UNIT – I

INTRODUCTION

Learning Objectives

After reading this lesson, you should be able to understand:

- **Meaning, objectives and types of research**
- **Qualities of researcher**
- **Significance of research**
- **Research process**
- **Research problem**
- **Features, importance, characteristics, concepts and types of Research design**
- **Case study research**
- **Hypothesis and its testing**
- **Sample survey and sampling methods**

1.1 Meaning of Research

Research in simple terms, refers to a search for knowledge. It is also known as a scientific and systematic search for information on particular topic or issue. It is also known as the art of scientific investigation. Several social scientists have defined research in different ways

In the Encyclopedia of Social Sciences, D. Slesinger and M. Stephension (1930) defined research as "the manipulation of things, concept or symbols for the purpose of generalizing to extend, correct or verify knowledge, whether that knowledge aids in construction of theory or in practice of an art".

According to Redman and Mory (1923), defined research is a "systematized effort to gain new knowledge". It is an academic activity and therefore the term should be used in a technical sense. According to Clifford Woody (Kothari 1988) research comprises "defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organizing and

evaluating data; making deductions and reaching conclusions; and finally, carefully testing the conclusions to determine whether they fit the formulating hypothesis".

Thus, research is an original addition to the available knowledge, which contributes to its further advancement. It is an attempt to pursue truth through the methods of study, observation, comparison and experiment. In sum, research is the search for knowledge, using objective and systematic methods to find solution to a problem.

1.1.1 Objectives of research

The objective of research is to discover answers to questions by applying scientific procedures. In the other words, the main aim of research is to find out truth which is hidden and has not yet been discovered. Although every research study has its own specific objectives, research objectives may be broadly grouped as follows:-

1. to gain familiarity with or new insights into a phenomenon (i.e., formulative research studies);
2. to accurately portray the characteristics of a particular individual, group, or a situation (i.e., descriptive research studies);
3. to analyse the frequency with which something occurs (i.e., diagnostic research studies); and
4. to examine a hypothesis of a causal relationship between two variables (i.e., hypothesis-testing research studies).

1.1.2 Research methods versus methodology

Research methods include all those techniques/methods that are adopted for conducting research. Thus, research techniques or methods are the methods the researchers adopt for conducting the research operations.

On the other hand, research methodology is the way of systematically solving the research problem. It is a science of studying how research is conducted scientifically. Under it, the researcher acquaints himself/herself with

the various steps generally adopted to study a research problem, along with the underlying logic behind them. Hence, it is not only important for the researcher to know the research techniques/methods, but also the scientific approach called methodology.

1.1.3 Research approaches

There are two main approaches to research, namely quantitative approach and qualitative approach. The quantitative approach involves the collection of quantitative data, which are put to rigorous quantitative analysis in a formal and rigid manner. This approach further includes experimental, inferential, and simulation approaches to research. Meanwhile, the qualitative approach uses the method of subjective assessment of opinions, behaviour and attitudes. Research in such a situation is a function of the researcher's impressions and insights. The results generated by this type of research is either in non-quantitative form or in the form which can not be put to rigorous quantitative analysis. Usually, this approach uses techniques like depth interviews, focus group interviews, and projective techniques.

1.1.4 Types of research

There are different types of research. The basic ones are as follows:

1) **Descriptive vs. Analytical:**

Descriptive research comprises surveys and fact-finding enquiries of different types. The main objective of descriptive research is describing the state of affairs as it prevails at the time of study. The term ex post facto research is quite often used for descriptive research studies in social sciences and business research. The most distinguishing feature of this method is that the researcher has no control over the variables here. He/she has to only report what is happening or what has happened. Majority of the ex post facto research projects are used for descriptive studies in which the researcher attempts to examine

phenomena, such as the consumers' preferences, frequency of purchases, shopping, etc. Despite the inability of the researchers to control the variables, ex post facto studies may also comprise attempts by them to discover the causes of the selected problem. The methods of research adopted in conducting descriptive research are survey methods of all kinds, including correlational and comparative methods.

Meanwhile in the analytical research, the researcher has to use the already available facts or information, and analyse them to make a critical evaluation of the subject.

2) **Applied vs. Fundamental**

Research can also be applied or fundamental research. An attempt to find a solution to an immediate problem encountered by a firm, an industry, a business organisation, or the society is known as applied research. Researchers engaged in such researches aim at drawing certain conclusions confronting a concrete social or business problem. On the other hand, fundamental research mainly concerns generalizations and formulation of a theory. In other words, "Gathering knowledge for knowledge's sake is termed 'pure' or 'basic' research" (Young in Kothari 1988). Researches relating to pure mathematics or concerning some natural phenomenon are instances of fundamental research. Likewise, studies focusing on human behaviour also fall under the category of fundamental research. Thus, while the principal objective of applied research is to find a solution to some pressing practical problem, the objective of basic research is to find information with a broad base of application and add to the already existing organized body of scientific knowledge.

3) **Quantitative vs. Qualitative**

Quantitative research relates to aspects that can be quantified or can be expressed in terms of quantity. It involves the measurement of quantity or

amount. The various available statistical and econometric methods are adopted for analysis in such research. They include correlation, regressions, time series analysis, etc.

Whereas, qualitative research is concerned with qualitative phenomenon, or more specifically, the aspects relating to or involving quality or kind. For example, an important type of qualitative research is 'Motivation Research', which investigates into the reasons for human behaviour. The main aim of this type of research is discovering the underlying motives and desires of human beings, using in-depth interviews. The other techniques employed in such research are story completion tests, sentence completion tests, word association tests, and other similar projective methods. Qualitative research is particularly significant in the context of behavioural sciences, which aim at discovering the underlying motives of human behaviour. Such research help to analyse the various factors that motivate human beings to behave in a certain manner, besides contributing to an understanding of what makes individuals like or dislike a particular thing. However, it is worth noting that conducting qualitative research in practice is considerably a difficult task. Hence, while undertaking such research, seeking guidance from experienced expert researchers is important.

4) **Conceptual vs. Empirical**

A research related to some abstract idea or theory is known as conceptual research. Generally, philosophers and thinkers use it for developing new concepts or for reinterpreting the existing ones. Empirical research, on the other hand, exclusively relies on observation or experience with hardly any regard for theory and system. Such research is data based. They often come up with conclusions that can be verified through experiment or observation. They are also known as experimental type of research. Under such research, it is

important to first collect facts, their source and actively do certain things to stimulate the production of desired information. In such a research, the researcher must first identify a working hypothesis or make a guess of the probable results. Next, he/she gathers sufficient facts to prove or disprove the stated hypothesis. Then he/she formulates experimental designs, which according to him/her would manipulate the individuals or the materials concerned, so as to obtain the desired information. This type of research is thus characterized by the researcher's control over the variables used to study their effects. Empirical research is most appropriate when an attempt is made to prove that certain variables influence the other variables in some way. Therefore, the results obtained using the experimental or empirical studies are considered as one of the most powerful evidences for a given hypothesis.

5) **Other types of research:** The remaining types of research are variations of one or more of the afore-mentioned methods. They vary in terms of the purpose of research, or the time required to complete it, or based on some other similar factor. On the basis of time, research may either be in the nature of one-time or longitudinal research. While the research is restricted to a single time-period in the former case, it is conducted over several time-periods in the latter case. Depending upon the environment in which the research is to be conducted, it may also be laboratory research or field-setting research, or simulation research, besides being diagnostic or clinical in nature. Under such research, in-depth approaches or case-study methods may be employed to analyse the basic causal relations. These studies usually conduct a detailed in-depth analysis of the causes of things or events of interest, and use very small samples and a sharp data collecting method. The research may also be explanatory in nature. Formalized research studies consist of substantial structure and specific hypotheses to be verified. As regards historical research,

sources like historical documents, remains, etc., are utilized to study past events or ideas. It also includes philosophy of persons and groups of the past or any remote point of time. Research is also categorized as decision-oriented and conclusion-oriented. In the case of decision-oriented research, it is always carried out for the need of a decision maker and hence, the researcher has no freedom to conduct the research as per his/her own desires. Whereas, under conclusion-oriented research, the researcher is free to choose the problem, redesign the enquiry as it progresses and even change conceptualization as he/she wishes to. Further, operations research is a kind of decision- oriented research, because it is a scientific method which provides the executive departments a quantitative basis for decision-making with respect to the activities under their purview.

1.1.5 Importance of knowing how to conduct research

The following are the importance of knowing how to conduct a research:

(i) the knowledge of research methodology provides training to new researchers and enables them to do research properly. It helps them to develop disciplined thinking or a 'bent of mind' to objectively observe the field.

(ii) the knowledge of doing research would inculcate the ability to evaluate and utilise the research findings with confidence;

(iii) the knowledge of research methodology equips the researcher with tools that help him/her to observe things objectively; and

(iv) the knowledge of methodology helps the research consumer to evaluate research and make rational decisions.

