In this case $n_1 = 13, S_1^2 = 27.5, n_2 = 9, S_2^2 = 11.2$

$$\frac{S_1^2}{S_2^2} = \frac{27.5}{11.2} = 2.455$$

Refering to the F-tables for the distribution with 12 and 8 d.f. we find the value of $\frac{S_1^2}{S_2^2}$ is 3.28 shown in below figure.



As this falls in the acceptance region. We cannot reject H_0 .

$\therefore$ we conclude that we do not have sufficient evidence to justify that unbranded bolts have a higher price variance than that of a given brand

16.4.5 A TWO TAILED TEST OF TWO VARIANCES :

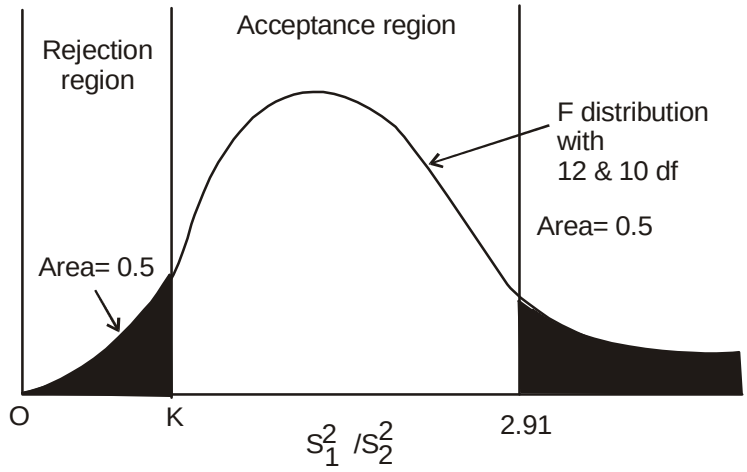
A two tailed test of equality of two varainces is similar to the one tailed test. The only difference is that the critical region would now be split into two parts under both the tails of F-distribution.

Let us take the decision problem by the marketing manager wanted to know if display at point of purchase helped in increasing sales. He picked up 13 retail shops with no display and found that the weekly sale in these shops had a mean of Rs. 6,000 and a standard deviation of Rs. 1004. Similarly he picked up a second sample of 11 retail shops with display at point of purchase and found that the weekly sale in these shops had a mean of Rs. 6500 and a sd of Rs. 1200. If he knew that the weekly sale in shops followed normal distributions, could he reasonably assume that the variances of weekly sale in shops with and without display were equal, if he used a significance level of 0.10?

Null Hypothesis : $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

We shall again use $\frac{S_1^2}{S_2^2}$ as test statistics which follows F - distribution with $(n_1 - 1) \& (n_2 - 1)$ d.f. The critical region splits into two parts figure given below.



Refering the F table the value of $\frac{S_1^2}{S_2^2}$ is 2.91.

Continue to goodness of fit.

16.5 CHI-SQUARE TEST FOR INDEPENDENCE

H_0 : The two-way table is independent.

H_a : The two-way table is not independent

Test Statistic

$$T = \sum_{i=1}^{\gamma} \sum_{j=1}^{c} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

where r = The number of rows in the contingency table.

c = The numberof columns in the contingency table.

O_{ij} = The observed frequency of the ith row and jth column.

E_{ij} = the expected frequency of the ith row and jth column

$$= \frac{R_i C_j}{N}$$

R_i = the sum of the observed frequencies for row i

C_j = the sum of the observed frequencies for column j

N = the total sample size

Significance Level α ,

Critical Region : $T > \text{CHSPPF}(\text{alpha}, (r-1)*(c-1))$

where CHSPPF is the percent point function of the chi-square distribution and $(r-1)*(c-1)$ is the degrees of freedom.

Conclusion : Reject the independence hypothesis if the value of the test statistic is greater than the chi-square value.

EXAMPLE 3 : A researcher wants to study how 7 methods of preparations affect students getting over 80% on a aptitude test. The researcher would also like to know how elapsed time after preparation affects student's performance on the test (> 80%) after 1, 2 and 3 months of final preparation and if these two criteria are related or independent.

The following table is the observed results of the study; contingency table.

Use Chi-square test for independent characteristics to evaluate this data.

Number of Months	Methods of preparation							Row totals
(M) after Prep.	1	2	3	4	5	6	7	
1. After 1 M	97	8	18	8	23	21	5	180
2. After 2 M	120	15	12	13	21	17	15	213
3. After 3 M	82	4	0	12	38	25	19	180
Column Totals	299	27	30	33	82	63	39	573

PROCEEDURE FOR CALCULATING CHI-SQUARE TEST OF INDEPENDENCE :

Step 1 : Make a problems statement : (becomes the hypothesis statement, H_0).

(1) Are the Methods of preparation and number of months after preparation related in terms of students achievement scores (> 80% on test)?

(2) Are the two criteria (methods of Preparation and number of months test is taken after preparation) independent or related to with respect to students performance scores (> 80% on test)?

HYPOTHESIS :

So H_0 **(null hypothesis) :** Methods of Preparation and Months after Preparation are independent. (not related or interacts with respect to students scores).

$$\chi^2 \leq \chi^2_{1-\alpha}$$

H_a : H_0 is not true. (Alternate hypothesis) : Methods of Prep and Months after Preparation are related. $\chi^2 > \chi^2_{1-\alpha}$.

Step 2 : Choose α, the significance level of the test.

If you want the be 95% certain that the test is true, then α = 0.05 = (100 - 95)/100

The df = (r - 1) (c - 1), or (m - 1) (k - 1) = (3 - 1) (7 - 1) = 12

So df = 12

Step 3. Look up $\chi^2_{1-\alpha}$ from Chi-square Table :

df	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22	28.3

for d.f. = 12, $\chi^2_{0.95} = 21.03$

Step 4 (4) Determine or compute $E = \dfrac{(\text{Row Total}) \cdot (\text{Column Total})}{\text{Grand Total}}$ **,The Expected Frequencies:**

The following table is the **Expected Frequencies, E** of each cells in the study :

Number of Months (M) after Prep.	Methods of Preparation							Row Freq-Total
	1	2	3	4	5	6	7	
1.	93.9267	8.4817	9.4241	10.3665	25.7592	19.7906	12.2513	180
2.	111.1466	10.0366	11.1518	12.267	30.4817	23.4188	14.4974	213
3.	93.9267	8.4817	9.4241	10.3665	25.7592	19.7906	12.2513	180
Column Freq. Totals	299	27	30	33	82	63	39	573

Step 5 : Compute $\chi^2 = \Sigma \frac{(O-E)^2}{E}$

Number of Months (M) after Prep.	Methods of Preparation							Row Freq-Total
	1	2	3	4	5	6	7	
After 1M	0.100558873	0.27354405	7.80408377	0.54022952	0.295544417	0.07390925	4.291907191	13.13358743
After 2M	0.705219497	2.45448959	0.064508517	0.04379761	2.949384084	1.759335536	0.017425536	7.994160371
After 3M	1.514438471	2.368095146	9.42408377	0.257401238	5.81688881	1.371263747	3.717548217	24.46971647
Totals	2.320216842	4.849939141	17.29267606	0.841428368	9.061814382	3.204508533	8.026880944	**45.5975**

$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 45.5974$ from **computational table above.**

Step 6 Perform test Chi-square Test :

$$\chi^2 \le \chi^2_{1-\alpha}$$

Since $\chi^2 \le \chi^2_{1-\alpha}$ i.e., 45.60 > 21.03, then we assume that the Null Hypothesis is not true (the types of preparations and number of weeks after preparation are related each other).

16.6 SUMMARY

In this lesson, using test statistic $\frac{(n-1)s^2}{\sigma^2}$ the variance of a normal population is tested. In this case s^2 not known directly. Procedure is developed for testing the equality of variance of two normal population : Goodness of fit test, and chi-square test for independent are discussed.

16.7 EXERCISE

1. A manufacturer claims that any of his lot of items cannot have a variance more than $1 cm^2$. A sample of 25 items has a variance of $1.2 cm^2$. Test whether the claim of the manufacturer is correct?

2. Having a received complaints about slow customer service in a certain public counters in a Bank branch, the chief executive of a most busy bank office ordered for a preliminary investigation. The investigation officer enquired about the time spent by 10 customers in receiving similar bank service in a peak day of banking activities. He has provided the following schedule of information.

S.No.	1	2	3	4	5	6	7	8	9	10
Time Spent By a Customer X (minutes)	60	45	25	75	60	30	90	15	80	60

3. Test the hypothesis that $\sigma = 10$, given that S = 15 for a random sample of size 50 from a normal population.

4. When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows.

No. of mistakes (x) in a page	0	1	2	3	4	5	6
No.of Pages(f)	275	72	30	7	5	2	1

Fit a poisson distribution to the above data and test the goodness of fit.

5. The following data show defective articles produced by 4 machines.

Machine :	A	B	C	D
Production time :	1	2	3	4
No. of defectives:	12	30	63	98

Do the figures indicate a significant difference in the performance of the machines.

6. A sample of size 13 gave an estimated population variance of 3.0 while another sample of size 15 gave an estimate of 2.5 could both samples be from populations with the same variance?

7. Two samples of size 9 and 8 gave the sums of squares of deviation from their respective mean are to 160 and 91 respectively. Can they be regarded as drawn from the same normal population?

Lesson Writer
Dr. K. Madhu Babu

Lesson - 17

BUSINESS FORE CASTING

Objectives:

After completion of this lesson, you should be able to

- Forecasting is the art and science of predicting future events.
- Methods of forecasting
- Difficulties facing in forecasting technology
- Forecasting is invented not predicted
- Difference between the quantitative and Quantitative methods
- Time series in forecasting
- Use moving averages and exponential smoothing in forecasting

Structure:

17.1 Introduction:

Forecasting is the art and science of predicting future events. It is estimating future event (variable), by casting forward past data. Past data are systematically combined in predetermined way to obtain the estimate. Forecasting is not guessing or prediction. Forecasts are required throughout an organization and at all levels of decision making in order to plan for the future and make effective decisions. The principal use of forecasts in operations management is in predicitng the demand for manufactured products and services for time horizons ranging from several years down to 1 day. Forecasting help managers to plan the system, plan the use of system.

Common Features of Forecasting:

1. Forecasting is rarely perfect (deviation is expected)
2. All forecastign techniques assume that there is some degree of stability in the system and "what happend in the past will continue to happen in the future".
3. Forecasting for a group of items is more accurate than the forecast for individuals.
4. Forecasting accuracy increases as time horizon increases.

Assumptions:

a. There is no way to state what the future will be with complete certainty. Regradless of the methods that we use there will always be an element of uncertainty until the forecast horizon has come to pass.

b. There will always be blind spots in forecasts. We cannot, for example, forecast completely new technologies for which there are no existing paradigms.

c. Providing forecasts to policy - makers will help them formulate social policy. The new social policy, in turn, will affect the future, thus changing the accuracy of the forecast.

17.2 Forecasting Methods:

(i) Genius Forecasting: This method is based on a combination of intuition, insight and luck. Science fiction writers have sometimes described new technologies with uncanny accuracy.

There are many examples where men and women have been remarkable successful at predicting the future. There are also many examples of wrong forecasts. The weakness in genious forecasting is that its impossible to recognize a good forecast until the forecast has come to pass.

(ii) Trend Extrapolation: These methods examine trends and cycles in historical data and then use mathematical techniques to extrapolate to the future. The assumption of all these techniques is that the forces responsible for creating the past, will continue to operate in the future. This is often a valid assumption when forecasting short term horizons, but it falls short when creating medium and long term forecasts. The further out we attempt to forecast, the less certain we become of the forecast.

The stability of the environment is the key factor in determining whether trend extrapolation is an appropriate forecasting model. The concept of "developmental inertia" embodies the idea that some items are more easily changed than others. Clothing styles is

an example of an area that contains little inertia. it is difficult to produce reliable mathematical forecasts for clothing. Energy consumption, on the other hand, contains substantial inertia and mathematical techniques work well. The developmental inertia of new industries or new technology cannot be determined because there is not yet a history of data to draw from.

There are many mathematical models for forecasting trends and cycles. Choosing an appropriate model for a particular forecasting application depends on the historical data. The study of the historical data is called exploratory data analysis. Its purpose is to identify the trends and cycles in the data so that appropriate model can be choosen.

The most common mathematical models involve various forms of weighted smoothing methods. Another type of model is known as decomposition. This technique mathematically separates the historical data into trend, seasonal and random components. A process known as a "turning point analysis" is used to produce forecasts. ARIMA models such as adaptive filtering and Box - Jenkins analysis constitute a third class of mathematical mode while simple linear regression and curve fitting is a fourth.

The common feature of these mathematical models is that historical data is the only criteria for producing a forecast. One might think then, that if two people use the same model on the same data that the forecasts will also be the same, but this is not necessarily the case. Mathematical models involve smoothing constants, coefficietns and other parameters that must decided by the forecaster. To a large degree, the choice of these parameters determines the forecast.

It is vogue today to diminish the value of mathematical extrapolation. Makridakis (one of the gurus of quantitative forecasting) correctly points out that judgmental forecasting is superior to mathematical models, however, there are many forecasting applications where computer generated forecasts are more feasible. For example, large manufacturing companies often forecast inventory levels for thousands of items each month. It would simply not be feasible to use judgmental forecasting in this kind of application.

(iii) Consensus Methods: Forecasting complex systems often involves seeking expert opinions from more than one person. Each is an expert in his own discipline, and it is through the synthesis of these opinions that a final forecast is obtained.

A better method is known as the Delphi Technique. This method seeks to rectify the problems of face - to - face confrontation in the group, so the responses and respondnets remain anonymous. The Classical technique proceeds in well - defined sequence. In the first round, the participants are asked to write their predictions. Their responses are collated and a copy is given to each of the participants. The participants are asked to comment on extreme views and to defend or modify their original opinion based on what the other participants have written. Again, the answers are collated and fed back to the participants. In the final round, participants are asked to reassess their original opinion in view of those presented by other participants.

The Delphi method general produces a rapid narrowing of opinions. It provides more accurate forecasts than group discussions. Furthermore, a face - to face discussion following the application of the Delphi method generally degrades accuracy.

(iv) Simulation Method: Simulation methods involve using analogs to model complex systems. These analogs can take on several forms. A mechanical analog might be a wind tunnel for modeling aircraft performance. An equation to predict an economic measure would be a mathematical analog. A metaphorical analog could involve using the growth of a bacteria colony to describe human population growth. Game analogs are used where the interactions of the players are symbolic of social interactions.

Mathematical analogs are of partiticular importance to futures research. They have been extremely successful in many forecasting applications, especially in the physical sciences. In the social sciences however, their accuracy is somewhat diminished. The extraordinary complexity of social systems makes it difficult to include all the relevant factors in any model.

Clarke reminds us of a potential danger in our reliance on mathematical models. As he points out, these techniques often begin with an initial set of assumptions, and if these are incorrect, then the forecasts will reflect and amplify these errors.

One of the most common mathematical analogs in societal growth is the S - curve. The model is based on the concept of the logistic or normal probability distribution. All processes experience exponential growth and reach an upper asymptopic limit. Models has hypothesized that chaos like states exist at the beginning and end of the S - curve. The disadvantage of this S - curve model is that it is difficult to know at any point in time where you currently are on the curve, or now close you are to the asymtopic limit. The advantage of the model is that it forces planners to take a long - term look at the future.

Another common mathematical analog involves the use of multivariate statistical techniques. These techniques are used to model complex systems involving relationships between two or more variables. Multiple regression analysis is the most common technique. Unlike trend extrapolation models, which only look at the history of the variable being forecast, multiple regression models look at the relationship between the variable being forecast and two or more other variables.

Multiple regression is the mathematical analog of a systems approach, and it has become the primary forecasting tool of economists and social scientists. The object of multiple regression is to be able to understand how a group of variables (working in unison) affect another variable.

The multiple regression problem of collinearity mirrors the practical problems of a systems approach. Paradoxically, strong correlations between predictor variables create unstable forecasts, where a slight change in one variable can have dramatic impact on another variable. In a multiple regression (and systems) approach, as the relationships between the components of the system increase, our ability to predict any given component decreases.

(v) Cross - Impact Matrix Method: Relationships often exist between events and developments that are not revealed by univariate forecasting techniques. The cross - impact matrix method recognizes that the occurrence of an event can, in turn, effect the likelihoods of other events. Probabilities are assigned to reflect the likelihood of an event in the presence and absence of other events. The resultant inter - correlational structure can be used to examine the relationships of the components to each other and within the overall system. The advantage

of this technique is that it forces forecasters and policy - makers to look at the relationships between system components, rather than viewing any variable as working independently of the others.

(vi) Scenario: The scenario is a narrative forecast that describes a potential course of evetns. Like the cross - impact matrix method, it recognizes the interrelationships of system components. The scenario descibes the impact on the other components and the system as a whole. It is a "script" for defining the particulars of an uncertain future.

Scenarios consider events such as new technology, population shifts and changing consumer preferences. Scenarios are written as long - term predictions of the future. A most likely scenario is usually written, along with at least one optimistic and one pessimistic scenario. The primary purpose of a scenario is to provoke thinking of decision makers who can then posture themselves for the fulfillment of the scenario(s). The three scenarios force decision makers to ask: 1) Can we survive the pessimistic scenario, 2) Are we happy with the most likely scenario and 3) Are we ready to take advantage of the optimistic scenario?

(vii) Decision Trees: Decision trees originally evolved as graphical devices to help illustrate the structural relationships between alternative choices. These trees were originally presented as a series of yes/no (dichotomous) choices. As our understanding of feedback loops improved, decision trees became more complex. Their structure became the foundation of computer flow charts.