1.1.6 Qualities of a researcher

It is important for a researcher to have certain qualities to conduct research. Foremost, the researcher being a scientist should be firmly committed to the

'articles of faith' of the scientific methods of research. This implies that a researcher should be a social science person in the truest sense.

Sir Michael Foster (Wilkinson and Bhandarkar 1979) identified a few distinctive qualities of a scientist. According to him, a true research scientist should possess the following main three qualities.

(1) First of all, the nature of a researcher must be of the temperament that vibrates in unison with the theme which he is searching. Hence, the seeker of knowledge must be truthful with truthfulness of nature, which is much more important, much more exacting than what is sometimes known as truthfulness. The truthfulness relates to the desire for accuracy of observation and precision of statement. Ensuring facts is the principle rule of science, which is not an easy matter. Such difficulty may arise due to untrained eye, which fails to see anything beyond what it has the power of seeing and sometimes even less than that. This may also be due to the lack of discipline in the method of science. An unscientific individual often remains satisfied with expressions like approximately, almost, nearly, etc., which is never what nature, is. It cannot see two things which differ, however minutely, as the same.

(2) A researcher must possess an alert mind. The Nature is constantly changing and revealing itself through various ways. A scientific researcher must be keen and watchful to notice such changes, no matter how small or insignificant they may appear. Such receptivity has to be cultivated slowly and patiently over time by the researcher through practice. No individual who is not alert and receptive, or is ignorant or has no keen eyes or mind to observe the unusual behind the routine, can make a good researcher. Research demands a systematic immersion into the subject matter for the researcher to be able to grasp even the slightest hint that may culminate into significant research problems. In this context, Cohen and Negal (Wilkinson and Bhandarkar 1979)

state that "The ability to perceive in some brute experience the occasion of a problem is not a common talent among men… It is a mark of scientific genius to be sensitive to difficulties where less gifted people pass by untroubled by doubt" (Selltiz, et. al.,1965).

(3) Scientific enquiry is pre-eminently an intellectual effort. It requires the moral quality of courage, which reflects the courage of a steadfast endurance. The science of conducting research is not an easy task. There are occasions when a research scientist might feel defeated or completely lost. This is a stage when the researcher would need immense courage and a sense of conviction. The researcher must learn the art of enduring intellectual hardships. In the words of Darwin, "It's dogged that does it" (Wilkinson and Bhandarkar 1979).

In order to cultivate the afore-mentioned three qualities of a researcher, a fourth one may be added. This is the quality of making statements cautiously. According to Huxley, "The assertion that outstrips the evidence is not only a blunder but a crime" (Thompson 1975). A researcher should cultivate the habit of reserving judgment when the required data are insufficient.

1.1.7 Significance of research

According to a famous Hudson Maxim, "All progress is born of inquiry. Doubt is often better than overconfidence, for it leads to inquiry, and inquiry leads to invention" (Wilkinson and Bhandarkar 1979). It brings out the significance of research, increased amounts of which makes progress possible. Research encourages scientific and inductive thinking, besides promoting the development of logical habits of thinking and organisation.

The role of research in applied economics in the context of an economy or business is greatly increasing in modern times. The increasingly complex nature of government and business has raised the use of research in solving

operational problems. Research assumes significant role in the formulation of economic policy, for both the government and business. It provides the basis for almost all government policies of an economic system. Government budget formulation, for example, depends particularly on the analysis of needs and desires of people, and the availability of revenues, which requires research. Research helps to formulate alternative policies, in addition to examining the consequences of these alternatives. Thus, research also facilitates the decision-making of the policy-makers, although in itself it is not a part of research. In the process, research also helps in the proper allocation of a country's scarce resources. Research is also necessary for collecting information on the social and economic structure of an economy to understand the process of change occurring in the country. Collection of statistical information, though not a routine task, involves various research problems. Therefore, large staff of research technicians or experts is engaged by the government these days to undertake this work. Thus, research as a tool of government economic policy formulation involves three distinct stages of operation, viz., (i) investigation of economic structure through continual compilation of facts; (ii) diagnosis of events that are taking place and the analysis of the forces underlying them; and (iii) the prognosis, i.e., the prediction of future developments (Wilkinson and Bhandarkar 1979).

Research also assumes a significant role in solving various operational and planning problems associated with business and industry. In several ways, operations research, market research, and motivational research are vital and their results assist in taking business decisions. Market research is refers to the investigation of the structure and development of a market for the formulation of efficient policies relating to purchases, production and sales. Operational research relates to the application of logical, mathematical, and analytical

techniques to find solution to business problems such as cost minimization or profit maximization, or the optimization problems. Motivational research helps to determine why people behave in the manner they do with respect to market characteristics. More specifically, it is concerned with the analyzing the motivations underlying consumer behaviour. All these researches are very useful for business and industry, who are responsible for business decision-making.

Research is equally important to social scientists for analyzing social relationships and seeking explanations to various social problems. It gives intellectual satisfaction of knowing things for the sake of knowledge. It also possess practical utility for the social scientist to gain knowledge so as to be able to do something better or in a more efficient manner. This, research in social sciences is concerned with both knowledge for its own sake, and knowledge for what it can contribute to solve practical problems.

1.2 Research process

Research process comprises a series of steps or actions required for effectively conducting research and for the sequencing of these steps. The following are the various steps that provide useful procedural guideline regarding the conduct research.

(1) formulating the research problem;
(2) extensive literature survey;
(3) developing hypothesis;
(4) preparing the research design;
(5) determining sample design;
(6) collecting data;
(7) execution of the project;
(8) analysis of data;
(9) hypothesis testing;
(10) generalization and interpretation, and

(11) preparation of the report or presentation of the results. In other words, it involves the formal write-up of conclusions.

1.3 Research Problem

The first and foremost stage in the research process is to select and properly define the research problem. A researcher should firstly identify a problem and formulate it, so as to make it amenable or susceptible to research.

In general, a research problem refers to some kind of difficulty the researcher might encounter or experience in the context of either a theoretical or practical situation, which he/she would like to resolve and find a solution to. A research problem is generally said to exist if the following conditions emerge (Kothari 1988):

(i) there should be an individual or an organisation, say X, to whom the problem can be attributed. The individual or the organization is situated in an environment Y, which is governed by certain uncontrolled variables Z_i.

(ii) there should be atleast two courses of action to be pursued, say A_1 and A_2. These courses of action are defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

(iii) there should be atleast two alternative possible outcomes of the said course of actions, say B_1 and B_2. Of them, one alternative should be preferable to the other. That is, atleast one outcome should be what the researcher wants, which becomes an objective.

(iv) the courses of possible action available must offer a chance to the researcher to achieve the objective, but not the equal chance. Therefore, if $P(B_j / X, A, Y)$ represents the probability of the occurrence of an outcome B_j when X selects A_j in Y, then $P(B_1 / X, A_1, Y) \neq P(B_1 / X, A_2, Y)$. Putting it in simple words, it means that the choices must not have equal efficiencies for the desired outcome.

Above all these conditions, the individual or organisation may be said to have arrived at the research problem only if X does not know what course of action to be taken is the best. In other words, X should have a doubt about the solution. Thus, an individual or a group of persons can be said to have a problem if they have more than one desired outcome. They should have two or more alternative courses of action, which have some but not equal efficiency for probing the desired objectives, such that they have doubts about the best course of action to be taken.

Thus, the various components of a research problem may be summarised as:

(i) there should be an individual or a group who have some difficulty or problem.
(ii) there should be some objective(s) to be pursued. A person or an organization who want nothing cannot have a problem.
(iii) there should be alternative ways of pursuing the objective the researcher wants to pursue. This implies that there should be more than one alternative means available to the researcher. This is because if the researcher has no choice of alternative means, he/she would not have a problem.
(iv) there should be some doubt in the mind of the researcher about the choice of alternative means. This implies that research should answer the question relating to the relative efficiency or suitability of the possible alternatives.
(v) there should be a context to which the difficulty relates.

Thus, identification of a research problem is the pre-condition to conducting research. A research problem is said to be the one which requires a researcher to find the best available solution to the given problem. That is, the researcher needs to find out the best course of action through which the research objective may be achieved optimally in the context of a given situation. Several factors may contribute to making the problem complicated. For example, the environment may alter, thus affecting the efficiencies of the alternative course of actions taken or the quality of the outcomes. Or, the number of alternative

course of actions may be very large and the individual not involved in making the decision may be affected by the change in environment, and may react to it favorably or unfavorably. Other similar factors are also likely to cause such changes in the context of research, all of which may be considered from the point of view of a research problem.

1.4 Research Design

The most important problem after defining the research problem is preparing the design of the research project, which is popularly known as the 'research design'. A research design helps to decide upon issues like what, when, where, how much, by what means, etc., with regard to an enquiry or a research study. "A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure. In fact, the research design is the conceptual structures within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data" (Selltiz, et.al. 1962)**.** Thus, research design provides an outline of what the researcher is going to do in terms of framing the hypothesis, its operational implications, and the final data analysis. Specifically, the research design highlights decisions which include:

(i) the nature of the study
(ii) the purpose of the study
(iii) the location where the study would be conducted
(iv) the nature of data required
(v) from where the required data can be collected
(vi) what time period the study would cover
(vii) the type of sample design that would be used
(viii) the techniques of data collection that would be used
(ix) the methods of data analysis that would be adopted
(x) the manner in which the report would be prepared

In view of the stated research design decisions, the overall research design may be divided into the following (Kothari 1988)

(a) the sampling design that deals with the method of selecting items to be observed for the selected study;

(b) the observational design that relates to the conditions under which the observations are to be made;

(c) the statistical design that concerns with the question of how many items are to be observed, and how the information and data gathered are to be analysed; and

(d) the operational design that deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

1.4.1 Features of research design

The important features of research design may be outlined as follows:

(i) it constitutes a plan that identifies the types and sources of information required for the research problem;

(ii) it constitutes a strategy that specifies the methods of data collection and analysis which would be adopted; and

(iii) it also specifies the time period of research and monetary budget involved in conducting the study, which comprise the two major constraints of undertaking any research.

1.4.2 Concepts relating to research design

It is also important to be familiar with the important concepts relating to research design. Some of them are discussed here.

1. Dependent and independent variables: A magnitude that varies is known as a variable. The concept may assume different quantitative values, like height, weight, income, etc. Qualitative variables are not quantifiable in the strictest

sense or objectively. However, the qualitative phenomena may also be quantified in terms of the presence or absence of the attribute(s) considered. Phenomena that assumes different values quantitatively even in decimal points are known as 'continuous variables'. But, all variables need not be continuous. Values that can be expressed only in integer values are called 'non-continuous variables'. In statistical term, they are also known as 'discrete variables'. For example, age is a continuous variable, whereas the number of children is a non-continuous variable. When changes in one variable depends upon the changes in one or more other variables, it is known as a dependent or endogenous variable, and the variables that cause the changes in the dependent variable are known as the independent or explanatory or exogenous variables. For example, if demand depends upon price, then demand is a dependent variable, while price is the independent variable. And, if more variables determine demand, like income and prices of substitute commodity, then demand also depends upon them in addition to the own price. Then, demand is a dependent variable which is determined by the independent variables own price, income and price of substitutes.

2 .Extraneous variable: The independent variables which are not directly related to the purpose of the study but affect the dependent variable are known as extraneous variables. For instance, assume that a researcher wants to test the hypothesis that there is a relationship between children's school performance and their self-concepts, in which case the latter is an independent variable and the former the dependent variable. In this context, intelligence may also influence the school performance. However, since it is not directly related to the purpose of the study undertaken by the researcher, it would be known as an extraneous variable. The influence caused by the extraneous variable(s) on the dependent variable is technically called as an 'experimental error'. Therefore, a

research study should always be framed in such a manner that the dependent variable(s) that completely influence the change in the independent variable and any other extraneous variable or variables.

3. Control: One of the most important features of a good research design is to minimize the effect of extraneous variable(s). Technically, the term 'control' is used when a researcher designs the study in such a manner that it minimizes the effects of extraneous independent variables. The term 'control' is used in experimental research to reflect the restrain in experimental conditions.

4. Confounded relationship: The relationship between the dependent and independent variables is said to be confounded by an extraneous variable(s), when the dependent variable is not free from its effects.

5. Research hypothesis: When a prediction or a hypothesized relationship is tested by adopting scientific methods, it is known as research hypothesis. The research hypothesis is a predictive statement which relates to a dependent variable and an independent variable. Generally, a research hypothesis must consist of at least one dependent variable and one independent variable. Whereas, the relationships that are assumed but not to be tested are predictive statements that are not to be objectively verified are not classified as research hypotheses.

6. Experimental and non-experimental hypothesis testing research: When the objective of a research is to test a research hypothesis, it is known as a hypothesis-testing research. Such research may be in the nature of experimental design or non-experimental design. A research in which the independent variable is manipulated is known as 'experimental hypothesis-testing research', whereas a research in which the independent variable is not manipulated is termed as 'non-experimental hypothesis-testing research'. For example, assume that a researcher wants to examine whether family income influences the school

attendance of a group of students, by calculating the coefficient of correlation between the two variables. Such an example is known as a non-experimental hypothesis-testing research, because the independent variable family income is not manipulated here. Again assume that the researcher randomly selects 150 students from a group of students who pay their school fees regularly and then classifies them into two sub-groups by randomly including 75 in Group A, whose parents have regular earning, and 75 in group B, whose parents do not have regular earning. Assume that at the end of the study, the researcher conducts a test on each group in order to examine the effects of regular earnings of the parents on the school attendance of the student. Such a study is an example of experimental hypothesis-testing research, because in this particular study the independent variable regular earnings of the parents has been manipulated.

7. Experimental and control groups: When a group is exposed to usual conditions in an experimental hypothesis-testing research, it is known as 'control group'. On the other hand, when the group is exposed to certain new or special condition, it is known as an 'experimental group'. In the afore-mentioned example, the Group A can be called a control group and the Group B an experimental group. If both the groups A and B are exposed to some special feature, then both the groups may be called as 'experimental groups'. A research design may include only the experimental group or both the experimental and control groups together.

8. Treatments: Treatments are referred to the different conditions to which the experimental and control groups are subject to. In the example considered, the two treatments are the parents with regular earnings and those with no regular earnings. Likewise, if a research study attempts to examine through an experiment the comparative impacts of three different types of fertilizers on the

yield of rice crop, then the three types of fertilizers would be treated as the three treatments.

9. Experiment: An experiment refers to the process of verifying the truth of a statistical hypothesis relating to a given research problem. For instance, experiment may be conducted to examine the yield of a certain new variety of rice crop developed. Further, Experiments may be categorized into two types, namely, absolute experiment and comparative experiment. If a researcher wishes to determine the impact of a chemical fertilizer on the yield of a particular variety of rice crop, then it is known as absolute experiment. Meanwhile, if the researcher wishes to determine the impact of chemical fertilizer as compared to the impact of bio-fertilizer, then the experiment is known as a comparative experiment.

10. Experiment unit(s): Experimental units refer to the pre-determined plots, characteristics or the blocks, to which the different treatments are applied. It is worth mentioning here that such experimental units must be selected with great caution.

1.4.3 Types of research design

There are different types of research designs. They may be broadly categorized as:

(1) exploratory research design;
(2) descriptive and diagnostic research design; and
(3) hypothesis-testing research design.

1. Exploratory research design:

The exploratory research design is known as formulative research design. The main objective of using such a research design is for formulating a research problem for an in-depth or more precise investigation, or for developing a working hypothesis from an operational aspect. The major purpose of such

studies is the discovery of ideas and insights. Therefore, such a research design suitable for such a study should be flexible enough to provide opportunity for considering different dimensions of the problem under study. The in-built flexibility in research design is required as the initial research problem would be transformed into a more precise one in the exploratory study, which in turn may necessitate changes in the research procedure for collecting relevant data. Usually, the following three methods are considered in the context of a research design for such studies. They are (a) a survey of related literature; (b) experience survey; and (c) analysis of 'insight-stimulating' instances.

2. Descriptive and diagnostic research design:

A descriptive research design is concerned with describing the characteristics of a particular individual, or a group. Meanwhile, a diagnostic research design determines the frequency with which a variable occurs or its relationship with another variable. In other words, the study analyzing whether a certain variable is associated with another comprises a diagnostic research study. On the other hand, a study that is concerned with specific predictions or with the narration of facts and characteristics relating to an individual, group or situation, are instances of descriptive research studies. Generally, most of the social research design falls under this category. As a research design, both the descriptive and diagnostic studies share common requirements, and hence they may grouped together. However, the procedure to be used must be planned carefully, and so the research design should also be planned carefully. The research design must also make appropriate provision for protection against bias and thus maximize reliability, with due regard to the completion of the research study in as economical manner as possible. The research design in such studies should be rigid and not flexible. Besides, it must also focus attention on the following:

(a) formulation of the objectives of the study,

(b) proper designing of the methods of data collection ,
(c) sample selection,
(d) data collection,
(e) processing and analysis of the collected data, and
(f) Reporting the findings.

3. Hypothesis-testing research design:

Hypothesis-testing research designs are those in which the researcher tests the hypothesis of causal relationship between two or more variables. These studies require procedures that would not only decrease bias and enhance reliability, but also facilitate deriving inferences about the causality. Generally, experiments satisfy such requirements. Hence, when research design is discussed in such studies, it often refers to the design of experiments.