Computer technology has made it possible create very complex decision trees consisting of many subsystems and feedback loops. Decisions are no longer limited to dichotomies they now involve assigning probabilities to the likelihood of any particular path.

Decision theory is based on the concept that an exprected value of a discrete variable can be calculated as the average value for that variable. The expected value is especially useful for decision makers because it represents the most likely value based on the probabilities of the distribution function. The application of Bayee's theorem enables the modification of initial probability estimates, so the decision tree becomes refined as new evidence is introduced.

Utility theory is often used in conjuction with decision theory to improve the decision making process. It recognizes that dollar amounts are not the only consideration in the decision process. Other factors, such as risk, are also considered.

(viii) Combining Forecasts: It seems clear that no forecasting technique is appropriate for all situations. There is substantial evidence to demonstrate that combining individual forecasts produces gains in forecasting accuracy. There is also evidence that adding quantitative forecasts to qualitative forecasts reduces accuracy. Research has not yet revealed the conditions or methods for the optimal combinations of forecasts.

Judgmental forecasting usually involves combining forecasts from more than one source. Informed forecasting begins with a set of key assumptions and then uses a combination of historical data and expert opinions. Involved forecasting seeks the opinions of all those directly affected by the forecast (e.g., the sales force would be included in the forecasting process). These techniques generally produce higher quality forecasts than can be attained from a single source.

Combining forecasts provides us with a way to compensate for deficiencies in a forecasting technique. By selecting complementary methods, the shortcomings of one technique can be offset by the advantages of another.

17.3 Difficulties in Forecasting Technology:

(a) Nearly all futurists describe the past as unchangeable, consisting as a collection of knowable facts. We generally perceive the existence of only one past. When two people give conflicitng stories of the past, we tend to believe that one of them must be lying or mistaken.

(b) This widely accepted view of the past might not be correct. Historians often interject their own beliefs and biases when they write about the past. Facts become distorted and altered over time. It may be that past is a reflection of our current conceptual reference. In the most extreme viewpoint, the concept of time itself comes into question.

(c) Cognitive dissonance theory in psychology has helped us understand that resistance to change is a natural human characteristic. It is extremely difficult to venture beyond our latitudes of acceptance in forecasting new techologies.

(d) When a major technological breakthrough does occur, it takes conviction and courage to accept the implications of the finding. Even when the truth is starring us in the face, we often have difficulty accepting its implications.

(e) The future, on the other hand, is filled will uncertainty. Facts give way to opinions. The facts of the past provide the raw materials from which the mind makes estimates of the future. All forecasts are opinions of the future (some more carefully formulated than others). The act of making a forecast is the expression of an opinion. The future, consists of a range of possible future phenomena or events. These futuribles are those things that might happen.

17.4 Forecasts Create The Future:

A paradox exists in preparing a forecast. If a forecast results in an adaptive change then the accuracy of the forecast might be modified by that change. Suppose the forecast is that our business will experience a ten percent drop in sales next month. We adapt by increasing our promotion effort to compensate for the predicted loss. This action, in turn, could affect our sales, thus changing the accuracy of the original forecast.

Many futurists that the way we contemplate the future is an expression of oyr desire to create that future. Physicist Dennis Gabor, discoverer of holography, claimed that the future is invented, not predicted. The implication is that the future is an expression of our present thoughts. The idea that we create our own reality is not a new concept. It is easy to imagine how thoughts might translate into actions that affect the future.

An incredible discovery was made at the University of Paris in 1982. A team of researchers lead by Alain Aspect found that under certain conditions, electrons could instantaneously communicate with each other across long distances. The results of this experimetn have been confirmed by many other researchers, although the implications are exceedingly hard to accept. Three explanations are possible: 1) Information can be transferred at speeds exceeding the speed of light, 2) the passage of time tis an illusion, 3) the distance between the electrons is an illusion. All three explanations rock our perception of reality.

David Bohm has explained Aspect's experiment by hypothesizing a holographic universe in which reality is essentially a projection of some deeper dimension that we are not able to comprehend. Instantaneous communication is possible because the distance between the particles is an illusion. neurophysiologist Karl Pribram has also theorized about the holographic nature of reality. His theory is based on a study of the way that the brain recalls memory patterns, but the implications are the same. Reality is a phantasm.

The phenomena of being able to see the future is known as precognition. Most people believe that (to some degree) they can predict the future . Fortune - tellers, however, believe they can view the future. There is a major difference. We predict the future based on knowledge, intuition and logic. Precognitive persons claim to "see" the future. Knowledge and logic are not involved.

Throughout history, there have been many reports of gifted psychics with precognitive powers. Through some unknown mechanism, these people are able predict things that will happen in the future. If we admit that even a single person in history has possessed this capability, the we must accept the fact that our concept of reality needs dramatic alteration. Time itself may not exist as we currently perceive it. Forecasting may be a method of creating illusions.

Forecasting can and often does contribute to the creation of the future but it is clear that other factors are also operating. A holographic theory would stress the interconnectedness of all elements in the system. At some level, everything contributes to the creation of the future. The degree to which a forecast can shape the future (or our perception of the future) has yet to be determined experimentally and experientially.

Sometimes forecasts become part of a creative process, and sometimes they don't. When two people make mutually exclusive forecasts, both of them cannot be true. At least one forecast is wrong. Does one person's forecast create the future, and the other does not? The mechanisms involved in the construction of the future are not well understood on an individual or social level.

Modis believes that the media provides the mechanism by which social forecasts take on a creative context. In this theory, extensive media coverage acts as a resonating cavity for public opinion, and creates a "cultural epidemic" that modifies social behavior.

17.5 Good Forecast:

1. **Timely:** Forecasting horizon must cover the time necessary to implement possible changes.
2. **Reliable:** It should work consistently.
3. **Accurate:** Degree of accuracy should be stated.
4. **Meaningful:** Should be expressed in meaningful units. Finanical planners should know how much amount needed, production should know how many units to be produced, and schedulers need to know what machines and skills will be required.
5. **Written:** To guarantee use of the same information and to make easier comparison to actual results.

6. Easy to use: Users should be comfortable working with forecast.

Classification of forecasting by time:

- Short - range (days - weeks - months) job scheduling, work assignments.
- Time spans ranging from a few days to a few weeks.
- Cycles, seasonality and trend may have little effect.
- Random fluctuation is main data component.
- Medium term (1 - 2 years) Sales, production
- Long range forecast (> 2 years) : Change location
- Time spans usually greater than one year.

Example of sales forecast:

Month	Forecast of Sales	Actual Sales	Error	Squared Error $= E^2$
1	10	8	2	4
2	8	12	4	16
3	11	7	4	16
4	14	16	2	4
5	10	8	2	4
Total			14	44

17.6 Forecast Accuracy:

- Total Absolute Deviation (TAD) = 14
- Mean Absolute Deviation (MAD) = $\sum (\text{Actual} - \text{Forecast}) / n = 14/5 = 2 \cdot 8$
- Total Squared Error (TSE) = 44
- Mean Square Error (MSE) = $44 / 5 = 8 \cdot 8$

Steps in forecast development:

1. Determine purpose of forecast
2. Establish a time horizon - time limit, accuracy decreases with shorter durations.
3. Select forecasting technique
4. Gather and Analyze Data
5. Prepare the forecast
6. Monitor forecast

17.7 Types of Forecasts:

In general, a contemporary business organization employs three distrinct types of forecasts:

These are given under:

1. Economic Forecasts
2. Technological Forecasts
3. Demand Forecasts

Economic Forecasts address the business cycle by predicting infation rates, money supplies, housing starts and other planning indicators.

Technological forecasts are concerned with rates of technological progress which can result in the birth of exciting new products, requiring new plants and equipment.

Demand forecasts are projections of demand for a company's products or services. These forecats, also called sales forecasts, drive a company's production, capacity and scheduling systems and serve as inputs to finanical marketing and personnel planning.

17.8 Strategic importance of forecasting:

Forecasting plays a very important role in the following areas:

- Human resource management

  (- hiring, training and laying - off workers all depend on anticipated demand)

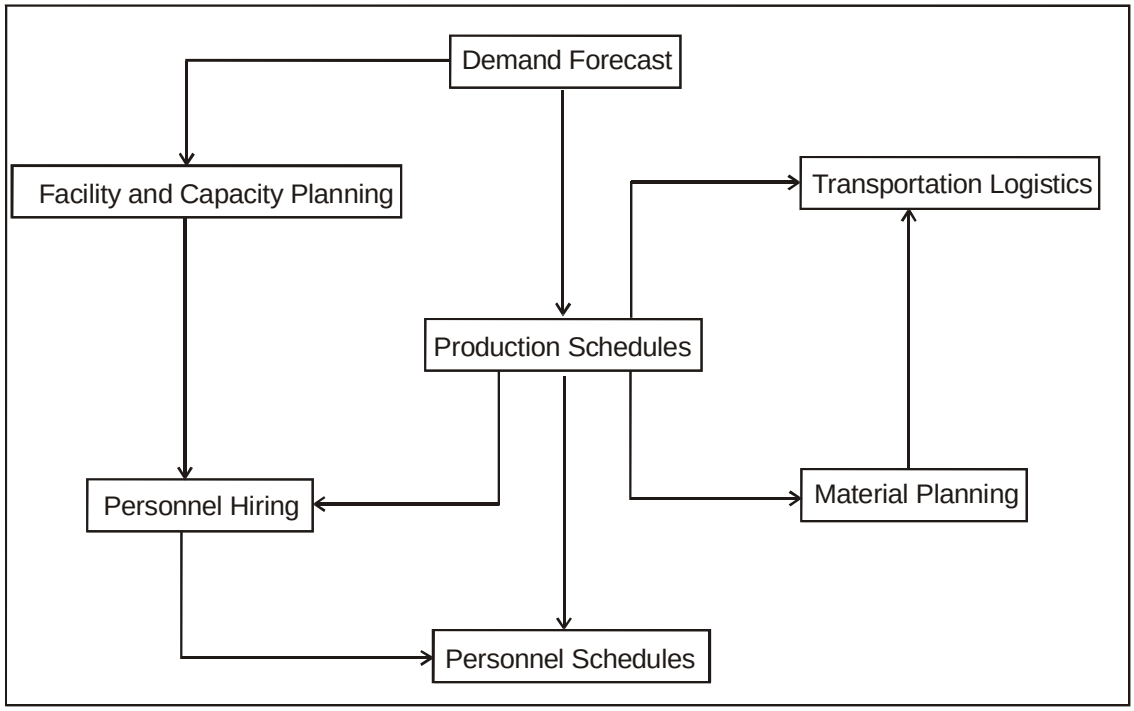
- Capacity planning

  (- when capacity is inadequate the resulting shortages can mean undependable delivery loss of customers and loss of market share)

- Supply - Chain management

  (- good supplier relations and the ensuring price advantages for materials and parts depend on accurate forecasts)

Now that we have a clear idea of forecasting and its significance, let us try to focus on the different facets of forecasting:



17.9 Forecasting Approaches:

There are numerous approaches to forecasting depending on the need of the decision maker, Broadly speaking, these can be categorized in two ways:

- Quantitative Forecasting
- Qualitative Forecasting

In general, we should consider using qualitative forecasting techniques when one or more of the following conditions exist:

1. Little or no historical data on the phenomenon to be forecast exist.
2. The relevant environment is likely to be unstable during the forecast horizon
3. The forecast has a long time horizon, such as more than three to five years.

Qualitative Methods of Forecasting:

The various qualitative methods are as follows:

1. Jury of executive opinion: This method takes the opinions of a small group of high - level managers, often in combination with statistical models, and results in a group estimate of demand.

2. Delphi Method:

(a) Panel of experts queried,

(b) Chosen experts to participate should be a variety of knowledgable people in different areas,

(c) Through questionnare the coordinator obtains estimates from all participants

(d) Coordinator summar rizes results and redistributes them to participants along with appropriate new questions.

(e) Summarize again and refine forecasts and develop new questions.

3. Consumer market survey: This method take input from customers or potential customers regarding their future purchasing plans. It can help not only in preparing a forecast but also in improving product design and planning for new products.

4. Naive approach: The forecast for any period equals the previous period's actual value.

- Simple to use
- Virtually no cost
- Quick and easy to prepare
- Easily understandable
- Cannot provide higher accuracy
- Can be a standard for accuracy and cost
- Can be applied in stable demand

Example:

1. Sales of air conditionning units next july, will be the same as the sales in last july (seasonal)

2. Highway traffic next tuesday will be the same as last tuesday

It assumes that demand in the next period is the same as demand in the most recent period. In other words, if sales of a product, say, Reliance WLL phones, were 100 units in January, we can forecast that February's sales will also be 100 phones. Does this make any sense? It turns out that for some product lines, selecting this naive approach is a cost - effective and efficient forecasting model.

To illustrate, let us see how theses techniques are put into practice. In the following practical problem, we would examine the role of forecasting as applicable to POM in practice.

We shall see how Delphi method of forecasting is applied.

POM in practice: Forecasting with the Delphi Method - American Hoist and Derrick is a manufacturer of construction equipment, with annual sales of several million dollars. Their sales forecast is an actual planning figure and is used to develop the master production schedule, cash flow projections, and work - force plans. One of the important components of their forecasting process is the use of the Delphi method of judgmental forecasting.

In 1985, top management wanted an accurate 5 year forecast of their sales in order to plan for expansion of production capacity. The Delphi method was used in conjunction with regression models and exponential smoothing in order to generate a forecast. A panel of 23 key personnel was established, consisting of those who had been making subjective forecasts those who had been using them or were affected by the forecasts, and those who had a strong knowledge of the market and corporate sales. Three rounds of the Delphi method were performed, each requesting estimates of:

- Gross national product;
- Construction equipment industry shipments;
- American Hoist and derrick construction equipment group shipments; and
- American Hoist and Derrick corporate value of shipments.

As the Delphi technique progressed, responses for each round were collected, analyzed, and summarized and reported back to the panel. In the third - round questionnaire, not Only were the responses of the first two rounds included, but - in addition - related facts, figures, and views of external experts are sent.

As a result of the Delphi experiment, the 1995 sales forecast error was less than 0.33 percent; in 1996 the error was under 4 percent. This was considerable improvement over previous forecast errors of plus or minus 20 percent. In fact, the Delphi forecasts were more accurate than regression models or exponential smoothing which had forecast errors of 10 to 15 percent. An additional result of the exercise was educational in nature. Managers developed a unifoprm outlook on business condiitons and corporate sales volume and thus had a common base for decision making.

Let us now discuss about quantitative approach of forecasting.

17.10 Quantitative Methods:

The chief quantitative methods are:

1. Moving averages
2. Exponential smoothing

— Time Series Models

3. Trend projection
4. Linear regression

— Casual Model

The time series models of forecasting predict on the basis of the assumption that the furture is a function of the past. In other words, they look at what has happened over a period of time and use a series of past data to make a forecast. If we are predicitng weekly sales of washing machine, we use the past weekly sales for washing machine in making the forecast.

A causal model incorporates into the model the variables or relationships that might influence the quantity being forecast. A casual model for washing machine sales might include relationships such as new housing, advertising budget, and competitors prices. Moving over to a structured approach to forecasting, let me introduce the basic steps involved in this process:

Steps in Forecasting:

There are eight steps to a forecasting system.

These are:

1. Determine the use of the forecast - (what objectives are we trying to achieve?)
2. Select the items that are to be forecasted
3. Determine the time horizon of the forecast (Is it short, medium, or long - range?)
4. Select the forecasting model
5. Gather the data needed to make the forecast
6. Validate the forecasting model
7. Make the forecast
8. Implement the results

We now focus our attention to one of the most widely used and effective method of forecasting.

Time Series Forecasting:

A time series is based on a sequence of evenly spaecd (weekly, monthly, quarterly and so on) data points. Forecasting time series data implies that furture values are predicted only from past values and that other variables, no matter how potentially valuable, are ignored.

Decomposition of a Time Series:

There are four main ways of decomposing the time series:

- Trend
- Seasonality
- Cycles
- Random variations

Two general forms of time series models are used in statistics. The most widely used is a multiplicative model, which assumes that demand is the product of the four components:

Demand = T x S x C x R,

where T denotes Trend

S denotes Season

C denotes Cycles

R denotes random variables

An additive model provides an estimate by adding the components together. It is stated as:

Differences between qualitative and quantitative

Qualitative Methods	Quantitative Methods
Use when situation is vague	Used instable situations
and little data available	Historical data available
New products	Existing products
New Technology	Current Technology
	Involves mathematical techniques
Example: Forecasting newly	Example: Sales of color TVs
introduced online sales	

17.11 Seasonalized Time Series Regression Analysis:

- Select a representative historical data set
- Develop a seasonal index for each season
- use the seasonal indexes to deseasonalize the data
- perform linear regression analysis on the deseasonalized data
- use the regression equation to compute the forecasts
- use the seasonal indexes to reapply the seasonal patterns to the forecasts.
- seasonalized Times, series regression analysis

An analyst at CPC wants to develop next year's quarterly forecasts of sales revenue for CPC's line of Epsilon Computers. She believes that the most recent 8 quarters of sales are representative of next year's sales.

Representative Historical Data Set

Year	Qtr.	($ mil.)	Year	Qtr.	($ mil).
1	1	7.4	2	1	8.3
1	2	6.5	2	2	7.4
1	3	4.9	2	3	5.4
1	4	16.1	2	4	18.0

1. Compute the seasonal indexes

Year	Q1	Q2	Q3	Q4	Total
1	7.4	6.5	4.9	16.1	34.9 (year 1)
2	8.3	7.4	5.4	18.0	39.1 (y2)
Totals	**15.7**	**13.9**	**10.3**	**34.1**	**74.0**
Qtr. Avg	7.85	6.95	5.15	17.05	9.25
Seas. Ind =					
Q. average/ 9.25	. 849	. 751	. 557	1.843	4.000

2. Deseasonalize the data

Quarterly Sales (= actual quarter sales / seasonality index)

Year	Q1	Q2	Q3	Q4
1	8.72	8.66	8.80	8.74
2	9.78	9.85	9.69	9.77

Notice that results have no seasonal variations.