1.4.4 Importance of research design

The need for a research design arises out of the fact that it facilitates the smooth conduct of the various stages of research. It contributes to making research as efficient as possible, thus yielding the maximum information with minimum effort, time and expenditure. A research design helps to plan in advance of the methods to be employed for collecting the relevant data and the techniques to be adopted for their analysis, so as to pursue the objectives of the research in the best possible manner, given the available staff, time and money. Hence, the research design should be prepared with utmost care, so as to avoid any error that may disturb the entire project. Thus, research design plays a crucial role in attaining the reliability of the results obtained, which forms the strong foundation of the entire process of the research work.

Despite its significance, the purpose of a well-planned design is not realized at times. This is because it is not given the importance that this problem deserves. As a consequence, many researchers are not able to achieve the purpose for which the research designs are formulated, due to which they end up

arriving at misleading conclusions. Therefore, faulty designing of the research project tends to render the research exercise meaningless. This makes it imperative that an efficient and suitable research design must be planned before commencing the process of research. The research design helps the researcher to organize his/her ideas in a proper form, which would facilitate him/her to identify the inadequacies and faults in them. The research design may also be discussed with other experts for their comments and critical evaluation, without which it would be difficult for any critic to provide a comprehensive review and comment on the proposed study.

1.4.5 Characteristics of a good research design

A good research design often possesses the qualities such as being flexible, suitable, efficient, economical, and so on. Generally, a research design which minimizes bias and maximizes the reliability of the data collected and analysed is considered a good design (Kothari 1988).

A research design which involves the smallest experimental error is said to be the best design for investigation. Further, a research design that yields maximum information and provides an opportunity of viewing the various dimensions of a research problem is considered to be the most appropriate and efficient design. Thus, the question of a good design relates to the purpose or objective and nature of the research problem studied. While a research design may be good, it may not be equally suitable to all studies. In other words, it may be lacking in one aspect or the other in the case of some other research problems. Therefore, no single research design can be applied to all types of research problems.

A research design suitable for a specific research problem would usually involve the following considerations:

(i) the methods of gathering the information;

(ii) the skills and availability of the researcher and his/her staff, if any;
(iii) the objectives of the research problem being studied;
(iv) the nature of the research problem being studied; and
(v) the available monetary funds and time duration for the research work.

1.5 Case Study Research

The method of exploring and analyzing the life or functioning of a social or economic unit, such as a person, a family, a community, an institution, a firm or an industry, is called a case study method. The objective of a case study method is to examine the factors that cause the behavioural patterns of a given unit and its relationship with the environment. The data for a study are always gathered with the purpose of tracing the natural history of a social or economic unit, and its relationship with the social or economic factors, besides the forces involved in its environment. Thus, a researcher conducting a study using the case study method attempts to understand the complexity of factors that are operative within a social or economic unit as an integrated totality. Burgess (Kothari 1988) described the special significance of the case study in understanding the complex behaviour and situations in specific detail. In the context of social research, he called these data as a social microscope.

1.5.1 Criteria for evaluating adequacy of case study

John Dollard (Dollard 1935) specified seven criteria for evaluating the adequacy of a case or life history in the context of social research. They are as follows: -

(i) The subject being studied must be viewed as a specimen in a cultural set up. That is, the case selected from its total context for the purpose of study should be considered a member of the particular cultural group or community. The scrutiny of the life history of the individual must be carried out with a view to identify the community values, standards and shared ways of life.

(ii) The organic motors of action should be socially relevant. This is to say that the action of the individual cases should be viewed as a series of

reactions to social stimuli or situations. Putting in simple words, the social meaning of behaviour should be taken into consideration.

(iii) The crucial role of the family-group in transmitting the culture should be recognized. This means that as the individual is a member of a family, the role of the family in shaping his/her behaviour should never be ignored.

(iv) The specific method of conversion of organic material into social behaviour should be clearly demonstrated. For instance, case-histories that discuss in detail how basically a biological organism, that is man, gradually transform into a social person are particularly important.

(v) The constant transformation of character of experience from childhood to adulthood should be emphasised. That is, the life-history should portray the inter-relationship between the individual's various experiences during his/her life span. Such a study provides a comprehensive understanding of an individual's life as a continuum.

(vi) The 'social situation' that contributed to the individual's gradual transformation should carefully and continuously specified as a factor. One of crucial the criteria for life-history is that an individual's life should be depicted as evolving itself in the context of a specific social situations and partially caused by it.

(vii) The life-history details themselves should be organized according to some conceptual framework, which in turn would facilitate their generalizations at higher levels.

These criteria discussed by Dollard emphasise the specific link of co-ordinated, related, continuous and configured experience in a cultural pattern that motivated the social and personal behaviour. Although, the criteria indicated by Dollard are principally perfect, but some of them are difficult to put to practice.

Dollard (1935) attempted to express the diverse events depicted in the life-histories of persons during the course of repeated interviews by utilizing psycho-analytical techniques in a given situational context. His criteria of life-

history originated directly from this experience. While the life-histories possess independent significance as research documents, the interviews recorded by the investigators can afford, as Dollard observed, "rich insights into the nature of the social situations experienced by them".

It is a well-known fact that an individual's life is very complex. Till date there is hardly any technique that can establish in some kind of uniformity, and as a result ensure the cumulative of case-history materials by isolating the complex totality of a human life. Nevertheless, although case history data are difficult to put to rigorous analysis, a skilful handling and interpretation of such data could help in developing insights into cultural conflicts and problems arising out of cultural-change.

Gordon Allport (Kothari 1988) has recommended the following aspects so as to broaden the perspective of case-study data as follows:

(i) if the life-history is written in first person, it should be as comprehensive and coherent as possible.

(ii) Life-histories must be written for knowledgeable persons. That is, if the enquiry of study is sociological in nature, the researcher should write it on the assumption that it would be read largely by sociologists only.

(iii) It would be advisable to supplement case study data by observational, statistical and historical data, as they provide standards for assessing the reliability and consistency of the case study materials. Further, such data offer a basis for generalizations.

(iv) Efforts must be made to verify the reliability of life-history data by examining the internal consistency of the collected material, and by repeating the interviews with the person, besides having personal interviews with the persons of the subject's own group who are well-acquainted with him/her.

(v) A judicious combination of different techniques for data-collection is crucial for collecting data that are culturally meaningful and scientifically significant.

(vi) Life-histories or case-histories may be considered as an adequate basis for generalization to the extent that they are typical or representative of a certain group.

(vii) The researcher engaged in the collection of case study data should never ignore the unique or atypical cases. He/she should include them as exceptional cases.

Case histories are filled with valuable information of a personal or private nature. Such information not only help the researcher to portray the personality of the individual, but also the social background that contributed to it. Besides, it also helps in the formulation of relevant hypotheses. In general, although Blummer (in Wilkinson and Bhandarkar 1979) was critical of documentary materials, he gave due credit to case histories by acknowledging the fact that the personal documents offer an opportunity to the researcher to develop his/her spirit of enquiry. The analysis of a particular subject would be more effective if the researcher acquires close acquaintance with it through personal documents. However, Blummer also acknowledges the limitations of the personal documents. According to him, independently such documents do not entirely fulfill the criteria of adequacy, reliability, and representativeness. Despite these shortcomings, avoiding their use in any scientific study of personal life would be wrong, as these documents become necessary and significant for both theory-building and practice.

In spite of these formidable limitations, case study data are used by anthropologists, sociologists, economists and industrial psychiatrists. Gordon Allport (Kothari 1988) strongly recommends the use of case study data for in-depth analysis of a subject. For, it is one's acquaintance with an individual that

instills desire to know his/her nature and understand them. The first stage involves understanding the individual and all the complexity of his/her nature. Any haste in analyzing and classifying the individual would create the risk of reducing his/her emotional world into artificial bits. As a consequence, the important emotional organizations, anchorages, and natural identifications characterizing the personal life of the individual might not yield adequate representation. Hence, the researcher should understand the life of the subject. Therefore, the totality of life-processes reflected in the well-ordered life-history documents become invaluable source of stimulating insights. Such life-history documents provide the basis for comparisons that contribute to statistical generalizations and help to draw inferences regarding the uniformities in human behaviour, which are of great value. Even if some personal documents do not provide ordered data about personal lives of people, which is the basis of psychological science, they should not be ignored. This is because the final aim of science is to understand, control and make predictions about human life. Once they are satisfied, the theoretical and practical importance of personal documents must be recognized as significant. Thus, a case study may be considered as the beginning and the final destination of abstract knowledge.

1.6 Hypothesis

"Hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of phenomenon either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts" (Kothari 1988). A research hypothesis is quite often a predictive statement, which is capable of being tested using scientific methods that involve an independent and some dependent variables. For instance, the following statements may be considered:

i) “students who take tuitions perform better than the others who not receive tuitions” or,
ii) “the female students perform as well as the male students”.
These two statements are hypotheses that can be objectively verified and tested. Thus, they indicate that a hypothesis states what one is looking for. Besides, it is a proposition that can be put to test in order to examine its validity.

1.6.1 Characteristics of hypothesis:

A hypothesis should have the following characteristic features:-

(i) a hypothesis must be precise and clear . If it is not precise and clear, then the inferences drawn on its basis would not be reliable.

(ii) a hypothesis must be capable of being put to test. Quite often, the research programmes fail owing to its incapability of being subject to testing for validity. Therefore, some prior study may be conducted by the researcher in order to make a hypothesis testable. A hypothesis “is tested if other deductions can be made from it, which in turn can be confirmed or disproved by observation” (Kothari 1988).

(iii) a hypothesis must state relationship between two variable, in the case of relational hypotheses.

(iv) a hypothesis must be specific and limited in scope. This is because a simpler hypothesis generally would be easier to test for the research. And therefore, he/she must formulate such hypotheses.

(v) as far as possible, a hypothesis must be stated in the most simple language, so as to make it understood by all concerned. However, it should be noted that simplicity of a hypothesis is not related to its significance.

(vi) a hypothesis must be consistent and derived from the most known facts. In other words, it should be consistent with a substantial body of established facts. That is, it must be in the form of a statement which judges accept as being the most likely to occur.

(vii) a hypothesis must be amenable to testing within a stipulated or reasonable period of time. No matter how excellent a hypothesis, a researcher should not use if it cannot be tested within a given period of time, as none can afford to spend a life-time on collecting data to test it.

(viii) a hypothesis should state the facts that gave rise to the necessity of looking for an explanation. This is to say that by using the hypothesis, and other known and accepted generalizations, a researcher must be able to derive the original problem condition. Therefore, a hypothesis should explain what it actually wants to explain, and for this it should also have an empirical reference.

1.6.2 Concepts relating to testing of hypotheses

Testing of hypotheses requires a researcher to be familiar with various concepts concerned with it. They are discussed here.

1) Null hypothesis and alternative hypothesis:

In the context of statistical analysis, hypothesis is of two types, viz., null hypothesis and alternative hypothesis. When two methods A and B are compared on their relative superiority, and it is assumed that both the methods are equally good, then such a statement is called as the null hypothesis. On the other hand, if method A is considered relatively superior to method B, or vice-versa, then such a statement is known as an alternative hypothesis. The null hypothesis is expressed as H_0, while the alternative hypothesis is expressed as H_a. For example, if a researcher wants to test the hypothesis that the population mean (μ) is equal to the hypothesized mean (H_0) = 100, then the null hypothesis should be stated as the population mean is equal to the hypothesized mean 100. Symbolically it may be written as:-

$$H_0: \quad = \quad \mu \; = \mu \; H_0 \; = 100$$

If sample results do not support this null hypothesis, then it should be concluded that something else is true. The conclusion of rejecting the null hypothesis is called as alternative hypothesis. To put it in simple words, the set of alternatives to the null hypothesis is termed as the alternative hypothesis. If H_0 is accepted, then it implies that H_a is being rejected. On the other hand, if H_0 is rejected, it means that H_a is being accepted. For H_0: $\mu \; = \mu \; H_0 \; = 100$, the following three possible alternative hypotheses may be considered (Kothari 1988).

Alternative hypothesis	to be read as follows
H_a: $\mu \neq \mu \; H_0$	the alternative hypothesis is that the population mean is not equal to 100, i.e., it could greater than or less than 100
H_a : $\mu > \mu \; H_0$	the alternative hypothesis is that the population mean is greater than 100
H_a : $\mu < \mu \; H_0$	the alternative hypothesis is that the population mean is less than 100

Before the sample is drawn, the researcher has to state the null hypothesis and the alternative hypothesis. While formulating the null hypothesis, the following aspects need to be considered:

(a) alternative hypothesis is usually the one which a researcher wishes to prove, whereas the null hypothesis is the one which he/she wishes to disprove. Thus, a null hypothesis is usually the one which a researcher tries to reject, while an alternative hypothesis is the one that represents all other possibilities.

(b) the rejection of a hypothesis when it is actually true involves great risk, as it indicates that it is a null hypothesis because then the probability of rejecting it when it is true is α (i.e., the level of significance) which is chosen very small.

(c) Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value.

(2) The level of significance:

In the context of hypothesis testing, the level of significance is a very important concept. It is a certain percentage that should be chosen with great care, reason and thought. If for instance, the significance level is taken at 5 per cent, then it means that H_0 would be rejected when the sampling result has a less than 0.05 probability of occurrence when H_0 is true. In other words, the five per cent level of significance implies that the researcher is willing to take a risk of five per cent of rejecting the null hypothesis, when (H_0) is actually true. In sum, the significance level reflects the maximum value of the probability of rejecting H_0 when it is actually true, and which is usually determined prior to testing the hypothesis.

(3) Test of hypothesis or decision rule

Suppose that the given hypothesis is H_0 and the alternative hypothesis Ha, then the researcher has to make a rule known as the decision rule. According to the decision rule, the researcher accepts or rejects H_0. For example, if the H_0 is that certain students are good against the H_a that all the students are good, then the researcher should decide the number of items to be tested and the criteria on the basis of which to accept or reject the hypothesis.

(4) Type I and Type II errors

As regards the testing of hypotheses, a research can make basically two types of errors. He/she may reject H_0 when it is true, or accept H_0 when it is not true. The former is called as Type I error and the latter is known as Type II error. In other words, Type I error implies the rejection of a hypothesis when it must have been accepted, while Type II error implies the acceptance of a hypothesis which

must have been rejected. Type I error is denoted by α (alpha) and is known as α error, while Type II error is usually denoted by β (beta) and is known as β error.

(5) One-tailed and two-tailed tests

These two types of tests are very important in the context of hypothesis testing. A two-tailed test rejects the null hypothesis, when the sample mean is significantly greater or lower than the hypothesized value of the mean of the population. Such a test is suitable when the null hypothesis is some specified value, the alternative hypothesis is a value that is not equal to the specified value of the null hypothesis.

1.6.3 Procedure of hypothesis testing

Testing a hypothesis refers to verifying whether the hypothesis is valid or not. Hypothesis testing attempts to check whether to accept or not to accept the null hypothesis. The procedure of hypothesis testing includes all the steps that a researcher undertakes for making a choice between the two alternative actions of rejecting or accepting a null hypothesis. The various steps involved in hypothesis testing are as follows:-

(i) Making a formal statement: This step involves making a formal statement of the null hypothesis (H_0) and the alternative hypothesis (H_a). This implies that the hypotheses should be clearly stated within the purview of the research problem. For example, suppose that a school teacher wants to test the understanding capacity of the students which must be rated more than 90 per cent in terms of marks. In this case, the hypotheses may be stated as follows:-

Null Hypothesis H_0 : = 100

Alternative Hypothesis H_a : > 100

(ii) Selecting a significance level: The hypotheses should be tested on a pre-determined level of significance, which should be specified. Usually, either

5% level or 1% level is considered for the purpose. The factors that determine the levels of significance are: (a) the magnitude of difference between the sample means; (b) the sample size: (c) the variability of measurements within samples; and (d) whether the hypothesis is directional or non-directional (Kothari 1988). In sum, the level of significance should be sufficient in the context of the nature and purpose of enquiry.

(iii) Deciding the distribution to use: After making decision on the level of significance for hypothesis testing, the research has to next determine the appropriate sampling distribution. The choice to be made generally relates to normal distribution and the t-distribution. The rules governing the selection of the correct distribution are similar to the ones already discussed with respect to estimation.

(iv) Selection of a random sample and computing an appropriate value: Another step involved in hypothesis testing is the selection of a random sample and then computing a suitable value from the sample data relating to test statistic by using the appropriate distribution. In other words, it involves drawing a sample for furnishing empirical data.

(v) Calculation of the probability: The next step for the researcher is to calculate the probability that the sample result would diverge as far as it can from expectations, under the situation when the null hypothesis is actually true.

(vi) Comparing the probability: Another step involved consists of making a comparison of the probability calculated with the specified value for α, the significance level. If the calculated probability works out to be equal to or smaller than the α value in case of one-tailed test, then the null hypothesis is to be rejected. On the other hand, if the calculated probability is greater, then the null hypothesis is to be accepted. In case the null hypothesis H_0 is rejected, the researcher runs the risk of committing the Type I error. But, if the null

hypothesis H_0 is accepted, then it involves some risk (which cannot be specified in size as long as H_0 is vague and not specific) of committing the Type II error.

1.7 Sample Survey

A sample design is a definite plan for obtaining a sample from a given population (Kothari 1988). Sample constitutes a certain portion of the population or universe. Sampling design refers to the technique or the procedure the researcher adopts for selecting items for the sample from the population or universe. A sample design helps to decide the number of items to be included in the sample, i.e., the size of the sample. The sample design should be determined prior to data collection. There are different kinds of sample designs which a researcher can choose. Some of them are relatively more precise and easier to adopt than the others. A researcher should prepare or select a sample design, which must be reliable and suitable for the research study proposed to be undertaken.