3. Perform Regression on Deseasonalized Data

X	y	x^2	x y
1	8.72	1	8.72
2	8.66	2	17.32
3	8.8	9	26.40
4	8.74	16	34.96
5	9.78	25	48.90
6	9.85	36	59.1
7	9.69	49	67.83
8	9.77	64	78.16

$$a = \frac{204(74\cdot 01) - 36(341\cdot 39)}{8(204) - (36)^2} = 8\cdot 357$$

$$b = \frac{8(341\cdot 39) - 36(74\cdot 01)}{8(204) - (36)^2} = 0\cdot 199$$

$$y = 8\cdot 357 + 0\cdot 199\ x$$

4. Compute the deseasonalized forecasts:

$$y_9 = 8 \cdot 357 + 0 \cdot 199(9) = 10 \cdot 148$$

$$y_{10} = 8 \cdot 357 + 0 \cdot 199(10) = 10 \cdot 347$$

$$y_{11} = 8 \cdot 357 + 0 \cdot 199(11) = 10 \cdot 546$$

$$y_{12} = 8 \cdot 357 + 0 \cdot 199(12) = 10 \cdot 745$$

Note: Average sales are expected to increase by 0.199 million (about $ 200,000) per quarter.

5. Seasonalize the forecasts: (= deseasonalized forecasts x seasonality index)

Yr	Quarter	Index	Deseasonalized Forecast	Seasonalized Forecast
3	1	.849	10.148	8.62
3	2	.751	10.347	7.77
3	3	.557	10.546	5.87
3	4	1.843	10.745	19.80

17.12 Forecasting by smoothing techniques:

A time series is a sequence of observations which are ordered in time. Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of cancelling the effect due to random variation. Widely used techniques are "smoothing". These techniques, when properly applied, reveals more clearly the underlying trends.

Moving Average:

Technique that averages a number of recent actual values updated as new values become available. It can be calculated using the following equation:

$$F_t = MA_n = \sum A\ i/n$$

where: Number of periods = n

Actual values in period i = A_i

Moving Average = MA

Index corresponds to period = i

Forecast for time period t = F_t

Example: MA_3 refers to a three period moving average forecast, and MA_5 would refer to a five period moving average forecast.

Calculate three period moving average for:

Period	Demand
1	42
2	40
3	43
4	40
5	41

$$F_6 = (43 + 40 + 41)/3 = 41 \cdot 33$$

If actual demand in period 6 turns out to be 39, so

$$F_7 = (40 + 41 + 39)/3 = 40 \cdot 00$$

Note That: The forecast is updated by adding the newest actual value and dropping the oldest.

Advantage of movign average: Easy to use and to compute

Disadvantage: Values in the average are weighted equally. For example in a ten period moving average.

Weighted moving average:

More recent values in a series are given more weight in computing a forecast.

Example:

The weight of most recent value = 0.40, next most recent weight = 0.30, next = 0.20 and next = 0.10

Total weights always = 1

In the last example: forecast of period 6 will be

$$F_6 = 0 \cdot 40(41) + 0 \cdot 30(40) + 0 \cdot 20(43) + 0 \cdot 10(40) = 41$$

If actual demand of period 6 is 39. Forecast of period 7 will be:

$$F_7 = 0 \cdot 40(39) + 0 \cdot 30(41) + 0 \cdot 20(40) + 0 \cdot 10(43) = 40 \cdot 2$$

Advantage: More reflective of the most recent occurrences.

Exponential smoothing:

Weighted averaging method based on previous forecast plus a percentage (α) of the forecast error:

Next forecast = Previous forecast + α (Actual - previous forecast)

where (Actual - previous forecast) = forecast error, α is a percentage of the error

$$F_t = F_{t-1} + \alpha\left(A_{t-1} - F_{t-1}\right)$$

where

F_t = Forecast for period t

F_{t-1} = Forecast for previous period

α = Smoothing constant

A_{t-1} = Actual demand or sales for the previous period

Example:

If the previous forecast was 42 units, actual demand was 40 units, and $\alpha = 0{\cdot}10$. The new forecast would be:

$$F_t = 42 + 0{\cdot}10\left(40 - 42\right) = 41{\cdot}8$$

then if the actual demand turns out to be 43, the next forecast would be:

$$F_t = 41{\cdot}8 + 0{\cdot}10\left(43 - 41{\cdot}8\right) = 41{\cdot}92$$

Period	Actual Demand	$\alpha = 0\times10$		$\alpha = 0\times40$	
		Forecast	Error	Forecast	Error
1	42	-	-	-	-
2	40	42	- 2	42.00	- 2
3	43	41.8	1.2	41.20	1.8
4	40	41.92	- 1.92	41.92	- 1.92
5	41	41.73	- 0.73	41.15	- 0.15
6	39	41.66	- 2.66	41.09	- 2.09
7	46	41.39	4.61	40.25	5.75
8	44	41.85	2.15	42.55	1.45
9	45	42.07	2.93	43.13	1.87

Relation between the smoothing constant and response to error:

- Exponential smoothing is one of the most widely used techniques in forecsting.
- The quickness of the forecast adjustment to error is determined by the smoothing constant α.
- The closer the value of α to zero, the slower the forecast will respond to error $\rightarrow$ more smoothing.
- The closer the value of α to 1.00, the greater the forecast will respond to error $\rightarrow$ less smoothing.
- Smoothing means that values are less variable $\rightarrow$ smooth curve
- To choose the best forecasting method $\rightarrow$ calculate forecasts and choose method with least MAD. So, steps will be: make forecasting by various methods $\rightarrow$ calculate MAD for each method $\rightarrow$ method with the least MAD is the best (in exam question) Notice that

17.13 Summary:

If reality is an illusion, then the future is also an illusion. The sucess of any business depends on its future estimates. Forecasting is different from predication and projection. Forecasting is a method of forecelling the course of business activity based on the analysis of past and present data mixed with the consideration of ensuring economic policies and circumstances. Forecasting means fore warning. Forecasts based on statistical analysis are much more reliable than a mere guess work.

17.14 Exercises:

1. Explain the forecasting methods
2. What are the difficulties in forecasting technology
3. Explain the Good forecasts
4. Describe the types of forecasts
5. What are the strategies of forecasting
6. Mentions the fore casting approaches
7. What are the difference between quantitative and Qualitative methods
8. Give suitable example for seasonalized timeseries and regression analysis
9. Explain the forecasting by smoothing techniques

9.9 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.
2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.
3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Prof: M. KOTESWARA RAO

Lesson - 18

CORRELATION

Objectives:

After completion of this lesson, you should be able to:

- To learn how many business decisions depend in knowing the specific relationship between two variables.
- To use scatter diagrams to visualize the relationship between two variables.
- To learn how correlation analysis describes the degree to which two variables are linearly related to each other.
- Significance of Rank correlation.
- Appreciate some practical applications of correlation.
- Concept of auto correlation and time series.

Structure:

18.1 Introduction:

Modern business requires managers to make professionals decisions every day. Which should depend upon predictions of future event. To make better use of foreccast they rely on relationships (intuitive and calculated) between related events. If decision makers can determine the strength of relationship that exists between variables. It can aid the decision making process considerably.

Examples:

The relationship between the age of husband and age of wife, price of a commodity and the amount demanded, heights and weights of a group of persons, income and expenditure of a group of persons etc.

Meaning:

The term carrelation indicates the relationship between two such variables in which with change in the values of one variable. The values of the other variable also changes. According to croxton and cowden "when the relationship is of a quantitative nature the appropriate statistical tool for discovering and expressing it in a brief farmula is known as correlation".

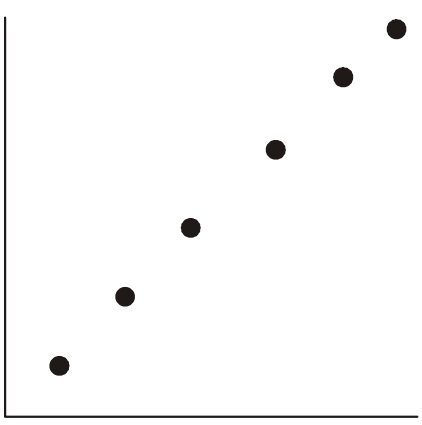
Uses of Correlation:

The study of correlation is useful in practical life because of the reasons as follows:

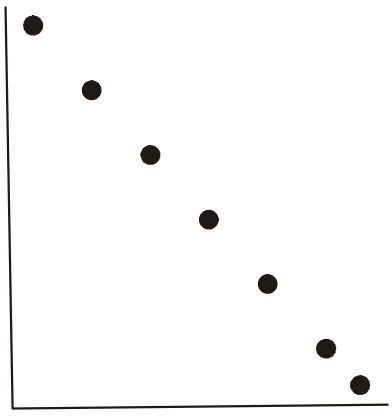
1. With the help of correlation analysis one can measure the degree of relationship that exists between the variable in one figure.
2. From one variable we can estimate the other variable by the help of regression analysis only when we establish the variables are related.
3. Correlation study helps us in identifying such factors which can stabilize a disturbed economic situation.
4. Interrelationship studies between different variables are helpful tools in promoting research.

Scatter Diagram Method:

The scatter diagram is the simplest method of studying relationship between two variables. The simplest device for ascertaining whether two variables are related is to prepare a dot chart harizontal axis representing one variable and vertical axis representing the other. The diagram of dots so obtained is known as scatter diagram. From the scatter diagram we can form a fairly good though rough idea about the relationship between two variables. The following diagrams of the scattered data depiet different type of correlation.



Perfect positive
linear correlation



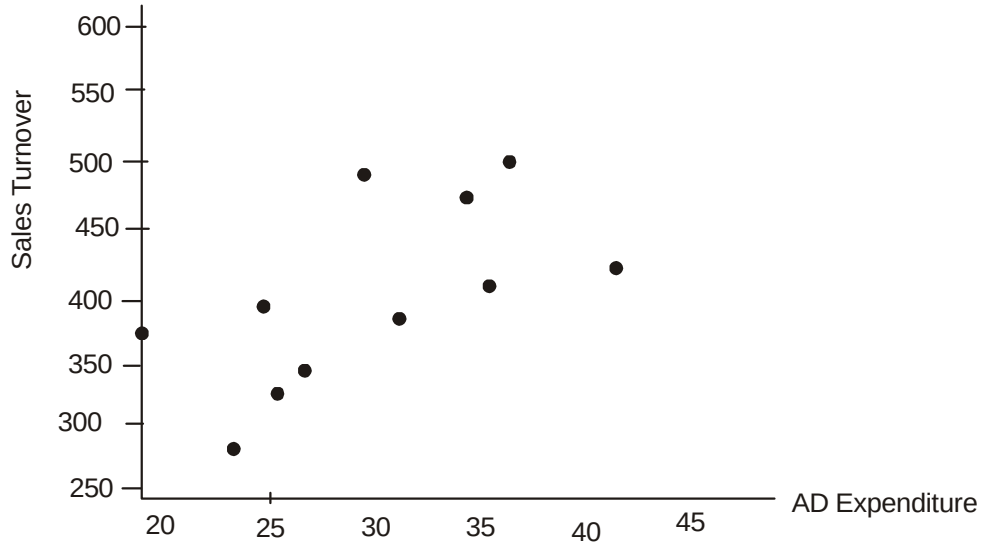
Perfect Negative
Linear correlation

Example 1:

The consultancy organisation wants to know whether the largerod budget has a broader bose on the sales revenue. The data are as follows:

Sales and advertising Expenditure Details

Unit	A	B	C	D	E	F	G	H	I	J	K
Sales Turnover	410	370	380	390	490	470	420	340	360	480	290
Add Exp	36	20	31	24	37	35	42	26	27	29	23



18.2 THE CORRELATION COEFFICIENT:

Define:

Correlation coefficient is a measure of degree or extent of linear relationship between two variables x and y. The population correlation coefficeint is denoted by P and its estimate by r.

Farmula:

Let (x_i, y_i) be the n paired sample values the formula for the sample carrelation coefficient r is

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad \text{far} \quad i = 1, 2, \cdots, n$$

$$= \frac{COV\ (x, y)}{\sqrt{V(x)\ V(y)}} = \frac{S_{xy}}{S_x \cdot S_y}$$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}}{\sqrt{\left\{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}\right\}\left\{\sum_i y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right\}}}$$

Types of Correlation:

Correlation may be classified into three categeries namely

(1) Positive Correlation

(2) Negative Correlation

(3) Linear (or) Non - Linear Correlation

(1) Positive or Negative Correlation: It is otherwise called as direct (positive) or inverse (negative) correlation. The correlation between two variables is said to be direct or positive when they make in the same direction. So that an increase or decrease in the value of one variable is associated with increase or decrease in the value of the other on the other hand if the variables move in the opposite direction i.e., increase in one variable is associated with decrease in the other and vice versa the correlation is said to inverse or negative the following example would example the difference between positive and negative correlation.

Positive Correlation:

(1)

X	10	15	20	25
Y	29	35	40	49

(2)

X	17	13	7	5
Y	34	22	15	7

Negative Correlation:

(1)

X	20	30	40	50
Y	50	40	35	20

(2)

X	60	40	30	20
Y	100	120	125	135

Karl Pearson's Coefficient of Correlation Method:

As a measuring of intensity or degree of linear relationship between two variables Karl Pearson (1867 - 1936). British Biometrician developed a formula called correlation coefficient.

Correlation coefficient between two random variables x and y useually denoted by $r(x, y)$ or simply r_{xy} is a numerical measure of linear relationship between them and is defined as. The pearsons coefficient is denoted by 'r'. The formula for computing r is

$$r = \frac{\sum xy}{N\sigma x\, \sigma y}$$

where $= (X - X')$

$$Y = (Y - Y')$$

σ_x = Standard deviation of x series

σ_y = Standard deviation of y series

N = Number of pair observations

r = Correlation Coefficient

It should be noted that this method is to be applied only where the deviations are taken from the actual mean. The value of coefficient of correlation as obtained by the above method shall always lie between ± 1 when $r = \pm 1$ it means there is a perfect positive correlation when r = - 1 it means there is a perfect negative correlation and when $r = 0$ there is no relationship between the variables. The coefficient of correlation not only describes the magnitude of correaltion but also its direction the above formula can be transformed into another formula were standard deviations of the two series need not be calculated and which is easier to apply.

$$r = \frac{\sum xy}{N\sqrt{\frac{\sum x^2}{N}}\sqrt{\frac{\sum y^2}{N}}} \quad \text{where} \quad X = (X - X'),\ Y = (Y - Y')$$

where $r = \dfrac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$

Assumptions Underlying Karl Pearson's Correlation Coefficient:

(i) The variables x and y under study are linearly related in other wards the scatter diagram of the data will give a straight line curve.

(ii) Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution. For example the variable relating to ages, heights, weight, supply, price etc.

(iii) The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion. In other word : cause and effect relationship exists between different forces operation on the items of the two variable series these forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion then there cannot be any correlation between the variables under study.

Example:

The correlation coefficient between

(a) The series of heights and incomes of individuals over a period of time

(b) The series of marriage rate and the rate of agricultural production in a country over a period of time.

(c) The series relating to the size of the shoe and inteligence of a group of individuals.

Example 2:

Calculate Karl Pearson's coefficeint of correlation from the following data:

Roll No. of Students	1	2	3	4	5	6	7	8
Marks in Statistics	65	66	67	67	68	69	70	72
Marks in Accountancy	67	68	65	68	72	72	69	71

Solution:

Let marks in statistics be denoted by x and accountancy by y.

R.N.	**X**	**$X - X'$**	**X^2**	Y	**$Y - Y'$**	**Y^2**	**X Y**
1	65	- 3	9	67	- 2	4	6
2	66	- 2	4	68	- 1	1	2
3	67	- 1	1	65	- 4	16	4
4	67	- 1	1	68	- 1	1	1
5	68	0	0	72	+ 3	9	0
6	69	1	1	72	+ 3	9	3
7	70	2	4	69	0	0	0
8	72	4	16	71	+ 2	4	8
	$\sum x = 544$	$\sum X - X' = 0$	$\sum x^2 = 36$	$\sum y = 552$	$\sum Y - Y' = 0$	$\sum y^2 = 44$	$\sum xy = 24$

$$r = \frac{\sum xy}{N\sigma_x \sigma_y}; \sqrt{x} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{36}{8}} = \sqrt{4 \cdot 5} = 2 \cdot 121$$

$$\sqrt{y} = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{44}{8}} = \sqrt{5 \cdot 5} = 2 \cdot 345$$

$$= \frac{24}{8 \times 2 \cdot 121 \times 2 \cdot 345} = + 0 \cdot 603$$

Calculation of r by second formula

$$r = \frac{24}{\sqrt{36 \times 44}} = + 0 \cdot 603$$

Correlation of Grouped Data:

When the number of observations is large the data are often classified into two way frequency distribution, otherwise called as bivariate frequency distribution since it shows the frequency distribution of two related variables. The class intervals of y are listed in the column headings and those of x are listed in rows of the table. The frequencies for each cell of the table are determined by tallying the frequency of the distribution of a single variable.

The formula for calculating the coefficient of correlation in such case

$$r = \frac{N \sum fdx\, dy - \sum fdx \cdot \sum fdy}{\sqrt{N \sum fdx^2 - (\sum fdx)^2}\ \sqrt{N \sum fdy^2 - (\sum fdy)^2}}$$

(OR)

$$r = \sum fdx\, dy - \frac{\frac{\sum fdx \sum fdy}{N}}{\sqrt{\sum fdx^2 - \frac{(\sum fdx)^2}{N}} \times \sqrt{\sum fdy^2 - \frac{(\sum fdy)^2}{N}}}$$

Note:

This formula is the same as the one discussed above for assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Steps:

(i) Take the step deviations of x and y series from assumed means and denote them by dx and dy.

(ii) Multiply dx and dy and the frequency of the cell. The value fdxdy may be put at the top on right or left hand side of the cell.

(iii) Add all values of $f\,dxdy$ as calculated in step (ii) and thus obtain the value of $\sum f\,dxdy$.

(iv) Multiply the frequency of x series with the step deviation and total all such product to get $\sum fdx$. Similarly get $\sum fdy$.

(v) Take the square of the step deviations of x series and multiply them with the frequency and total the products to get $\sum fdx^2$. Similarly get $\sum fdy^2$.

(vi) Substitute the values so obtained in the formula given above to get the value of r.

The following example would clarify the above points.

Example 3:

From the following table given below calculate the coefficient of correlation between family income and food expenditure of 100 families.

Bivariate Table

Food Expenditure in %	Family Income				
	2000 - 3000	3000 - 4000	4000 - 5000	5000 - 6000	6000 - 7000
10 - 15	---	---	---	03	07
15 - 20	---	04	09	04	03
20 - 25	07	06	12	05	---
25 - 30	03	10	19	08	---

$$r = \sum fdxdy - \frac{\frac{\sum fdx \sum fdy}{N}}{\sqrt{\sum fdx^2 - \frac{(\sum fdx)^2}{N}}\ \sqrt{\sum fdy^2 - \frac{(\sum fdy)^2}{N}}}$$

$$r = \frac{-48 - 0 \times 100}{\sqrt{120 - \frac{(0)^2}{100}}\ \sqrt{200 - \frac{(100)^2}{100}}}$$

$$r = \frac{-48}{10 \cdot 95 \times 10} = \frac{48}{109 \cdot 5} = -0 \cdot 4383$$

Properties of correlation coefficient:

(i) The correlation coefficient is a pure number i.e. if has no unit.

(ii) The correlation coefficient P(or r) ranges from - 1 to 1

(iii) The correlation between two variables is known as simple correlation or correlation of zero order.

(iv) It is not affected by coding of variables or variate values.

(v) The relation between the correlation coefficient 'r' and the two regression coefficients byx and bxy is

$$r = \sqrt{bxy \times byx}$$

(vi) The sign of r will be the same as that of byx or bxy.

(vii) If the two variables are independent the correlation coefficient between them is zero but the converse is not true.

(viii) If $P(\text{or } r) = 0$ it shows that the relation ship between the variables x and y is not linear.

(ix) The relationship between the correlation coefficient with a regression coefficient is

$$\beta_{yx} = P\frac{\sigma_y}{\sigma_x}$$

$$\beta_{xy} = P\frac{\sigma_x}{\sigma_y}$$

(x) Arithmatic mean of two regression coefficient is always greater than the positive correlation coefficient between the vaiables symbolically.

$$\frac{1}{2}\left(P\frac{\sigma_y}{\sigma_x} + P\frac{\sigma_x}{\sigma_y}\right) \geq P$$

$$\sigma_y^2 + \sigma_x^2 \geq 2\,\sigma_x\sigma_y$$

$$\left(\sigma_y - \sigma_x\right)^2 \geq 0$$

This confirms the property.

(xi) The quantity $\sqrt{1 - r^2}$ is referred to coefficient of nienation.

Merits and Demerits of Karl Pearson's Coefficient of Correlation:

Karl Pearson's method is the most popular method for measuring the degree of relationship among all the mathematical methods used for the same. The merit of this coefficient is that it gives the degree of the relationship among the variables as well as the direction of the correlation.

However it suffers from some limitations also these are:

(1) The correlation coefficient always assumes linear relationship even though it may not be there

(2) It is liable to be misinterpreted as a high degree of correlation does not necessarily mean very close relationship. Thus great care should be exercised in interpreting the values.

(3) The values of the coefficient is unduly affected by the extreme items.

(4) As compared to other methods it is tedious to calculate.

18.3 TESTING OF SIGNIFICANCE OF CORRELATION COEFFICIENT:

Onece the correlation coefficient has been calculates from the sample data we are interested in knowing whether the association established between the variables is enough to make predictions about the population or with what confidence we can make a statement about the association between variables?

To answer the said questions the calculated sample correlation coefficient is to be tested for its significance against a null hypothesis of population correlation beeing equal to zero by using the t statistic.

Testing Against a $H_0 = 0$:

One sample procedure used in estimating the true population correlation 'P' by using the value of sample correlation coefficient is the use of charts specially prepared as shown in the following figure. The chart being prepared separately for different confidence intervals of 95's and 99% level would be helpful to estimate the upper and lower bounds of true population parameter.

t - statistic:

$$t = \frac{r}{\sqrt{\left(1 - r^2\right) / \left(n - 2\right)}}$$

r = sample correlation coefficient

(n - 2) = degree of freedom

t = number of standard deviation that 'r' is away from 'zero' the calculated 't' value is compared to the table value for (n - 2) degrees of freedom at any desired level of significance if calculated t value is lessthan table value we accept the null hypothesis (H_0: The population

correlation coefficient equals to zero) stating that the association between variables in not significantly different from zero.

Suppose the correlation coefficient between sales and advertising for 12 observations is found to be 0.7343 in a particular industry. Then the following calculation determines the significance or otherwise of such a correlation.

Testing of Hypothesis:

H_0 : Null hypothesis : calculated correlation coefficient is not significantly different from zero.

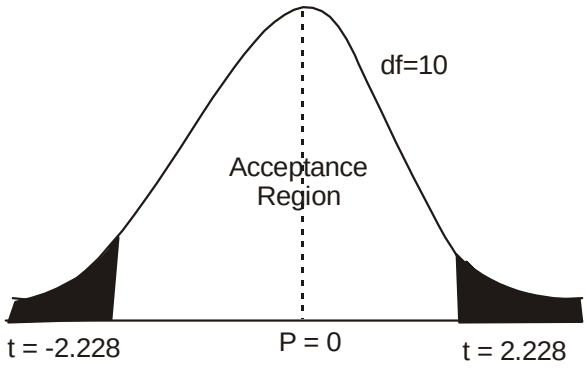
significance level : 5%

Degree of freedom $= (n - 2) : (12 - 2) = 10$

Calculated 't' value

$$t = \frac{r}{\sqrt{(1 - r^2) / (n - 2)}}$$

$$= \frac{0 \cdot 7343}{1 - 0 \cdot 5392} = 3 \cdot 421$$



Decision:

If t calculated value is greater than t table value of $\pm 2 \cdot 228$ to 10 df rejected H_0

Since 't' calculated > 't' table value

3.421 > 2.228 reject H_0

H_1 : The relationship between sales and advertisement is significant

18.4 Rank Correlation:

Let us suppose that a group of n individuals is arranged in order or merit of proficiency in possession of two characteristics A and B. These ranks in the two characterisitcs will in general be different. For example, if we consider the relation between intelligence and beauty it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i) ; $i = 1, 2, \cdots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coeficient of correlation between A and B for that group of individuals.

Assuming that no two individuals are breakeated equal in either classification each of the variables x and y takes the values 1, 2, . . ., n

Hence $\bar{x} = \bar{y} = \frac{1}{n}(1+2+3+\cdots+n) = \frac{n+1}{2}$

$$\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2} = \frac{1}{n}\left(1^2 + 2^2 + \cdots + n^2\right) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6} - \left[\frac{(n+1)}{2}\right]^2$$

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

$$\sigma_{x^2} = \frac{n^2-1}{12} = \sigma_{y^2}$$

In general $x_i \neq y_i$ let $d_i = x_i - y_i$

$$d_i = (x_i - \bar{x}) - (y_i - \bar{y})$$

Squaring and summing ever i from 1 to n we get

$$\sum d_i^2 = \sum\left\{(x_i - \bar{x}) - (y_i - \bar{y})\right\}^2$$

$$= \sum (x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 - 2\sum(x_i - \bar{x})(y_i - \bar{y})$$

Dividing bothe sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \sigma_{x^2} + \sigma_{y^2} - 2\ COV\left(x, y\right)$$

$$= \sigma_{x^2} + \sigma_{y^2} - 2P\sigma_x\,\sigma_y$$

Where P is the rank correaltion coefficeint between A and B

$$\frac{1}{n}\sum d_i^2 = 2\,\sigma_{x^2} - 2P\sigma_{y^2} = 1 - P\,\frac{\sum d_i^2}{2\pi\sigma^2 x}$$

$$P = 1 - \frac{6\sum di^2}{n\left(n^2 - 1\right)}$$

P = Coefficient of rank correlation

d = difference between ranks of each pair of observation

n = number of pairs observations

Correlation based on ranks:

This method of rank correlation would simply the process of computing from a very loose set of data for each the two variables. This method is called spearman's rank correlation. In honor of a statistician who developed it. The coefficient is expressed as 'rho' measures the strength of an increasing relationship between variable and it treats all observations equally so out liers are no more effect on its value then any other observation have.

Example 4:

Two women customers are randomly selected in asuper market and are asked to test five alternative shampoos and rank them in order of preference from 5 (best) to 1 (least desirable) the results are as follows:

Brand of shampoo

	A	B	C	D	E
Customer 1	1	4	3	2	5
Customer 2	3	5	2	1	4

Calculation of Rank correlation coefficient

	Ranks of Customer 1(R_1)	Ranks of Customer 2(R_2)	Difference in Ranks (d)	d^2
A	1	3	- 2	4
B	4	5	- 1	1
C	3	2	1	1
D	2	1	1	1
E	5	4	1	1
				$\sum d^2 = 8$

$$P = 1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)}$$

$$= 1 - \frac{6 \times 8}{5\left(5^2 - 1\right)} = 1 - 0 \cdot 4 = 0 \cdot 60$$

There exists some consistency in ranking the brands by customers.

Example 5:

Ten competitors in a beauty contest were ranked by three judges int he following order.

First Judge	1	6	5	9	2	3	4	10	7	8
Second Judge	3	5	8	7	4	9	2	1	6	10
Third Judge	6	4	8	9	1	3	2	10	7	5

Use the method of rank correlation to determine which pair of judges has the nearest approach to common tasts in beauty.

Solution:

With a view to find out which pair of judges have the nearest approach to common taste in beauty we will compare the rank correlation between the judgements of

(i) 1st and 2nd judge

(ii) 1st and 3rd judge

(iii) 2nd and 3rd judge

The rank given by the three judges would be denoted by R_1, R_2 and R_3.

R_1	R_2	R_3	$R_1 - R_2$ $d1 \cdot 2$	$R_1 - R_3$ $d1 \cdot 3$	$R_2 - R_3$ $d2 \cdot 3$	d_{12}^2	d_{13}^2	d_{20}^2
1	3	- 6	- 2	- 5	- 3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	8	- 3	- 3	0	9	9	0
9	7	9	2	0	- 2	4	0	4
2	4	1	- 2	1	3	4	1	9
3	9	3	- 6	0	6	36	0	36
4	2	2	2	2	0	4	4	0
10	1	10	9	0	- 9	81	0	81
7	6	7	1	0	- 1	1	0	1
8	10	5	- 2	3	5	4	9	25
						148	54	166

We have n = 10

Spearman's rank correlation coefficient are given by

$$r_{1\cdot2} = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 148}{(10)^3 - 10}$$

$$= 1 - \frac{888}{990} = 1 - 0 \cdot 896 = 0 \cdot 106$$

$$r_{1\cdot3} = 1 - \frac{6 \times 54}{(10)^3 - 10}$$

$$= 1 - \frac{324}{990} = 1 - 0 \cdot 327 = 0 \cdot 673$$

$$r_{2\cdot3} = 1 - \frac{6\sum d^2}{N3 - N} = 1 - \frac{6 \times 166}{(10)^3 - 10} = 1 - \frac{996}{990} = 1 - 1 \cdot 006 = -0 \cdot 006$$

Since $r_{1.3}$ is the highest the pair of 1st and 3rd judges has the nearst approach to common taste in beauty.

When ranks are not given:

Rank formula can be used even if we are dealing with variables which can be measured quantitatively. If we are given the actual data (not the ranks) we have to convert the data into ranks. The highest (or smallest) value is given rank 1 i.e., either ranking is to be done by discending or ascending order. Whatever may be the order of ranking it has to be uniformly followed in case of both the variables.

Example 6:

Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Advertisement cost (in '000 Rs.)	30	68	65	92	85	76	27	98	36	79
Sales (in lakhs)	47	59	59	86	63	68	60	91	51	84

Solution:

Let x denote the advertisement cost ('000 Rs.) and y denote the sales (lakhs).

x	y	R_x	R_y	d	d^2
39	47	8	10	- 2	4
68	54	6	8	- 2	4
65	59	7	7	0	0
92	86	2	2	0	0
85	63	3	5	- 2	4
76	68	5	4	1	1
27	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
79	84	4	3	1	1
				$\sum d = 0$	$\sum d^2 = 30$

Here n = 10

Therefore

$$r_s = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 30}{10^3 - 10}$$

$$= 1 - \frac{180}{990} = 1 - 0 \cdot 181 = 0 \cdot 819$$

Equal Ranks:

In some cases it may be found necessary to rank or more individuals or entries as equal. In such cases common ranks are assigned to the repeated items. This common ranks are the arithmetic mean of ranks which these items would have got if they were different from each other and the next will get the rank next to rank used in computing the common rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank 5 + 6/2, that is 5.5, if 4 are ranked equal at 6th place they are given the rank. $6 + 7 + 8 + 9/4 = 30/4 = 7 \cdot 5$ and the next rank would be 10th where equal ranks are assigned to some entries, an adjustment in the above formula is made calculating rank correlation coefficient.

The adjustment consists of adding $1/12\left(m^3 - m\right)$ to the value of $\sum d^2$ where m stands for the number of items whose ranks are common. If there are more than one such group of items in common, this value is added as many times as the number of such groups. The formula can thus be written as

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \cdot \cdot \cdot\right\}}{N^3 - N}$$

Example 7:

A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair here approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below.

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

Solution:

A	B	Rank of A	Rank of B	d	d^2
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	- 1	1
14	26	10	4	6	36
30	23	1.5	5	- 3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	- 3	9.00
30	20	1.5	7	- 5.5	30.25
20	16	7	9.5	- 2.5	6.25
28	11	4	11	- 7	49.00
11	21	11	6	5	25.00
				$\sum d = 0$	$\sum d^2 = 225$

In the A series the value 30 occurs twice. The common rank assigned to each of those values is 1.5 the arithmetic mean of 1 and 2 the ranks which these observations would have taken if they were different. Thenext value 29 gets the next rank i.e. 3. Again the value 19 occurs twice. The common rank assigned to it is 8.5. The arithmetic mean of 8 and 9 and the next value 14 gets the rank 10. Similarly in B series the value 16 occurs twice and the common rank assigned to each is 9.5 the arithmetic mean of 9 and 10. The next value 11 gets the rank 11.

Hence we see that in the A series item 19, 30 are repeated, each occurs twice and in B series the item 16 is repeated. Thus in each of the three cases m = 2. Hence on applying the correction factor $m^3 - m/12$ for each repeated items we get

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{2^3-2}{12} + \frac{2^3-2^3-2}{12} + \frac{2^3-2}{12}\right]}{11^3 - 11}$$

$$= 1 - \frac{6 \times 226 \cdot 5}{11 \times 120} = 1 - 1 \cdot 0225$$

$$= -0 \cdot 0225$$

Merits and Demerits of the rank method:

Merits:

1. This method is very simple to calculate and to understand.
2. Where the data are of a qualitaitve nature like beauty honesty intelligence etc this method can be employed usefully.
3. When the ranks of different item values in the variables only are given this is the only method for finding out degree of correlation.

4. If it is desired to use this formula when actual values are given ranks can be ascertained and correlation can be found out.

5. Since in this method $\sum d$ or sum of the differences provides a check on calculation

6. It can be interpreted in the same way like Karl Pearson's coefficient

Demerits:

1. This method cannot be used for finding out correlation in a grouped frequecny distribution.

2. It can be conveniently used only when n is small say 30 or less if it exceeds 30 the calculations becomes quite tedious and require a lot of time.

3. As all the information concerning the variables is not utilised this method lacks precision as compared to pearson's method.

Concurrent Deviation Method:

This method is one of the ways of ascertaining the coefficient of correlation by an extremely simple calculation. It is based on the direction of change or variation in the two paired variables. In this method correlation is calculated between the direction of deviation and not their magnitude. In majority of Pearson's coefficient with much less calculations.

To calculate the coefficient of concurrent deviations the deviations are not calculated from an average or assumed or moving average but only their direction from the preceding item and only the direction of the deviation (i.e. positive or negative) and not the extent of deviation are considered. The formula calculation of coefficient of concurrent deviation is given below:

Coefficent of concurrent deviation or

$$r_c = \pm \sqrt{\pm \left(2c - n\right)/2}$$

where r_c stands for coefficient of concurrent deviation.

c stands for the number of pairs of concurrent deviation and n for number of pair observations. The value of this coefficient of correlation also varies between +1 to -1. The plus and minus signs given in the formula should be carefully noted. If the value of (2c-n)/n is negative its square cannot be calculated and so a minus sign is placed before the sign of the root so that the square root may be calculated and the minus sign may be kept before the value of the coefficient of correlation.