1.8.1 Steps in sampling design

A researcher should take into consideration the following aspects while developing a sample design:

(i) Type of universe: The first step involved in developing sample design is to clearly define the number of cases, technically known as the Universe, to be studied. A universe may be finite or infinite. In a finite universe the number of items is certain, whereas in the case of an infinite universe the number of items is infinite (i.e., there is no idea about the total number of items). For example, while the population of a city or the number of workers in a factory comprise finite universes, the number of stars in the sky, or throwing of a dice represent infinite universe.

(ii) Sampling unit: Prior to selecting a sample, decision has to be made about the sampling unit. A sampling unit may be a geographical area like a state, district, village, etc., or a social unit like a family, religious community, school, etc., or it may also be an individual. At times, the researcher would have to choose one or more of such units for his/her study.

(iii) Source list: Source list is also known as the 'sampling frame', from which the sample is to be selected. The source list consists of names of all the items of a universe. The researcher has to prepare a source list when it is not available. The source list must be reliable, comprehensive, correct, and appropriate. It is important that the source list should be as representative of the population as possible.

(iv) Size of sample: Size of the sample refers to the number of items to be chosen from the universe to form a sample. For a researcher, this constitutes a major problem. The size of sample must be optimum. An optimum sample may be defined as the one that satisfies the requirements of representativeness, flexibility, efficiency, and reliability. While deciding the size of sample, a researcher should determine the desired precision and the acceptable confidence level for the estimate. The size of the population variance should be considered, because in the case of a larger variance generally a larger sample is larger required. The size of the population should considered, as it also limits the sample size. The parameters of interest in a research study should also be considered, while deciding the sample size. Besides, costs or budgetary constraint also plays a crucial role in deciding the sample size.

(a) Parameters of interest: The specific population parameters of interest should also be considered while determining the sample design. For example, the researcher may want to be estimating the proportion of persons with certain characteristic in the population, or may be interested in knowing some average

regarding the population. The population may also consist of important sub-groups about whom the researcher would like to make estimates. All such factors have strong impact on the sample design the researcher selects.

(b) Budgetary constraint: From the practical point of view, cost considerations exercise a major influence on the decisions relating to not only the sample size, but also on the type of sample selected. Thus, budgetary constraint could also lead to the adoption of a non-probability sample design.

(c) Sampling procedure: Finally, the researcher should decide the type of sample or the technique to be adopted for selecting the items for a sample. This technique or procedure itself may represent the sample design. There are different sample designs from which a researcher should select one for his/her study. It is clear that the researcher should select that design which, for a given sample size and budget constraint, involves a smaller error.

1.7.2 Criteria for selecting a sampling procedure

Basically, two costs are involved in a sampling analysis, which govern the selection of a sampling procedure. They are:-

(i) the cost of data collection, and
(ii) the cost of drawing incorrect inference from the selected data.

There are two causes of incorrect inferences, namely systematic bias and sampling error. Systematic bias arise out of errors in the sampling procedures. They cannot be reduced or eliminated by increasing the sample size. Utmost, the causes of these errors can be identified and corrected. Generally a systematic bias arises out of one or more of the following factors:

a. inappropriate sampling frame,
b. defective measuring device,
c. non-respondents,
d. indeterminacy principle, and
e. natural bias in the reporting of data.

Sampling errors refers to the random variations in the sample estimates around the true population parameters. Because they occur randomly and likely to be equally in either direction, they are of compensatory type, the expected value of which errors tend to be equal to zero. Sampling error tends to decrease with the increase in the size of the sample. It also becomes smaller in magnitude when the population is homogenous.

Sampling error can be computed for a given sample size and design. The measurement of sampling error is known as 'precision of the sampling plan'. When the sample size is increased, the precision can be improved. However, increasing the sample size has its own limitations. The large sized sample not only increases the cost of data collection, but also increases the systematic bias. Thus, an effective way of increasing the precision is generally to choose a better sampling design, which has smaller sampling error for a given sample size at a specified cost. In practice, however, researchers generally prefer a less precise design owing to the ease in adopting the same, in addition to the fact that systematic bias can be controlled better way in such designs.

In sum, while selecting the sample a researcher should ensure that the procedure adopted involves a relatively smaller sampling error and helps to control systematic bias.

1.7.3 Characteristics of a good sample design

The following are the characteristic features of a good sample design:

(a) the sample design should yield a truly representative sample;
(b) the sample design should be such that it results in small sampling error;
(c) the sample design should be viable in the context of budgetary constraints of the research study;
(d) the sample design should be such that the systematic bias can be controlled; and
(e) the sample must be such that the results of the sample study would be applicable, in general, to the universe at a reasonable level of confidence.

1.7.4 Different types of sample designs

Sample designs may be classified into different categories based on two factors, namely, the representation basis and the element selection technique. Under the representation basis, the sample may be classified as:-

I. non-probability sampling
II. probability sampling

While probability sampling is based on random selection, the non-probability sampling is based on 'non-random' sampling.

I. Non-probability sampling:

Non-probability sampling is the sampling procedure that does not afford any basis for estimating the probability that each item in the population would have an equal chance of being included in the sample. Non-probability sampling is also known as deliberate sampling, judgment sampling and purposive sampling. Under this type of sampling, the items for the sample are deliberately chosen by the researcher; and his/her choice concerning the choice of items remains supreme. In other words, under non-probability sampling the researchers select a particular unit of the universe for forming a sample on the basis that the small number that is thus selected out of a huge one would be typical or representative of the whole population. For example, to study the economic conditions of people living in a state, a few towns or village may be purposively selected for an intensive study based on the principle that they are representative of the entire state. In such a case, the judgment of the researcher of the study assumes prime importance in this sampling design.

Quota sampling: Quota sampling is also an example of non-probability sampling. Under this sampling, the researchers simply assume quotas to be

filled from different strata, with certain restrictions imposed on how they should be selected. This type of sampling is very convenient and is relatively less expensive. However, the samples selected using this method certainly do not satisfy the characteristics of random samples. They are essentially judgements samples and inferences drawn based on the would not be amenable to statistical treatment in a formal way.

II. Probability Sampling:

Probability sampling is also known as 'choice sampling' or 'random sampling'. Under this sampling design, every item of the universe has an equal chance of being included in the sample. In a way, it is a lottery method under which individual units are selected from the whole group, not deliberately, but by using some mechanical process. Therefore, only chance determines whether an item or the other would be included in the sample or not. The results obtained from probability or random sampling would be assured in terms of probability. That is, the researcher can measure the errors of estimation or the significance of results obtained from the random sample. This is the superiority of random sampling design over the deliberate sampling design. Random sampling satisfies the law of Statistical Regularity, according to which if on an average the sample chosen is random, then it would have the same composition and characteristics of the universe. This is the reason why the random sampling method is considered the best technique of choosing a representative sample.

The following are the implications of the random sampling:

(i) it provides each element in the population an equal probability chance of being chosen in the sample, with all choices being independent of one another; and

(ii) it offers each possible sample combination an equal probability opportunity of being selected.

1.7.5 Method of selecting a random sample

The process of selecting a random sample involves writing the name of each element of a finite population on a slip of paper and putting them into a box or a bag. Then they have to be thoroughly mixed and then the required number of slips for the sample should be picked one after the other without replacement. While doing this, it has to be ensured that in successive drawings each of the remaining elements of the population has an equal chance of being chosen. This method would result in the same probability for each possible sample.

1.7.6 Complex random sampling designs

Under restricted sampling technique, the probability sampling may result in complex random sampling designs. Such designs are known as mixed sampling designs. Many of such designs may represent a combination of non-probability and probability sampling procedures in choosing a sample. Few of the prominent complex random sampling designs are as follows:

(i) Systematic sampling: In some cases, the best way of sampling is to select every ith item on a list. Sampling of this kind is called as systematic sampling. An element of randomness is introduced in this type of sampling by using random numbers to select the unit with which to start. For example, if a 10 per cent sample is required, the first item would be selected randomly from the first and thereafter every 10^{th} item. In this kind of sampling, only the first unit is selected randomly, while rests of the units of the sample are chosen at fixed intervals.

(ii) Stratified sampling: When a population from which a sample is to be selected does not comprise a homogeneous group, stratified sampling technique is generally employed for obtaining a representative sample. Under stratified sampling, the population is divided into many sub-populations in such a manner that they are individually more homogeneous than rest of the total population.

Then, items are selected from each stratum to form a sample. As each stratum is more homogeneous than the remaining total population, the researcher would be able to obtain a more precise estimate for each stratum and by estimating more accurately each of the component parts, he/she is able to obtain a better estimate of the whole. In some stratified sampling method yields a more reliable and detailed information.