The steps in the calculation of this coefficient are:

(1) Examine the fluctuation of each series and find whether each item increses or diminishes in comparison with the item just preceding. All increases are noted as plus and all decreases as minus. If here is neither increase or decrease the direction is zero. The direction of change on both series is denoted by dx and dy respectively.

(2) Multiply dx and dy and determine the value of C which would be the number of positive product of duty i.e. $\left(-x-\right)$ or $\left(+x+\right)$

(3) Count number of paired observation 'n'

(4) Use the formula given below to obtain the value of the coefficient or r_c

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

We will apply the formula in the following example to illustrate the steps.

Example 8:

The following are the marks obtained by a group of 10 students in Economics and Statistics.

Students:	1	2	3	4	5	6	7	8	9	10
Economics:	10	36	98	25	75	85	91	65	68	34
Statistics:	49	39	92	60	68	62	86	58	53	47

Calculate r_c by the method of concurrent deviation

Solution:

Students	**Marks in Economics**	**Deviation from Preceding item (dx)**	**Marks in Statistics**	**Deviation from preceding item (dy)**	**dxdy**
1	10	-	49	-	-
2	36	+	39	-	-
3	98	+	92	+	+
4	25	-	60	-	+
5	75	+	62	+	+
6	85	+	62	-	-
7	91	+	86	-	+
8	65	-	58	-	-
9	68	+	53	-	-
10	34	-	47	-	+
	N = 9		N = 9		C = 6

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

$$r_c = \pm\sqrt{\pm\frac{2(6)-9}{9}}$$

$$r_c = \sqrt{\frac{12-9}{9}}$$

$$= \sqrt{3/9} \ = 0\cdot 58$$

Merits and Demerits of concurrent Deviation Method :

Merits :

1) As compared to other methods it is the simplest and easiest.
2) When the number of items is very large, this method may be used to form a quick idea about the degrees of relationship before making use of complicated methods.

Demerits :

1. The Chief demerit is that it is only a very rough indicator of correlation.
2. It does not differentiate between small and big obviation i.e., an increase from 10 to 11 is given the same weight as to 10 to 10,000 i.e., it takes into consideration the direction of change and not the magnitude. It should be however, remembered that the results obtained by this method are not very different from those obtained by the use of Karl Pearson's coefficient in the case of short term oscillation only.

18.5 TESTING THE SIGNIFICANCE OF AN OBSERVED PARTIAL CORRELATION COEFFICIENT:

$P = 0$ (sawkin's method) the sample correlation coefficient of order k observed in a sample of size n from a multivariate normal population Prof R.A. Fisher proved that under the null hypothesis $H_0 : P_{12\cdot 34\ \cdot\ \cdot\ \cdot (K+2)} = 0$ i.e. the population correlation coefficeint is zero. The sampling distribution $(k+2)$ is same as that of r_{12} under the null hypothesis with sample size n reduced by K. Hence the test of for H_0 becomes:

$$t = \frac{r_{12\cdot 34\ \cdot\ \cdot\ \cdot (K+2)}}{\left[1-r^2_{12\cdot 34\ \cdot\ \cdot\ \cdot (k+2)}\right]^{1/2}} \cdot \sqrt{(n-k-2)}$$

with following student's 't' - distribution with $(n-k-2)$ d.f. .....

18.6 Auto correlation and time series:

An important guide to the properties of a time series is provided by a series of quantites called sample auto correlation coefficients or serial correlation coefficient. which measures the corelation between observations at different distance apart. These coefficients often provide insight into the probability model which generate the data. The simple auto correlation coefficient is similar to the ordinary correlation coefficient between two variables (x) and (y) except that it is applied to a single time series to see if successive observations are correlated.

Correlation: A useful aid in interpreting a set of auto correlation coefficients is a graph called a correlation and it is plotted against the lg (k) : where is the auto correlation coefficient at lg (k). A correlation can be used to get a general understanding on the following aspects of our time series.

(1) **A random series:** If a time series is completely random then for Large (N); will be approximately zero for all non - zero values of k.

(2) **Short - term correlation:** Stationary series often exhibit short term correlation characterized by a fairly large value of 2 or 3 more correlation coefficients which while significantly greater than zero tend to get successively smaller.

(3) **Non stationary series:** If a time series contains a trend, then the values of will not come to zero except for very large values of the lag.

(4) **Seasonal Fluctuations:** Common auto regressive models with seasonal flutuation of periods are

$$X(t) = a + bX(t-s) + \sum t$$

and

$$X(t) = a + b \times (t - s) + c \times (t - 2s) + \sum t$$

where $\sum t$ is a white noise series.

$$X(26) = 14 \cdot 44 + 0 \cdot 715 \times (25)$$

$$= 14 \cdot 44 + 0 \cdot 715 (51 \cdot 0) = 50 \cdot 91$$

$$X(27) = 14 \cdot 44 + 0 \cdot 715 \times (26)$$

$$= 14 \cdot 44 + 0 \cdot 715 (50 \cdot 91) = 50 \cdot 84$$

$$X(28) = 14 \cdot 44 + 0 \cdot 715 \times (27)$$

$$= 14 \cdot 44 + 0 \cdot 715 (50 \cdot 84) = 50 \cdot 79$$

18.7 Summary:

In this lesson the concept of correlation or the association between two variables has been explained the correlation coefficient r may assume values between - 1 and 1 the sign indicates whether the association is direct (+v) or inverse (-ve). If r equal to 1 indicates perfect association. While a value of zero indicates no association lenearly tests for significance of the correaltion coefficient are explained. Finally the concept of auto correlation and time series and defined with an example.

18.8 Exercises:

1. Define correlation coefficient
2. Give the formula for sample correlation coefficient r
3. Discuss the properties of correlation coefficient
4. What is the uses of correlation
5. Explain different types of correlation coefficient
6. Expalin karl pearson's coefficient of correlation method.
7. Assumptions underlying karl pearson's correlation coefficient
8. Merits and demerits of karl pearsons correlation coefficient
9. Explain rank correlation method

18.9 Reference Books:

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.
2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.
3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. K. MADHU BABU

Lesson - 19

REGRESSION

Objectives:

- To understand the concept of regression and estimate the relationship between development and independent variables.
- To estimate model parameters using least square method.
- To estimate standard error, t-test, F-test, Goodness of fit test.

Structure:

19.1 Introduction:

Regression analysis means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which is extensively used in almost all sciences - natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves. cost functions, productions and consumption functions.

The aim of regression model is to estimate as best as possible, the dependent variable Y from the independent vairbale(s) X. The line of average relationship is another name given to the regression line regression analysis will show us how to determine both the nature and the strength of a relationship between variables.

Before attempting such an analysis, a sequence of steps has to be followed, which includes.

STEPS FOR ANALYSIS :

(1) Determining type of relationships between variables and identifying independent and dependent variables.

(2) Establishing or guessing the forms of relationships between dependent and independent variables.

(3) Selecting and adopting appropriate functional models of relationship and estimating the parameters, which would help in forecasting.

(4) Testing the appropriateness of the functional relationship established, so as to enable its use for forecasting purposes.

What is regression?

Regression is a mathematical measure of the averages relationship between two or more vairables. The relationships between rainfall and agricultural production, consumer expenditure and disposable income, science entrance examination score and employee performance, interest rates and careful market activity, students performances in quantitative papers and problem solving ability are few exampels to have tendency of dependence on each other. Such relationships could be termed linear (the value of dependent variable increases by a constant absolute amount for a unit change in independent variable) or simple (when the study is limited to only between two variables).

We express the basic relationship between two variables in the following function form

$$Y = f(X)$$

It states that the value of the variable Y is a function of the value of the variable X. When it is assumed to be with linear relationship it is denoted by

$$Y = a + bX$$

where Y is dependent variable, a is intercept, b is sloap of the line and X is independent variable. Multiple Regression of

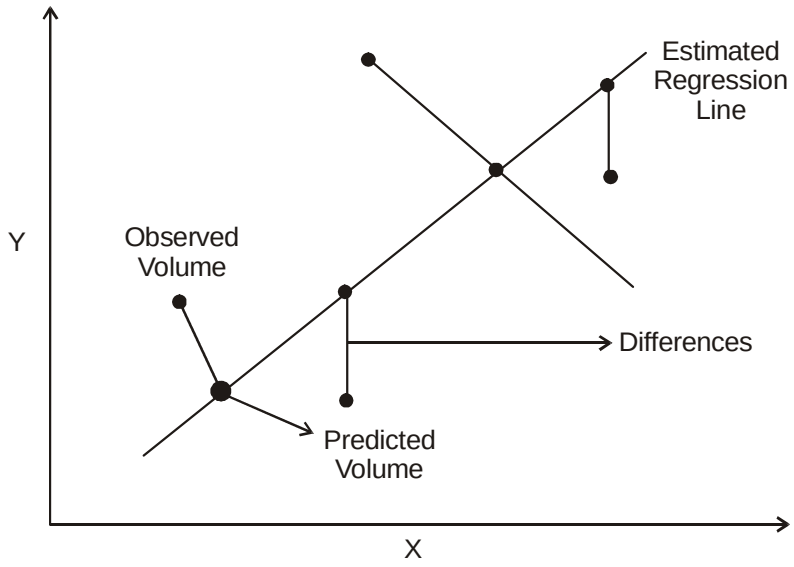
$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n$$

Where Y is dependent variable $X_1, X_2, \cdots, X_n$ are independent variables and $b_1, b_2, \cdots, b_n$ are regression coefficient of respective variables.

19.2 METHOD OF LEAST SQUARE

FITTING OF A STRAIGHT LINE :

Whenever we draw a estimated regression line, not all points will lie on the regression line. Some will be above it and some will be below. The difference between any point and the corresponding point on the regression line is called the deviation from the line. It represents the difference between the predicted value and the actual value. The least square method fits the regression line in such a way that the sum of squares of these deviations as small as possible.



Differences between Y values and the estimated regression lines.

Let us suppose that in the Bivariate distribution $(X_i, Y_i); i = 1, \cdots, n$, y is independent and x is the independent variable?

Let the line of regression of Y on X be

$$Y = a + bx$$

So the estimated line would be $\hat{Y} = a + bx + d$ where d is deviation or difference.

We have to minimise the sum of squares (SS) of deviation i.e.,

$$\underset{a,b}{\text{Min}}\ \Sigma d_i^2 \text{ or } \underset{a,b}{\text{Min}}\ \Sigma (y_i - a - bx_i)^2$$

We get the following normal equations

$$\begin{cases} \dfrac{\partial}{\partial a}\Sigma (y_i - a - bx_i)^2 = 0 \\ \dfrac{\partial}{\partial b}\Sigma (y_i - a - bx_i)^2 = 0 \end{cases}$$

or
$$\begin{cases} 2\Sigma (y_i - a - bx_i)(-1) = 0 \\ 2\Sigma (y_i - a - bx_i)(-x_i) = 0 \end{cases}$$

or
$$\begin{cases} \Sigma y_i - na - b\Sigma x_i = 0 \\ \Sigma y_i x_i - a\Sigma x_i - b\Sigma x_i^2 = 0 \end{cases}$$

Solving them we get

$$a = \frac{1}{n}(\Sigma y - b\Sigma x) = \bar{Y} - b\bar{X}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

So the equation of the estimated regression line is

$$\hat{Y} = a + bx$$

where $$b = \frac{n(\Sigma xy) - (\Sigma y)(\Sigma x)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{1}{n}(\Sigma y - b\Sigma x)$$

The straight line that best fits a set of data points according to the least squares criterion is called the regression line where as the equation of the regression line is called the regression equation.

Example 1 : From the following data obtain the two regression equations using the method of least square :

X	1	6	5	2	1	1	7	3
Y	6	1	0	6	1	2	1	5

Solution :

X	Y	X^2	Y^2	XY
1	6	1	36	6
6	1	36	1	6
5	0	25	0	0
2	6	4	36	12
1	1	1	1	1
1	2	1	4	2
7	1	49	1	7
3	5	9	25	15
$\Sigma X = 26$	$\Sigma Y = 22$	$\Sigma X^2 = 126$	$\Sigma Y^2 = 104$	$\Sigma XY = 49$

Regression equation Y on X $Y = a + bX$

To get the values of a and b the following two normal equations are used :

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values

$22 = 8a + b(26)$ ---------- (1)

$49 = a(26) + b(126)$ ---------- (2)

Multiplying equation (1) by 13 and equation (2) by 4 deducting (4) from (3) we get

$286 = 104a + 338b$ --------- (3)

$196 = -104a + 504b$ --------- (4)

$90 = -166b$

or $b = -0.54$ (approx)

substituting 'b' in equation (1) we get

$22 = 8a + 26(-0.54)$

$= 8a - 14.04$

$8a = 22 + 14.04$

$a = 36.04/8$

$= 4.51$ (approx)

$Y = 4.51 - (0.54)X$. This is the regression equation of Y on X.

Similarly the regression equation of X on Y is $X = a + by$. The two normal equations are

$$\Sigma X^2 = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values in the above equations

$26 = 8a + b(22)$ ---------- (1)

$4a = 22a + b(104)$---------- (2)

Multiplying equation(1) by 11 and (2) by 4 we get

$286 = 88a + 242b$ ----------(3)

$196 = 88a + 416b$ ----------(4)

Deducting (4) from (3) we get

90 = - 174b or

b = - 0.52 (approx)

Substituting b value in equation (1) we get

26 = 8a + 22(-0.52)

26 = 8a + (- 11.44) or

a = 26 + 11.44/ 8

= 4.68

Therefore X = 4.68 - (0.52) Y. This is the regression equation of X on Y.

19.2.1 DEVIATIONS TAKEN FROM ARITHMETIC MEANS OF X AND Y

The regression equations are written as follows :

Regression equations of X on Y : $(X - \bar{X}') = r\ \sigma x / \sigma y\ (Y - \bar{Y}')$

$\bar{X}$ is the mean of X series and $\bar{Y}$ is the mean of Y series.

r is the coefficient of correlation between X and Y series and σx and σy are the standard deviations of X and Y series respectively.

$r\sigma x / \sigma y$ is called the regression coefficient of X on Y and denoted by bxy. Thus

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{N\sigma x\,\sigma y} \times \frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{N\sigma y^2}$$

$$= \frac{\Sigma xy}{n\Sigma y^2 / n} = \frac{\Sigma xy}{\Sigma y^2}$$

Thus instead of finding out the values of r, $\sigma x, \sigma y$ we can directly find out the value of bxy by dividing the product of the deviations of X and Y series from their respective means by the sum of the squares of the deviations of the Y series from its mean.

Similarly regression equation of Y on X.

$(Y - \bar{Y}') = r\sigma y / \sigma x (X - \bar{X}')$

$r\sigma y / \sigma x$ is called the regression coefficient of Y on X or 'byx'

$r\sigma y / \sigma x$ can be calculated in the same way as above and its

value will be $\Sigma xy / \Sigma x^2$

Thus the two regression equations can be rewritten as follows :

(i) Regression equation of X on Y

$$X - \bar{X}' = \Sigma xy / \Sigma y^2 (Y - \bar{Y}')$$

(ii) Regression equation of Y on X

$$Y - \bar{Y}' = \Sigma xy / \Sigma x^2 (X - \bar{X}')$$

19.2.2 REGRESSION COEFFICIENT

As discussed above bxy and byx are called as the regression coefficient of regression equation X on Y and Y on X. The regression coefficients bxy and byx possess some important properties. They are

1. The underroot of the product of two regression coefficients gives us the value of correlation coefficients. Symbolically

$$\gamma = r\sqrt{bxy \times byx}$$

Proof : $bxy = r\sigma x / \sigma y$ and $byx = r\sigma y / \sigma x$

$$bxy \times byx = r\sigma x / \sigma y \times r\ \sigma y / \sigma x = r^2$$

$$r = \sqrt{bxy \times byx}$$

Then the geometric mean of bxy and byx gives the value of coefficient of correlation.

2. Both the regression coefficients will have signs i.e., either they will be positive or negative. The reason is that so far as standard deviation is concerned they are always positive. Only coefficient of correlation can be either positive or negative. If regression coefficient are negative r will also be negative. For example if bxy = - 1.3 and byx = - 0.65.

$$r = \sqrt{-1.3 \times -0.65} = -0.92 \text{ not } 0.92$$

3. Since the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one or more; in other words both the coefficients cannot be greater than one.

The following example would illustrate the use of the above method of obtaining regression equation.

EXAMPLE 2 : From the following data, obtain the two regression equations and their correlation coefficient.

Sales	46	42	44	40	43	41	45
Purchase	40	38	36	35	39	37	41

Solution : Let us denote the sales by the variable X and purchase by the variable Y.

X	$X' = X - \bar{X}$	x^2	Y	$Y' = Y - \bar{Y}$	y^2	XY
46	+3	9	40	+2	4	+6
42	-1	1	38	0	0	0
44	+1	1	36	-2	4	-2
40	-3	9	35	-3	9	9
43	0	0	39	+1	1	0
41	-2	4	37	-1	1	2
45	+2	4	41	+3	9	6
$\Sigma x = 301$		$\Sigma x^2 = 28$	$\Sigma y = 266$		$\Sigma y^2 = 28$	$\Sigma xy = 21$

$$\bar{X}' = \Sigma x / n = 301 / 7 = 43$$

$$\bar{Y}' = 266 / 7 = 38$$

Regression equation of Y on X

$$Y - \bar{Y}' = byx(X - \bar{X}') \text{ where } byx = \Sigma xy / \Sigma x^2$$

Putting the values

$$Y - 38 = 21 / 28(X - 43)$$

$$Y = 0.75(X - 43) + 38$$

$$= 0.75X + 5.73$$

Regression equation of X on Y

$$X - \bar{X}' = bxy(Y - \bar{Y}') \text{ where } bxy = \Sigma xy / \Sigma y^2$$

$$X - 43 = 21 / 28$$

$$= 0.75(Y - 38)$$

$$= 0.75Y + 14.50$$

We have

$$r^2 = byx \cdot bxy = 0.75 \times 0.75$$

$$r = 0.75$$

19.3 STANDARD ERROR

The standard error of regression coefficient measures the deviation of scattered points of the observed values around regression line. Generally, it can be symbolised as

$$S_e = \sqrt{\frac{\Sigma\left(Y - \hat{Y}\right)^2}{n-2}}$$

where Y = Actual value of the dependent variable

$\hat{Y}$ = estimated value from estimating equation.

n = number of data points.

The deviation are average by (n - 2) but not 'n'. It is said that since the 'a' and 'b' values are obtained from sample data in estimating the regression, line, we are left with only (n - 2) cases. It is also called degrees of freedom.

To illustrate, let us consider the data provided by a land development bank.

EXAMPLE 3 : Rate of interest and Rural Deposits.

Calculation of Equation

The changes in rates of interest and the size of deposits outstanding with a Land Development Bank, during the period of 1980 - 90 in the case of XYZ Land Development Bank in a semi-urban area is as follows :

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

$$= \frac{1135.7 - 10 \times 8.83 \times 12.27}{794.2 - 10 \times 8.83^2}$$

$$= \frac{52.26}{14.51} = 3.60$$

$$a = \bar{Y} - b\bar{X}$$

= 12.27 - 3.60 (8.83)

= - 19.518.

Table : Calculation of Regression Inputs

Interest Rate X	Size of Rural (in Lakhs) Y	XY XY	X^2
7.00	7.50	52.5	49.0
7.50	7.75	58.1	56.3
7.80	8.00	62.4	60.8
8.50	8.50	72.3	72.3
8.75	9.00	78.8	76.6
8.75	12.00	105.0	76.6
9.00	15.00	135.00	81.0
9.50	16.50	156.8	90.3
10.50	17.50	183.3	110.3
11.00	21.00	231.0	121.0
Σ 88.03	Σ 122.75	Σ 1135.7	Σ 794.2

The results indicate that the increasing interest rates have been positively attracting more and more advances from rural areas. The regression coefficient indicates that one percent increase in interest rate mobilising 3.60 percentage points of rural deposits in the said area.

CALCULATION OF STANDARD ERROR

However, the accuracy of the prediction equation depends on how widespread the scatter points of Y values are on the regression line. In order to verify the reliability of the estimates, let us calculate the deviation $\hat{Y}$ from the actual Y at the given levels of X. $\hat{Y}$ can be done separately for each of the X values by substituting $\bar{X}$ 'a actual value in the following.

$$\hat{Y} = a + bX$$

Y Actual '000	$\hat{Y}$ Estimated '000	e^2
7.50	5.68	3.3124
7.75	7.50	0.0625
8.00	8.50	0.2500
8.50	11.10	6.7600
9.00	12.00	9.0000

12.00	12.00	0.0000
15.00	12.90	4.4100
16.50	14.70	3.2400
17.50	18.30	0.6400
21.00	20.10	0.8100
Σ 122.75	Σ 122.75	Σ 28.4849

Standard Error

$$S.E. = \sqrt{\frac{(Y - \hat{Y})^2}{n - 2}}$$

$$= \sqrt{\frac{28.4849}{8}} = 1.8869$$

The standard error score can be interpreted on a normal curve. Generally, we would be expecting about 95 per cent of Y values to fall within plus or minus two standard errors of regression line.

The fitted regression line is shown in thick line and the dashed lines above and below the regression line are probability units of the estimates of Y within one standard error. As is the case standard deviation, the larger the standard error of the estimate, the greater the scattering (or dispersion) of points around the regression line. If Se = O, the estimating equation could be perfect estimator and all data points would lie on regression line.

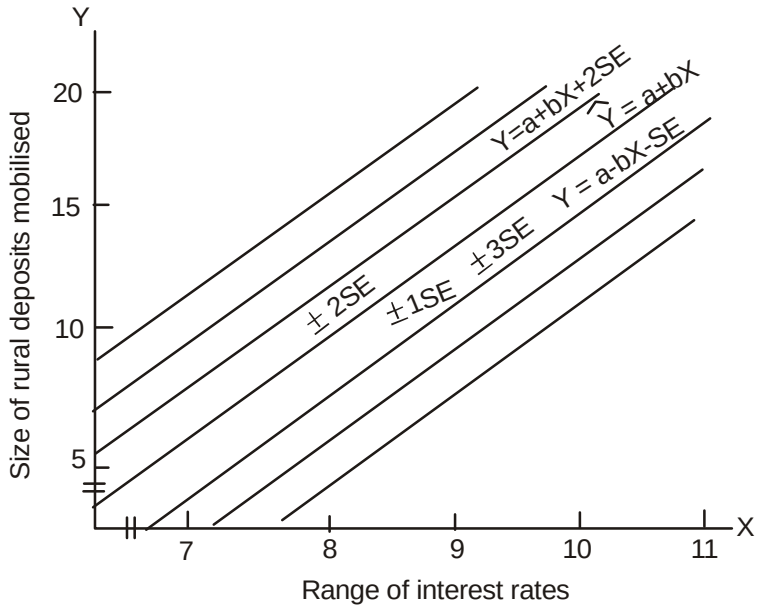


Fig. 1 : Observed and Predicted values with standard errors

Limits of Area under normal curve

The measure of standard error can be used and interpreted as standard deviation. If we assume that data follows normal distribution.

68% of the points would lie within ± 1 SE

95.5% of the points would lie within ± 2 SE

99.7% of the points would lie within ± 3 SE

19.4 'T' STATISTIC

TESTING THE SIGNIFICANCE OF REGRESSION COEFFICIENT

When there is no association between dependent variable Y and independent variable X, then the value of the slope B, of the regression line would be zero. Therefore, any evidence of relationship between Y and X should possess a significantly different 'B' value from zero. In other words, regression coefficient is to be tested for its signifiance or otherwise to validate a null hypothesis of B = 0. The 't' statistic to be adopted is :

$$t = \frac{b_1 - 0}{S_b}$$

where $b_1 =$ Value of the regression coefficient

when $S_b =$ Standard error of the slope coefficient of the sample.

$$S_b = \frac{S_e}{\sqrt{\Sigma X^2 - n\bar{X}^2}}$$

Calculation of 't' test

To illustrate, consider the earlier example of interest rates and rural deposits mobilised. To determine if the slope of the regression is statistically significant, we may determine the slope as zero or not. If a confidence level of 5 per cent is assumed, the critical t - score for (n - 2) degrees of freedom is $\pm$ 2.306. Hence, the decision rule is :

Reject the Hypothesis of B = O, when 't' score based on sample evidence is greater than $\pm$2.306. Otherwise Accept it.

$$S_b = \frac{S_e}{\sqrt{\Sigma X^2 - n\bar{X}^2}} = \frac{1.8869}{\sqrt{794.2 - 10 \times 8.83^2}}$$

$$= \frac{1.8869}{3.8093} = 0.4953$$

$$H_0 = B_1 = 0$$

$$H_1 = B_1 \neq 0$$

$$t = \frac{3.60 - 0}{0.4953} = 7.27$$

Since computed 't' value is greater than 2.306, the null hypothesis is rejected at 0.05 level of signifiance. Thus $B \neq 0$ therefore, we can conclude that there exists a significant relationship between interest rates and rural deposits mobilised.

19.5 F - STATISTIC

Testing the significance of the entire equation

Another statistical test for evaluating the significance of a simple regression analysis is the computation of F-statistic F test also called as analysis of variance (ANOVA) test reveals how the variation in one group can be the same as or different from the variation in a second group. in other words, we wish to compare variance associated with regression with that of the variance not associated with regression.

F ratio represents

$$\frac{\text{Sum of squares associated with regression} \div \text{Degrees of freedom}}{\text{Sum of squares not associated with regression} \div \text{Degrees of freedom}}$$

The numerator denotes the sum of squared deviations of the predicted variable about its mean (due to regression), the denomenator is the sum of squared deviations of differences between the actual and predicted values (due to error).

$$F = \frac{\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2 / (m-1)}{\Sigma\left(Y - \hat{Y}\right)^2 / n - m}$$

m = number of parameters.

Formula for F-test

The same can be written as :

$$F = \frac{b^2\left(\Sigma X^2 - n\bar{X}^2\right)}{S_e^2}$$

Calculation of F ratio

For the data in earlier illustration

$$F = \frac{3.60^2\left(794.2 - 10 \times 8.83^2\right)}{1.8869^2} = \frac{188.06}{3.56} = 52.82$$

If we wish to consider to test the H_0 (There is no evidence of significant linear relationship) the critical F score (from F tables at 5% level of signifiance) can be seen to be 5.03. The calculated F value is very high than the table value indicating the evidence of relationship in the said problem.

19.6 CO-EFFICIENT OF DETERMINATION - R^2 VALUE :

Goodness of fit test

The determination of 'Coefficient of determination', R^2 value is the test how closely the least squares regression line fits the data points in the scatter diagram. Statisticians also interpret it as the amount of variation in Y that is explained by the regression line. Consider the following figure. One actual Y value is taken at point $\bar{Y}'$. This $\bar{Y}'$ value, when compared to the average of Y values $\bar{Y}'$, shows the "Total deviation". Suppose we have used regression line to estimate the Y value at a given level of X, then the deviation of $\bar{Y}'$ from $\hat{Y}$ gives a partial deviation. Out of the total distance of $(Y' - \bar{Y})$ regression equation is found to have explained $(\hat{Y} - \bar{Y})$ distance and attributed such value to 'known' deviation and the remaining $(\bar{Y}' - \hat{Y})$ distance to unknown or unexplained deviation.

When one considers all the data points on a scatter diagram, the sum of total deviations can be expressed as

$\Sigma(y_1' - \bar{Y})^2$, it is squared to avoid the signs. Similarly explained variation can be denoted as

and unexplained part can be denoted as $\Sigma(Y_t - \hat{Y})^2$.

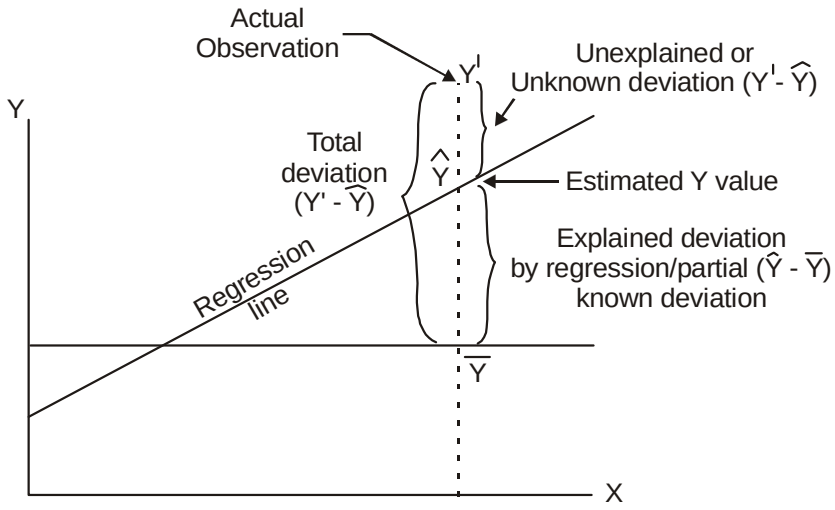


Fig. 2 : Details of explained and unexplained deviation relating to regression equations

Total Variance	=	Explained Variation	+	Unexplained Variation

$$\Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2 + \Sigma\left(Y_i - \hat{Y}\right)^2$$

Explained variation is equal to one minus unexplained variation

A regression equation can be said to be better fit only when the proportion of unexplained variation is less. Therefore, coefficient of determination is

$$r^2 = 1 - \frac{\Sigma\left(Y_i - \hat{Y}\right)^2}{\Sigma\left(Y_i - \bar{Y}\right)^2}$$

It gives the fraction of total variation that is explained.

To illustrate, consider the earlier example of interest rates and deposits mobilised. Instead of reworking the illustration, let us use a short-cut formula provided by statisticians. It is nothing but calculated 'r' correlation coefficient as per the procedure suggested in earlier lesson.

Formula for Correlation Coefficient (r)

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]}\ \sqrt{\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

and then $r^2 = r \cdot r$

Coefficient of determination is the square of correlation

This is only an extension to correlation formula. There is another direct formula, using regression constants.

Formula for r^2 using regression constants

$$r^2 = \frac{a\Sigma Y + b\Sigma XY - n\bar{Y}^2}{\Sigma Y^2 - n\bar{Y}^2}$$

where 'a' and 'b' are regression constants

n = number of data points

X = Value of independent variable

Y = Value of dependent variable

$\bar{X}$ = Mean of independent variable

$\bar{Y}$ = Mean of dependent variable

Using the above formula

$$r^2 = \frac{-19.518(122.75) + 3.60(1135.70) - 10(12.275)^2}{1722.06 - 10(12.275)^2}$$

$$= \frac{185.94}{215.25} = 0.86$$

Thus 86 per cent of the total variation in the deposits is associated with the regression on interest rates.

19.7 USING SIMPLE REGRESSION MODEL FOR FORECASTING

Applications of Forecasting

One of the basic objectives of learning regression analysis as a tool of managerial decision making is for purposes of 'forecasting'. Forecasting the demand for product, forecasting costs, forecasting return from securities/projects are some of the necessities for business managers. Then, how to use this regression analysis for forecasting purposes?

As we know, the regression equation established a relationship between an independent variable (X) and a dependent variable (Y).

Further, we can find the estimated value $\hat{Y}$ of dependent variable for changing values of independent variable. Given a particular new value of X, designated as X^* the estimated regression yields.

Formula for forecasting values

$$\hat{Y}^* = a + bX^*$$

where $\hat{Y}^*$ = Forecasted value of Y at the expected level of X.

X^* = Particular new value for X, for which the value of dependent variable to be estimated.

In order to validate the forecasted value of Y, it is customary to calculate the SE for the forecasted Y^* value for using the following formula.

Formula for testing the forecasted value

$$SE_{Y^*} = S_e\sqrt{\frac{1}{n} + \frac{(X_* - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}}$$

S_e = Standard error of the equation as a whole

$(X_* - \bar{X})^2$ = Refers to how far away X_*, is from the mean X.

$(X_i - \bar{X})^2$ = Refers to differences of n known X values.

In the above equation, if the new value of X, i.e., X, is equal to mean of the n known values of X, the equation yields lowest possible error for the estimate.

19.8 USES OF REGRESSION ANALYSIS

There are many uses of regression analysis.

(i) In many situations the dependent variable Y is such that it cannot be measured directly. But with the help of some concomitant or auxiliary variables as independent variables in a regression equation, Y can be estimated.

(ii) The effect of certain treatments can better be adjudged by estimating the effect of concomitant variables.

(iii) The length of a confidence interval of population mean can be reduced by considering the effect of dummy variables.

(iv) Regression equation is often used as a prediction equation.

(v) To measure the influence of independent variables on the dependent variable which may be having a casual relationship.

19.8.1 Examples of Use of Regression Analysis from different areas are given below :

(i) Height of a person at a given age can be estimated by finding the regression of height (Y) on age (X).

(ii) The yields of a crop can be predicted for different doses of a fertilizer, ofcourse within the non-toxic dose of fertilizer.

(iii) A scientist may be able to estimate the brain size on the basis of certain related external body measurements.

(iv) Future demand of food may be predicted with the help of regression models.

(v) Certain treatments can be compared accurately by eliminating the effect of certain variables like percentage germination per unit area, number of weeds per unit area etc.

19.8.2 LIMITATIONS OF REGRESSION ANALYSIS

With the help of regression the estimates are made but it should be noted that for estiamtion unless the assumption that has been taken remains unchanged since the equation was computed estimates may go wrong. Another point to be remembered is that the relationship shown in the scatter diagram may not be the same if the equation is extended beyond the values in computing the equation. For example, if we find a close linear relationship between yield of a crop and amount of fertilizer applied, it would not be logical to extend this equation beyond the limits of the experiment

for it is quite likely that if the amount of fertilizer were increased indefinitely, the yield would eventualy decline as too much fertilizer is applied.

19.9 SUMMARY

In this lesson the concept of regression analysis is understood. The estimation of the parameters of regression models accomplished by the least square method to minimize the sum of squares of the error for all the data points.

After the model is fitted to data the next logical question is to find out how good the quality of fit is. This question can best be answered by conducting statistical tests and determining the standard errors of estimates.

16.10 EXERCISES

1. Explain the concept of regression and indicate its utility in business forecasting.
2. Explain what is meant by the regression of one variable Y on another variable X. What is the relationship between regression coefficient and correlation coefficient?
3. What is testing of significance in the context of regression analysis? Explain the procedure relating to it.
4. What do you mean by standard error of an estimate? Give the expression for standard error of a linear equation. Also explain about the 'explained and unexplained' variations.
5. Explain how 't' test F - ratio, and r values would be helpful in testing the reliability of estimated equation.
6. A portfolio analyst of Chartered Bank of India is interested in the reliability and market sensitivity of a particular blue chip scrip order to evaluate this aspect, he has considered market price of the said scrip on the test saturday of the month for 12 months and Economic Times Index on the same days.

The data are as follows :

MPS of the scrip	234.5	255.0	202.0	312.0	350.0	310.0
ET Index	300.0	340.0	310.0	350.0	363.0	350.0
Contd	318.0	360.0	380.0	410.0	450.0	410.0
	340.0	350.0	380.0	400.0	400.0	400.0

Estimate a Regression Line.