(iii) Cluster sampling: When the total area of research interest is large, a convenient way in which a sample may be selected is to divide the area into a number of smaller non-overlapping areas and then randomly selecting a number of such smaller areas. In the process, the ultimate sample would consist of all the units in these small areas or clusters. Thus in cluster sampling, the total population is sub-divided into numerous relatively smaller subdivisions, which in themselves constitute clusters of still smaller units. And then, some of such clusters would be randomly chosen for inclusion in the overall sample.

(iv) Area sampling: When clusters are in the form of some geographic subdivisions, then cluster sampling is termed as area sampling. That is, when the primary sampling unit represents a cluster of units based on geographic area, the cluster designs are distinguished as area sampling. The merits and demerits of cluster sampling is equally applicable to area sampling.

(iv) Multi-stage sampling: A further development of the principle of cluster sampling is multi-stage sampling. When the researcher desires to investigate the working efficiency of nationalized banks in India and a sample of few banks is required for this purpose, the first stage would be to select large primary sampling unit like the states in the country. Next, certain districts may be selected and all banks interviewed in the chosen districts. This represents a two-stage sampling design, with the ultimate sampling units being clusters of districts.

On the other hand, if instead of taking census of all banks within the selected districts, the researcher chooses certain towns and interviews all banks in it, this would represent three-stage sampling design. Again, if instead of taking a census of all banks within the selected towns, the researcher randomly selects sample banks from each selected town, then it represents a case of using a four-stage sampling plan. Thus, if the researcher selects randomly at all stages, then it is called as multi-stage random sampling design.

(vi) Sampling with probability proportional to size: When the case of cluster sampling units does not have exactly or approximately the same number of elements, it is better for the researcher to adopt a random selection process, where the probability of inclusion of each cluster in the sample tends to be proportional to the size of the cluster. For this, the number of elements in each cluster has to be listed, irrespective of the method used for ordering it. Then the researcher should systematically pick the required number of elements from the cumulative totals. The actual numbers thus chosen would not however reflect the individual elements, but would indicate as to which cluster and how many from them are to be chosen by using simple random sampling or systematic sampling. The outcome of such sampling is equivalent to that of simple random sample. The method is also less cumbersome and is also relatively less expensive.

Thus, a researcher has to pass through various stages of conducting research once the problem of interest has been selected. Research methodology familiarizes a researcher with the complex scientific methods of conducting research, which yields reliable results that are useful to policy-makers, government, industries, etc., in decision-making.

References:

Claire Sellitiz and others, **Research Methods in Social Sciences**, 1962, p.50
Dollard,J., **Criteria for the life-history,** Yale University Press, New York,1935, pp.8-31.
C.R. Kothari, **Research Methodology, Methods and Techniques,** Wiley Eastern Limited, New Delhi, 1988.
Marie Jahoda, Morton Deutsch and Staurt W. Cook, **Research Methods in Social Relations**, p.4.
Pauline V. Young, **Scientific Social Surveys and Research**, p.30
L.V. Redman and A.V.H. Mory, **The Romance of Research**, 1923.
The Encylopaedia of Social Sciences, Vol. IX, MacMillan, 1930.
T.S. Wilkinson and P.L. Bhandarkar, **Methodology and Techniques of Social Research**, Himalaya Publishing House, Bombay, 1979.

Questions:

1. Define research.
2. What are the objectives of research?
3. State the significance of research.
4. What is the importance of knowing how to do research?
5. Briefly outline research process
6. Highlight the different research approaches.
7. Discuss the qualities of a researcher.
8. Explain the different types of research.
9. What is a research problem?
10. Outline the features of research design.
11. Discuss the features of a good research design.
12. Describe the different types of research design.
13. Explain the significance of research design.
14. What is a case study?
15. Discuss the criteria for evaluating case study.
16. Define hypothesis.
17. What are the characteristic features of a hypothesis?
18. Distinguish between null and alternative hypothesis.
19. Differentiate Type I error and Type II error.
20. How is a hypothesis tested?
21. Define the concept of sampling design.
22. Describe the steps involved in sampling design.
23. Discuss the criteria for selecting a sampling procedure.
24. Distinguish between probability and non-probability sampling.
25. How is a random sample selected?
26. Explain complex random sampling designs.

***

UNIT—II DATA COLLECTION

LESSON 1

SOURCES OF DATA

LESSON OUTLINE

- **Primary data-**
- **Methods of collecting primary data-**
- **Direct personal investigation**
- **Indirect oral interviews**
- **Information received through local agencies**
- **Mailed questionnaire method**
- **Schedules sent through enumerators**

Learning Objectives

After reading this lesson you should be able to

- **Understand the meaning of primary data**
- **Preliminaries of data collection**
- **Method of data collection**
- **Methods of collecting primary data**
- **Usefulness of primary data**
- **Merits and demerits of different methods of primary data collection**
- **Pre cautions while collecting primary data.**

Introduction

It is important for a researcher to know the sources of data which he requires for his different purposes. Data are nothing but the information. There are two sources of information or to say data- Primary data and Secondary data. Primary data mean the data collected for the first time, whereas secondary data mean the data that have already been collected and used earlier by somebody or some agency. For example, the statistics collected by the Government of India relating to the population, are primary data for the Government of India since it has been collected for the first time. Later when the same data are used by a researcher for his study of a particular problem, then the same data become the secondary data for the researcher.

Both the sources of information have their merits and demerits. The selection of a particular source depends upon–(a) Purpose and scope of enquiry ; (b) availability of time ;(c) availability of finance and;(d) accuracy required. (e) Statistical units to be used (f) Sources of information (data) and (g) Method of data collection. Let us discuss the above points in short.

(a) Purpose and scope of enquiry:-The purpose and scope of data collection or survey should be clearly set out at the very beginning. It requires the clear statement of the problem indicating the type of information which is needed and the use to which it is needed .If for example, the researcher is interested in knowing the nature of price change over a period of time, it would be necessary to collect data of commodity prices and it must be decided whether it would be helpful to study wholesale or retail prices and the possible uses to which such information could be put. The objective of an enquiry may be either to collect specific information relating to a problem or adequate data to test a

hypothesis. Failure to set out clearly the purpose of enquiry is bound to lead to confusion and waste of resources.

After the purpose of enquiry has been clearly defined, the next step is to decide about the scope of the enquiry. Scope of the enquiry means the coverage with regard to the type of information, the subject-matter and geographical area. For instance, an enquiry may relate to India as a whole or a state or an industrial town where in a particular problem related to a particular industry can be studied.

(b)Availability of time:- The investigation should be carried out within a reasonable period of time; otherwise the information collected may become outdated, and have no meaning at all. For instance, if a producer wants to know the expected demand of a product newly launched by him and the result of the enquiry that the demand would be meager, takes two years to reach to him then the whole purpose of enquiry would become useless because by that time he would have already received a huge loss. Thus in this respect the information is quickly required and hence the researcher has to choose the type of enquiry accordingly.

I Availability of resources:- The investigation will greatly depend on the resources available like number of skilled personnel, the amount etc. If the number of skilled personnel who will carry out the enquiry is quite sufficient and the amount is not a problem then the enquiry can be conducted over a big area covering a good number of samples otherwise a small sample size will do.

(d)The degree of accuracy desired:- Deciding the degree required is must for the investigator, because absolute accuracy in statistical work is seldom achieved. This is so because (a) statistics are based on estimates, (b) tools of measurement are not always perfect and (c) there may be unintentional bias on the part of the investigator,, enumerator or informant. Therefore, a desire of

100% accuracy is bond to remain unfulfilled. Degree of accuracy desired primarily depends upon the object of enquiry. For example when we buy gold even a difference of $1/10^{th}$ gram in its weight is significant whereas the same will not be the case when we buy rice or wheat. However, the researcher must aim at attaining a higher degree of accuracy otherwise the whole purpose of research would become meaningless.

(e) Statistical Units to be used: A well defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called a *statistical unit.* For example, in socio-economic survey the unit may be an individual person, a family, a household or a block of locality. A very important step before the collection of data begins is to define clearly the statistical units on which the data are to be collected. In number of situations the units are conventionally fixed like the physical units of measurement such as metres, kilometers, quintals, hours, days, week etc., which are well defined and do not need any elaboration or explanation. However in many statistical investigations, particularly relating to socio-economic studies, arbitrary units are used which must be clearly defined. This is must because in the absence of a clear cut and precise definition of the statistical units, serious errors in the data collection may be committed in the sense that we may collect irrelevant data on the items, which should have, in fact, been excluded and omit data on certain items which should have been included. This will ultimately lead to fallacious conclusions.

(f) Sources of information (data):- After decided about the unit, a researcher has to decide about the source from which the information can be obtained or collected. For any statistical inquiry, the investigator may collect the data first hand or he may use the data from other published sources such as the

publications of the government/semi-government organizations or journals and magazines etc.