7. A Delhi-based Instant Food Products manufacturing enterprise is estimating that the demand for its product would be high in those towns wherever the proportion of women employment is more. To validate this presupposition they have collected data 8 in towns and requested you to fit a linear line.

Demand for the Food Product (Tonnes)	200	360	410	320	450	510	600	650
Proportion of women employment (%)	24	26	32	30	45	40	35	35

8. A large petrochemical company has ten machines of similarly type. It is investigating the relationship between the annual cost of maintenance of these machines and their age. Figures for ten machines are as follows :

Machine No.	1	2	3	4	5	6	7	8	9	10
Age(years)	5	7	10	5	5	25	15	0	5	7
Maintenance cost (Rs '000)	20	15	25	10	15	50	15	40	10	10

Find least squares regression of maintenance cost on age.

9. The marketing executive of Bata, Power Shoe division is interested in evaluating the impact of continuous advertisement during last 12 months on sales. The past data are as follows :

Sales (Rs. Laks)	22	28	22	26	34	18	30	38	30
Advertising Expenditure (Rs. Lakhs)	0.8	1.0	1.6	2.0	2.2	2.6	3.0	3.0	4.0
Contd.				40	50	46			
				4.0	4.0	4.6			

Calculate a linear regression line and also test the hypothesis of no relationship between advertising and sales.

10. ECIL is estimating the demand for its computer hardware would be depending on the enrollment of students in different engineering colleges for MCA and DCA programmes. In order to forecast the demand for next five years the marketing executive is trying to fit a regression line for the said data.

Total value of Hardware sold (Rs. crores)	2.20	2.50	3.70	4.50	9.50	6.50	7.50
Total Number of students ('000) Joining MCA OCA courses	20	22	20	25	35	32	32

Fit a regression line and test the validity of the expected relationship.

19.11 REFERENCE BOOKS

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

Lesson Writer

Dr. K. MADHU BABU

Lesson - 20

TIME SERIES ANALYSIS

Objectives:

- After completion of this lesson, you should be able to :
- To understand the various components in a time series.
- To adopt the methodology of procedures in decomposing series components in different practical situation.
- To understand Box-Jenkin's Methodology
- To understand models of Auto regressive, Auto correlation and conelogram.

Structure:

20.1 Introduction:

Forecasting is a prerequistic for managerial decision making. Forecasting is important to firms as they face fluctuation in demand and inventory levels throughout the year; firm's sales are related to changes in economic environment in which it operates; and abrupt changes in political and economic conditions will have serious consequences at firm and industry levels. Its uses vary from determining the size of working capital requirements for a small-scale enterprise to estimating the annual sales of two wheelers in case of large sized manufacturing enterprise producing motor vehicles.

Forecasts can be made by considering a large number of factors that affect the results of managerial decisions. But identification of these factors, measurement and monitoring of their influence is impossible since the roles of these factors continuously change. Although the

determination of the factors that affect the future are uncertain, often the past data offers a good indication of what the future will hold. The time series analysis is one quantitative technique the decision makers use in considering and extracting meaningful information to make a decision about the future from past information.

A time series is series of measurements taken at successive time periods.

Examples of a time include daily share price quotations, weekly production figures, monthly inventory levels, annual turnover.

Uses of time series Analysis

There are several reasons to analyse the past performance. One reason is to detect patterns of change in statistical information over regular intervals of time. This examination is towards identification of seasonal variations and long term growth or decline. The second reason for studying time series data is to predict future behaviour. Although no crystal clear forecast be able to make, all that can be guessed is that the past patterns repeat themselves over time.

20.2 ANALYSING TREND

Long term trend represents the long term direction of the series. In order to identify this component, it is better to fit a line to a set of points of historical data. Such a line need not always be linear. We can adopt to fit certain non-linear shapes of lines like first-degree, exponential, modified exponential and Gompertz curves. However, we know fitting a straight line through the method of least squares. The other methods of analysis are similar. Different shapes of Trend lines can be observed in Fig.

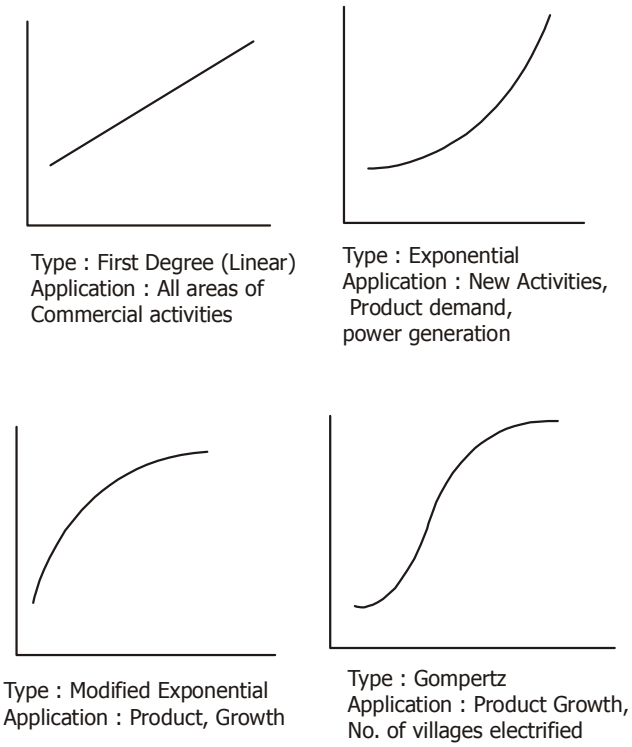


Fig. 20.1 Long term trend pattern curves

An analysis of trend component allows us to dscribe the historical pattern of past data project past trends into future. Further the study of trend of a time series makes it easier to study the impact of the other three components after eliminating the trend changes. Suppose we are interested in determining the seasonal sales of refrigerators, elimination of annual trend component gives more accurate idea of seasonal component. An analysis of this nature allows the manufacturer to adopt off-season discount sales to ensure continuous production schedule.

Estimating Trend

Estimating trend is the study of trend to determine the long term direction in time series. The determination of 'trend' would be of help to project it into future, or to eliminate it from original series. The estiamtion of trend should consider the length of the time series as lengthier projects require the use of a time series of longer duration.

The initial step of trend analysis is to chart the time series to observe the scatter plots. This chart gives the pattern of distribution of data, which enables to choose either linear or non-linear ways of measurement.

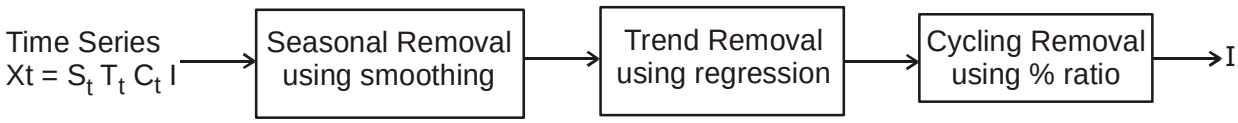
20.3 DECOMPOSITION ANALYSIS :

It is the pattern generated by the time series and not necessarily the individual data values that offers to the manager who is an observer, a planner, or a controller of the system. Therefore, the Decomposition Analysis is used to identify several patterns that appear simultaneously in a time series.

A variety of factors are likely influencing data. It is very important in the study that these different influences or components be separated or decomposed out of the 'raw' data levels. In general, there are four types of components in time series analysis : Seasonality, Trend, Cycling and Irregularity.

$$X_t = S_t \cdot T_t \cdot C_t \cdot I$$

The first three components are deterministic which are called "Signals", while the last component is a random variable, which is called "Noise". To be able to make a proper forecast, we must know to what extent each component is present in the data. Hence, to understand and measure these components, the forecast procedure involves initially removing the component effects from the data (decomposition). After the effects are measured, making a forecast involves putting back the components on forecast estimates(recomposition). The time series decomposition process is depicted by the following flowchart :



The Three Signals Decomposition and its Reversal Process for Forecasting

Definitions of the major components in the above flowchart:

SEASONAL VARIATION:

When a repetitive pattern is observed over some time horizon, the series is said to have

seasonal behavior. Seasonal effects are usually associated with calendar or climatic changes. Seasonal variation is frequently tied to yearly cycles.

TREND:

A time series may be stationary or exhibit trend over time. Long-term trend is typically modeled as a linear, quadratic or exponential function.

CYCLICAL VARIATION :

An upturn or downturn not tied to seasonal variation. Usually results from changes in economic conditions.

1. Seasonalities are regular fluctuations which are repeated from year to year with about the same timing and level of intensity. The first step of a times series decomposition is to remove seasonal effects in the data. Without deseasonalizing the data, we may, for example, incorrectly infer that recent increase patterns will continue indefinitely; i.e., a growth trend is present, when actually the increase is 'just because it is that time of the year'; i.e., due to regular seasonal peaks. To measure seasonal effects, we calculate a series of seasonal indexes. A practical and widely used method to compute these indexes is the ratio-to-moving-average approach. From such indexes, we may quantitatively measure how far above or below a given period stands in comparison to the expected or 'business as usual' data period (the expected data are represented by a seasonal index of 100%, or 1.0).

2. Trend is growth or decay that is the tendencies for data to increase or decrease fairly steadily over time. Using the deseasonalized data, we now wish to consider the growth trend as noted in our initial inspection of the time series. Measurement of the trend component is done by fitting a line or any other function. This fitted function is calculated by the method of least squares and represents the overall trend of the data over time.

3. Cyclic oscillations are general up-and-down data changes; due to changes e.g., in the overall economic environment (not caused by seasonal effects) such as recession-and-expansion. To measure how the general cycle affects data levels, we calculate a series of cyclic indexes. Theoretically, the deseasonalized data still contains trend, cyclic, and irregular components. Also, we believe predicted data levels using the trend equation do represent pure trend effects. Thus, it stands to reason that the ratio of these respective data values should provide an index which reflects cyclic and irregular components only. As the business cycle is usually longer than the seasonal cycle, it should be understood that cyclic analysis is not expected to be as accurate as a seasonal analysis.

4. Irregularities (I) are any fluctuations not classified as one of the above. This component of the time series is unexplainable; therefore it is unpredictable. Estimation of I can be expected only when its variance is not too large. *Otherwise, it is not possible to decompose the series.* If the magnitude of variation is large, the projection for the future values will be inaccurate. The best one can do is to give a probabilistic interval for the future value given the probability of I is known.

5. Making a Forecast: At this point of the analysis, after we have completed the study of the

time series components, we now project the future values in making forecasts for the next few periods. The procedure is summarized below.

Step 1: Compute the future trend level using the trend equation.

Step 2: Multiply the trend level from Step 1 by the period seasonal index to include seasonal effects.

Step 3: Multiply the result of Step 2 by the projected cyclic index to include cyclic effects and get the final forecast result.

20.4 FORECASTING

Forecasting has become an essential activity in modern business. Different functional managers are interested in different types of forecasts prepared within the organisational forecasts. For example, a finance executive is interested in developing cashflow forecast to plan for raising necessary funds for different requirements.

Let us now consider the procedure for forecasting basing on statistical methods explained earlier for decomposing the time series. Recall the basic relationship assumed was that.

$$Y = T.C.S.I.$$

But in preparing the forecast, the last term I, the random component can not be projected and therefore, the relationship for forecasting simply.

$$Y = T.C.S.$$

Let us consider an illustration of a soft drink parlour of a departmental store. The management of soft drink wishes to forecast the demand for the next summer season (II quarter) so as to expand the facilities in the stores.

Example 20.1 : Forecasting in soft drink parlour of a Departmental store

The details of quarterly sales for the last five years are given under.

Table - 20.1 : Quarterly Sales details of soft drinks parlour during 2002 - 2006

Year	Quarters			
	I	II	III	IV
2002	5.0	8.0	4.5	3.5
2003	5.0	8.5	5.0	3.5
2004	5.5	8.0	5.5	4.0
2005	4.5	9.0	6.0	4.5
2006	5.0	10.0	5.0	4.5

The steps involved in decomposing consists of:

(a) Deseasonalising the time series by constructing seasonal indices by using ratio-to-moving averages method.

(b) Fitting a trend line of the type $\hat{Y} = a + bX$ for deseasionalised time series.

(c) Identifying the cyclical variations around Trend Line and finding out the index of cyclical fluctuation.

(d) Use of above information for forecasting the sales of the next year (2007).

Step 1. Seasonal component

For finding out the seasonal component, seasonal index is working out by following the given steps.

1. Calculating the moving average considering four-quarterly data at a time.
2. Contering the moving average.
3. Calculating the 'ratio-to-moving' average.
4. Finding out the normalised seasonal index and estimating the deseasoned sales in each quarter.

Table 20.2 : Calculation of Ratio-to-moving average

Year	Quarter	Actual Sales	4-quarter Moving average	Centred Moving average	Ratio-to moving average	Ratio-to moving average(%)
2002	I	5.0				
	II	8.0				
			5.250			
	III	4.5		5.250	0.857	85.71
			5.250			
	IV	3.5		5.313	0.659	65.88
			5.375			
2003	I	5.0		5.440	0.919	91.92
	II	8.5		5.500	1.545	154.55
			5.500			

	Ⅲ	5.0		5.563	0.899	89.83
			5.625			
	Ⅳ	3.5		5.563	0.629	62.92
			5.500			
2004	Ⅰ	5.5		5.563	0.988	98.87
			5.615			
	Ⅱ	8.0		5.688	1.406	140.65
			5.750			
	Ⅲ	5.5		5.625	0.978	97.78
			5.500			
	Ⅳ	4.0		5.625	0.711	71.11
			5.750			
2005	Ⅰ	4.5		5.813	0.774	77.41
			5.875			
	Ⅱ	9.0		5.938	1.515	151.57
			6.000			
	Ⅲ	6.0		6.063	0.989	98.96
			6.125			
	Ⅳ	4.5		6.250	0.720	72.00
			6.375			
2006	Ⅰ	5.0	6.125	6.250	0.800	80.00
	Ⅱ	10.0		6.063	1.649	164.93
			6.000			
	Ⅲ	5.0				
	Ⅳ	4.0				

Considering either ratio-to-moving average or percentage of the ratio seasonal indices can be constructed as follows :

Table 20.3 Computation of Normalised Seasonal Index

	Quarters			
Year	I	II	III	IV
2002	-	-	0.857	0.659
2003	0.919	1.545	0.899	0.629
2004	0.988	1.406	0.978	0.711
2005	0.744	1.515	0.989	0.720
2006	0.800	1.640	-	-
Mean Seasonal Index	0.8700	1.5290	0.9310	0.6800
Normalisation factor	4/4.01= 0.4975			
Normalised Seasonal Index	0.8678	1.525	0.928	0.678

These seasonal indices indicate that there exists a seasonal demand during the second quarter in every year. Once the seasonal fluctuations are identified the time series can be deseasonalised to find the impact of other components. As

$$Y = T.S.C.I.$$

Deseasonalised time series $\frac{Y}{S} = \frac{T.S.C.I.}{S}$

Therefore, a deseasonalised time series can be obtained by dividing the Actual values by its corresponding seasonal index.

Step 2. Fitting a Trend line to deseasonalise time series

Using the procedure adopted for least squares method of fitting trend, let us estimate trend for deseasoned sales values.

Table 20.4 Computation of Trend for Deseasonalised sales

Year	Quarter	Actual Sales	Seasonal Index	Deseasonalised Sales (Y)	Coded value of time Period(X)	XY	X^2
2002	I	5.0	0.8678	5.76	-19	-109.44	361
	II	8.0	1.525	5.25	-17	-89.25	289
	III	4.5	0.928	4.85	-15	-72.75	225
	IV	3.5	0.678	5.16	-13	-67.08	169
2003	I	5.0	0.8678	5.76	-11	-63.36	121
	II	8.5	1.525	5.57	-9	-50.13	81
	III	5.0	0.928	5.39	-7	-37.73	49
	IV	3.5	0.678	5.16	-5	-25.80	25
2004	I	5.5	0.8678	6.34	-3	19.02	9
	II	8.0	1.525	5.25	-1	-5.25	1
	III	5.5	0.928	5.93	+1	5.93	1
	IV	4.0	0.678	5.90	+3	17.70	9
2005	I	4.5	0.8678	5.19	+5	25.95	25
	II	9.0	1.525	5.90	+7	41.30	49
	III	6.0	0.928	6.47	+9	58.23	81
	IV	4.5	0.678	6.64	+11	73.04	121
2006	I	5.0	0.8678	5.76	+13	74.88	169
	II	10.0	1.525	6.56	+15	98.40	225
	III	5.0	0.928	5.39	+17	91.63	289
	IV	4.0	0.678	5.90	+19	112.10	361
				$\Sigma Y=114.13$	$\Sigma X=0$	$\Sigma XY=59.35$	$\Sigma X^2=2660$

$$a = \bar{\gamma} = \frac{114.13}{20} = 5.7065$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{59.35}{2660} = 0.0223$$

Trend line : 5.7065 + 0.0223 X

Step 3 : Identifying Cyclical Component

The cyclical component is identified by measuring variation of deseasonalised sales around the trend line. It can be expressed as ratio of actual deseasonalised sales to the value predicted by the trend line.

Table 20.5 Identification of Cyclical Component

Year	Quarter	Deseasonalised Sales Y	Trend	Cyclical Index Y/T(%)
2002	I	5.76	5.28	109.1
	II	5.25	5.33	98.5
	III	4.85	5.37	90.3
	IV	5.16	5.41	90.4
2003	I	5.76	5.46	105.5
	II	5.57	5.51	102.0
	III	5.39	5.55	97.1
	IV	5.16	5.59	92.3
2004	I	6.34	5.64	112.41
	II	5.25	5.68	92.41
	III	5.92	5.73	103.5
	IV	5.90	5.77	102.3
2005	I	5.19	5.82	89.2
	II	5.90	5.86	100.7
	III	6.47	5.91	109.5
	IV	6.64	5.95	111.6

2006	I	5.76	5.99	96.2
	II	6.56	6.04	108.6
	III	5.39	6.09	88.5
	IV	5.90	6.13	96.3

The random of irregular component is assumed to be relatively insignificant. The following figure represents the original time series, trend and seasonal components.

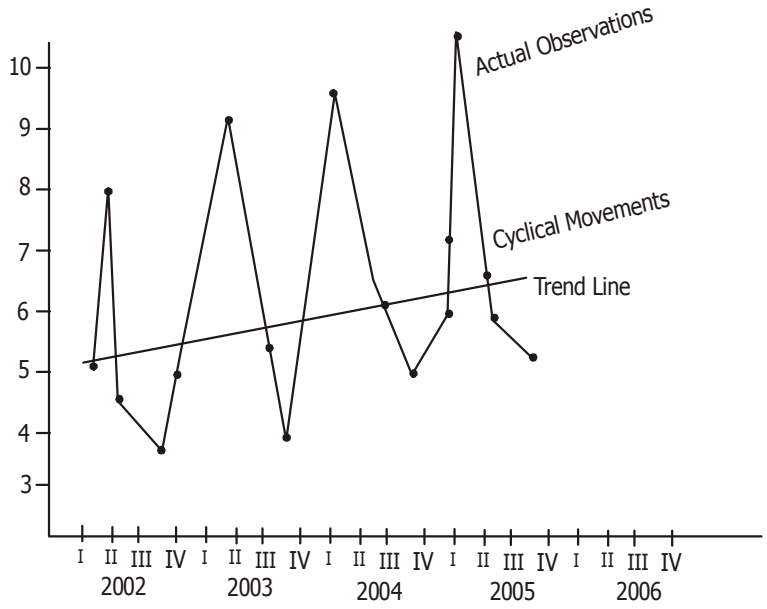


Fig. 20.2 Fitting a trend line observing seasonal variations

Thus, sales estimates for the soft drink parlour for the second and third quarters of 2007 would be Rs. 9.48 lakhs and Rs. 5.81 lakhs respectively.

This procedure has considered trend and seasonal fluctuations only into consideration for estimation. Whereas the cyclical fluctuations and irregular fluctuations are not much suited for forecasting.

20.5 BOX-JENKINS METHODOLOGY

INTRODUCTION

FORECASTING BASICS:

The basic idea behind self-projecting time series forecasting models is to find a mathematical formula that will approximately generate the historical patterns in a time series.

Time Series:

A time series is a set of numbers that measures the status of some activity over time. It is the historical record of some activity, with measurements taken at equally spaced intervals (exception: monthly) with a consistency in the activity and the method of measurement.

Approaches to time Series Forecasting:

There are two basic approaches to forecasting time series: the self-projecting time series and the cause-and-effect approach. Cause-and-effect methods attempt to forecast based on underlying series that are believed to cause the behavior of the original series. The self-projecting time series uses only the time series data of the activity to be forecast to generate forecasts. This latter approach is typically less expensive to apply and requires far less data and is useful for short, to medium-term forecasting.

Box-Jenkins Forecasting Method:

The univariate version of this methodology is a self- projecting time series forecasting method. The underlying goal is to find an appropriate formula so that the residuals are as small as possible and exhibit no pattern. The model- building process involves a few steps, repeated as necessary, to end up with a specific formula that replicates the patterns in the series as closely as possible and also produces accurate forecasts.

Jenkins Methodology

Box-Jenkins forecasting models are based on statistical concepts and principles and are able to model a wide spectrum of time series behavior. It has a large class of models to choose from and a systematic approach for identifying the correct model form. There are both statistical tests for verifying model validity and statistical measures of forecast uncertainty. In contrast, traditional forecasting models offer a limited number of models relative to the complex behavior of many time series, with little in the way of guidelines and statistical tests for verifying the validity of the selected model.

Data : The misuse, misunderstanding, and inaccuracy of forecasts are often the result of not appreciating the nature of the data in hand. The consistency of the data must be insured, and it must be clear what the data represents and how it was gathered or calculated. As a rule of thumb, Box-Jenkins requires at least 40 or 50 equally-spaced periods of data. The data must also be edited to deal with extreme or missing values or other distortions through the use of functions such as log or inverse to achieve stabilization.

Preliminary Model Identification Procedure: A preliminary Box-Jenkins analysis with a plot of the initial data should be run as the starting point in determining an appropriate model. The input data must be adjusted to form a stationary series, one whose values vary more or less uniformly about a fixed level over time. Apparent trends can be adjusted by having the model apply a tech-

nique of "regular differencing," a process of computing the difference between every two successive values, computing a differenced series which has overall trend behavior removed. If a single differencing does not achieve stationarity, it may be repeated, although rarely, if ever, are more than two regular differencing required. Where irregularities in the differenced series continue to be displayed, log or inverse functions can be specified to stabilize the series, such that the remaining residual plot displays values approaching zero and without any pattern. This is the error term, equivalent to pure, white noise.

Pure Random Series: On the other hand, if the initial data series displays neither trend nor seasonality, and the residual plot shows essentially zero values within a 95% confidence level and these residual values display no pattern, then there is no real-world statistical problem to solve and we go on to other things.

Model Identification Background

Basic Model : With a stationary series in place, a basic model can now be identified. Three basic models exist, AR (autoregressive), MA (moving average) and a combined ARMA in addition to the previously specified RD (regular differencing): These comprise the available tools. When regular differencing is applied, together with AR and MA, they are referred to as ARIMA, with the I indicating "integrated" and referencing the differencing procedure.

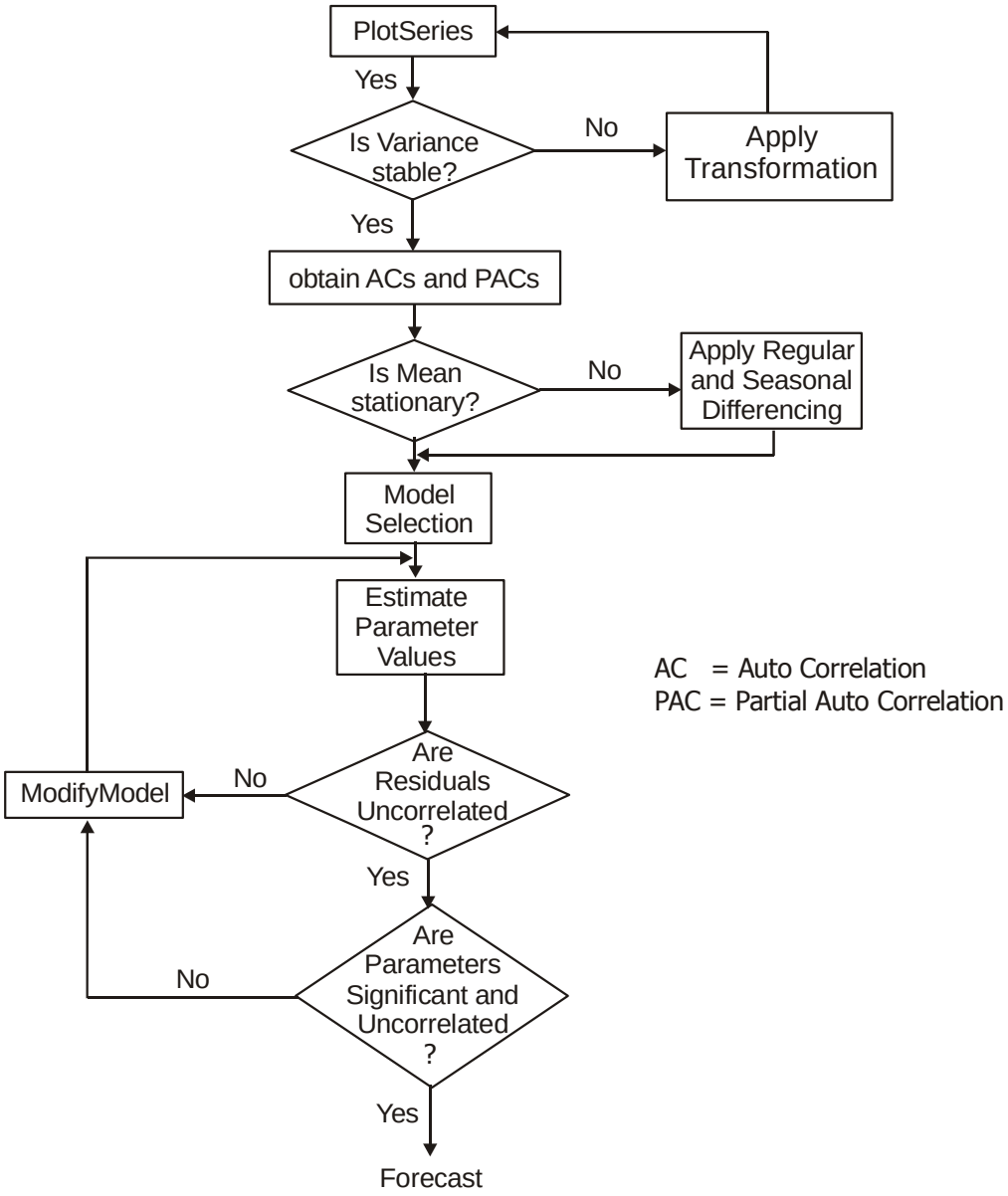
Seasonality : In addition to trend, which has now been provided for, stationary series quite commonly display seasonal behavior where a certain basic pattern tends to be repeated at regular seasonal intervals. The seasonal pattern may additionally frequently display constant change over time as well. Just as regular differencing was applied to the overall trending series, seasonal differencing (SD) is applied to seasonal non-stationarity as well. And as autoregressive and moving average tools are available with the overall series, so too, are they available for seasonal phenomena using seasonal autoregressive parameters (SAR) and seasonal moving average parameters (SMA).

Establishing Seasonality : The need for seasonal autoregression (SAR) and seasonal moving average (SMA) parameters is established by examining the autocorrelation and partial autocorrelation patterns of a stationary series at lags that are multiples of the number of periods per season. These parameters are required if the values at lags s, 2s, etc. are nonzero and display patterns associated with the theoretical patterns for such models. Seasonal differencing is indicated if the autocorrelations at the seasonal lags do not decrease rapidly.

B-J Modeling Approach to Forecasting

Referring to the below chart know that, the variance of the errors of the underlying model must be invariant, i.e., constant. This means that the variance for each subgroup of data is the same and does not depend on the level or the point in time. If this is violated then one can remedy this by stabilizing the variance. Make sure that there are no deterministic patterns in the data. Also, one must not have any pulses or one-time unusual values. Additionally, there should be no level or step shifts. Also, no seasonal pulses should be present.

The reason for all of this is that if they do exist, then the sample autocorrelation and partial autocorrelation will seem to imply ARIMA structure. Also, the presence of these kinds of model components can obfuscate or hide structure. For example, a single outlier or pulse can create an effect where the structure is masked by the outlier.



Box - Jenkins Modeling Approach

20.6 AUTOREGRESSIVE MODELS

The autoregressive model is one of a group of linear prediction formulas that attempt to predict an output of a system based on the previous outputs and inputs, such as:

$$Y(t) = \beta_1 + \beta_2 Y(t-1) + \beta_3 X(t-1) + \varepsilon_t ,$$

where X(t-1) and Y(t-1) are the actual value (inputs) and the forecast (outputs), respectively. These types of regressions are often referred to as *Distributed Lag Autoregressive Models, Geometric Distributed Lags,* and *Adaptive Models in Expectation* , among others.

A model which depends only on the previous outputs of the system is called an autoregressive model (AR), while a model which depends only on the inputs to the system is called a moving average model (MA), and of course a model based on both inputs and outputs is an autoregressive-moving-average model (ARMA). Note that by definition, the AR model has only poles while the MA model has only zeros. Deriving the autoregressive model (AR) involves estimating the coefficients of the model using the method of least squared error.

Autoregressive processes as their name implies, regress on themselves. If an observation made at time (t), then, p-order, [AR(p)], autoregressive model satisfies the equation:

$$X(t) = \phi_0 + \phi_1 X(t-1) + \phi_2 X(t-2) + \phi_2 X(t-3) + \cdots + \phi_p X(t-p) + \varepsilon_t,$$

where ε_t is a White-Noise series.

The current value of the series is a linear combination of the p most recent past values of itself plus an error term, which incorporates everything new in the series at time t that is not explained by the past values. This is like a multiple regressions model but is regressed not on independent variables, but on past values; hence the term “Autoregressive” is used.

20.6.1 AUTOCORRELATION:

An important guide to the properties of a time series is provided by a series of quantities called sample autocorrelation coefficients or serial correlation coefficient, which measures the correlation between observations at different distances apart. These coefficients often provide insight into the probability model which generated the data. The sample autocorrelation coefficient is similar to the ordinary correlation coefficient between two variables (x) and (y), except that it is applied to a single time series to see if successive observations are correlated.

Given (N) observations on discrete time series we can form (N - 1) pairs of observations. Regarding the first observation in each pair as one variable, and the second observation as a second variable, the correlation coefficient is called autocorrelation coefficient of order one.

20.6.2 CORRELOGRAM:

A useful aid in interpreting a set of autocorrelation coefficients is a graph called a correlogram, and it is plotted against the lag(k); where is the autocorrelation coefficient at lag(k). A correlogram can be used to get a general understanding on the following aspects of our time series:

A random series: if a time series is completely random then for Large (N), will be approximately zero for all non-zero values of (k).

Short-term correlation: stationary series often exhibit short-term correlation characterized by a fairly large value of 2 or 3 more correlation coefficients which, while significantly greater than zero, tend to get successively smaller.

Non-stationary series: If a time series contains a trend, then the values of will not come to zero except for very large values of the lag.

Seasonal fluctuations: Common autoregressive models with seasonal fluctuations, of periods are:

$X(t) - a + b\ X(t-s) + \varepsilon_t$ and

$X(t) = a + b\ X(t-s) + c\ X(t-2s) + \varepsilon_t$

where ε_t is a White-Noise series.

20.7 SUMMARY

Importance of time series are discussed different graphs ae also given to analysing trend curves. Detailed of decomposition analysis are discussed with an example.

Box - Jenkins Methodology for time series analysis has been given. Finally, Auto regressive, Auto correlation, and correlogram models have been out lined.

20.8 EXERCISES

1. What do you understand by time series analysis?
2. Explain the Analysing Trend?
3. Explain the Decomposition Analysis?
4. Explain the Box=Jenkins Methodology.
5. Explain the Autoregressive, Auto correlation and Correlogram Models.
6. Forecasting demand for Standard Chartered Bank Advances. The details of quarterly demand for loans for the last five years is as follows.

	Quarterly Demand for Loans (Rs. in Crores)			
Year	I	II	III	IV
2004	3.5	3.9	3.4	3.5
2005	3.5	10.0	3.5	4.0
2006	4.0	12.0	2.5	4.5
2007	4.5	6.0	2.9	4.0
2008	5.0	12.0	5.0	4.0

The steps involved in decomposing consists of

(a) Deseasonalising the time series by constructing seasonal indices by using ratio-to-moving avarages method.

(b) Fitting a trend line of the type $\hat{Y} = a + bX$ for deseasionalized time series.

(c) Identifying the cyclical variations around Trend line and finding out the index of cyclical fluctuation.

(d) Use of above information for forecasting the sales of the next year (2009).

7. An Engineering firm producing farm equipment wants to predict future sales based on the analysis of its past sales pattern. The sales of the company for the last five years are given in Table. 1

Table 1 : Quarterly Sales of an Engineering Firm during 1983 - 87 (Rs. in Lakhs)

Year	I	II	III	IV
2000	5.5	5.4	7.2	6.0
2001	4.8	5.6	6.3	5.6
2002	4.0	6.3	7.0	6.5
2003	5.2	6.5	7.5	7.2
2004	6.0	7.0	8.4	7.7

The steps involved in decomposing consists of :

(a) Deseasonalising the time series by constructing seasonal indices by using ratio-to-moving averages method.

(b) Fitting a trend line of the type $\hat{Y} = a + bX$ for deseasonalised time series

(c) Identifying the cyclical variations around trend line and finding out the index of cyclical fluctuation.

(d) Use of above information for forecasting the sales of the next year 2005.

8. The details of quarterly production of coffee in an Indian states for the last five years.

Year	Production (in Tonnes)			
	Quarter - I	Quarter - II	Quarter - III	Quarter - IV
2002	5	1	10	17
2003	7	1	10	16
2004	9	3	8	18
2005	5	2	15	19
2006	8	4	14	21

The steps involved in decomposing consists of :

a) Deseasonalising the time series by constructing seasonal indices by using ratio-to-moving averages method.

b) Fitting a trend line of the $\hat{Y} = a + bX$ for deseasionalised time series.

c) Identifying the cyclical variations around trend line and fitting out the index of cyclical fluctuation.

d) Use of above information for forecasting the production of the next year (2007)

20.9 REFERENCE BOOKS

1. Budnicks, F.S. 1983 Applied Mathematics For Business, Economics and Social Sciences, Mc Graw - Hill ; New York.

2. Hughes, A.J. 1983, Applied Mathematics For Business Economics and Social Sciences, Irwin; Homewood.

3. Weber, J.E. 1983 Mathematical Analysis, Business and Economics Applications, Harper and Row, New York.

LESSON WRITER

PROF. K. CHANDAN