(g) Method of data collection:- There is no problem if secondary data are used for the research . However, if primary data are to be collected a decision has to be taken whether (i) census method or (ii) sample technique, is to be used for data collection .In census method we go for total enumeration i.e. all the units of a universe have to be investigated. But in sample technique, we inspect or study only a selected representative and adequate fraction of the population and after analyzing the results of the sample data we draw conclusions about the characteristics of the population. Selection of a particular technique becomes difficult because where population or census method is more scientific and 100% accuracy can be attained through this method, choosing this becomes difficult because it is time taking, it requires more labor and after all it is very expensive. Therefore, for a single researcher or for a small institution it proves to be unsuitable. On the other hand, sample method is less time taking, less laborious and less expensive but a 100% accuracy cannot be attained through this method because of sampling and non sampling errors attached to this method. Hence, a researcher has to be very cautious and careful while choosing a particular method.

Methods of collecting Primary data

Primary data may be obtained by applying any of the following methods-

1. Direct Personal Interviews
2. Indirect oral interviews.
3. Information from correspondents.
4. Mailed questionnaire methods.
5. Scheduled sent through enumerators.

1. Direct personal interviews:- A face to face contact is made with the informants(persons from whom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information. Thus, if a person wants to collect data about the working conditions of the workers of the Tata Iron and Steel Company, Jamshedpur, he would go to the factory, contact the workers and obtain the desired information. The information collected in this manner is first hand and also original in character.

There are many merits and demerits of this method which is discussed below:-

Merits

1. Most often respondents are happy to pass on the information required from them when contacted personally and thus response is encouraging.
2. The information collected through this method is normally more accurate because the interviewer can clear up doubts of the informants about certain questions and thus obtain correct information. In case the interviewer apprehends that the informant is not giving accurate information, he may cross-examine him and thereby try to obtain the information.
3. This method also provides the scope for getting the supplementary information from the informant because while interviewing it is possible to ask some supplementary questions which may be of great use latter.
4. It is experienced that there are some difficult questions which normally becomes difficult to ask directly but a trained and experienced researcher can sandwiched the difficult questions between other questions and get the desired information. He can twist the questions keeping in mind the informant's reaction. Precisely, a delicate situation can usually he

handled more effectively by a personal interview than by other survey techniques.

5. The interviewer can adjust the language according to the status and educational level of the person interviewed, and thereby can avoid inconvenience and misinterpretation on the part of the informant.

Demerits: There are some demerits or limitations of this method which are explained below:

1. This method can prove to be expensive if the number of informants is large and the area is wide spread
2. There is a greater chance of personal bias and prejudice under this method as compared to other method.
3. The interviewers have to be thoroughly trained and experienced otherwise they may not be able to obtain the desired information. Untrained or poorly trained interviewers may spoil the entire work.
4. This method is more time taking as compared to others. This is because interviews can be held only at the convenience of the informants. Thus, if information is required to be obtained from the working members of households, interviews will have to be held in the evening or on week end. Even during evening only an hour or two can be used for interviews and hence, the work may have to be continued for a long time, or a large staff may have to be employed which may involve huge expense.

Conclusion:-Though there are some demerits in this method of data collection still we cannot say that it is not useful. The matter of fact is that this method is suitable for intensive rather than extensive field surveys. Hence, it should be used only in those cases where intensive study of a limited field is desired.

In the present time of extreme advancement in the communication system, the investigator instead of going personally and conducting a face to face interview may also obtain information on telephone. A good number of surveys are being conducted every day by newspapers and television channels by sending the reply either by e-mail or SMS. This method has become very popular nowadays as it is less expensive and the response is extremely quick. But this method suffers from some serious defects as – (a) very few people own a phone or a television and hence a limited type of people can be approached by this method,(b) only few questions can be asked over phone or through television,(c) the respondents may give a vague and reckless answers because answers on phone or through SMS would have to be very short.

2.Indirect Oral Interviews:- Under this method of data collection, the investigator contacts third parties generally called 'witnesses' who are capable of supplying necessary information. This method is generally adopted when the information to be obtained is of a complex nature and informants are not inclined to respond if approached directly. For example, when the researcher is trying to obtain data on drug addiction or the habit of taking liquor, there is high probability that the addicted person will not supply the desired data and hence disturb the whole research process. In this situation taking the help of such persons or agency or the neighbour who know them well becomes necessary. Since these people know the person well and hence, they can supply the desired data. Enquiry Committees and Commissions appointed by the Government generally adopt this method to get people's views and all possible details of facts relating to the enquiry.

Though this method is very popular, its correctness depends upon a number of factors which is discussed below:-

1. The person or persons or agency whose help is solicited must be of proven integrity otherwise any bias or prejudiced on the part of them will not bring the correct information and the whole process of research will become useless.

2. The ability of the interviewers to draw out the information from witnesses by means of appropriate questions and cross-examination.

3. It might happen that because of bribery, nepotism or certain other reasons those who are collecting the information give it such a twist that correct conclusions are nor arrived at.

Therefore, for the success of this method it is necessary that the evidence of one person alone is not relied upon. Views from other persons and related agencies should also be ascertained to find the real position .Utmost care must be exercised in the selection of these persons because it is on their views that the final conclusions are reached.

3. Information from Correspondents:- The investigator appoints local agents or correspondents in different places to collect information under this method. These correspondents collect and transmit the information to the central office where data are processed. This method is generally adopted by news paper agencies. Correspondents who are posted at different places supply information relating to such events as accidents, riots, strikes, etc., to the head office. The correspondents are generally paid staff or sometimes they may be honorary correspondents also. This method is also adopted generally by the government departments in such cases where regular information is to be collected from a wide area. For example, in the construction of a wholesale price index numbers regular information is obtained from correspondents appointed in different areas. The biggest advantage of this method is that it is cheap and appropriate for extensive investigation. But a word of caution is that it may not always ensure

accurate results because of the personal prejudice and bias of the correspondents.

As already stated earlier, this method is suitable and adopted in those cases where the information is to be obtained at regular intervals from a wide area.

1. **Mailed Questionnaire Method:-** Under this method, a list of questions pertaining to the survey which is known as 'Questionnaire' is prepared and sent to the various informants by post. Sometimes the researcher himself too contacts the respondents and gets the responses relating to the various questions in the questionnaire. The questionnaire contains questions and provides space for answers. A request is made to the informants through a covering letter to fill up the questionnaire and send it back within a specified time.

The questionnaire studies can be classified on the basis of:

i. The degree to which the questionnaire is formalized or structured.

ii. The disguise or lack of disguise of the questionnaire , and

iii. The communication method used.

When no formal questionnaire is used, interviewers adapt their questioning to each interview as it progresses or perhaps elicit responses by indirect methods such as showing pictures on which the respondent comments. When a researcher follows a prescribed sequence of questions, it is referred to as *structured* study. On the other hand, when no prescribed sequence of questions exists, the study is *non-structured.*

When questionnaires are constructed so that the objective is clear to the respondents then these questionnaires are known as *non- disguised*; on the other hand, when the objective is not clear the questionnaire is a *disguised* one. On the basis of these two classifications, four types of studies can he distinguished:

i. Non-disguised structured,

ii. Non-disguised non-structured,

iii. Disguised structured, and

iv. Disguised non-structured.

There are certain merits and demerits or limitations of this method of data collection which are discussed below:

Merits:

2. Questionnaire method of data collection can be easily adopted where the field of investigation is very vast and the informants are spread over a wide geographical area.
3. This method is relatively cheap and expeditious provided the informants respond in time.
4. This method is proved to be superior when a question of a personal nature or questions requiring reaction by the family, than other methods as personal interviews or telephone method.

Demerits:

1. This method can be adopted only where the informants are literate people so hat they can understand written questions and wend the answers in writing.

2. It involves some uncertainty about the response. Co-operation on the part of informants may difficult to presume.

3. The information supplied by the informants may not be correct and in may be difficult to verify the accuracy.

However by following the following guidelines this method can be made more effective.

i. The questionnaire should be made in such a manner that it does not become an undue burden on the respondents; otherwise they may not return them back.

ii. Prepaid postage stamp should be affixed

iii. The sample should be large

iv. It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest in the enquiry.

v. It should be preferred in such enquiries where there could be a legal compulsion to supply the information so that the risk of non-response is eliminate.

5. **Schedules sent through Enumerators:-A**nother method of data collection is through sending schedules through the enumerators or interviewers. The enumerators contact the informants, get replies to the questions contained in a schedule and fill them in their own handwriting in the questionnaire form. There is difference between questionnaire and schedule. Questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in him self, whereas Schedule is the name usually applied to a set of questions which are asked and filed in a face-to face situation with another person. This method is free from most of the limitations of the mailed questionnaire method.

Merits

The main merits or advantages of this method are listed below:-

i. It can be adopted in those cases where informants are illiterate.

ii. There is very little scope of non-response as the enumerators go personally to obtain the information.

iii. The information received is more reliable as the accuracy of statements can be checked by supplementary questions wherever necessary.

This method too like others is not free from defects or limitations. The main limitations are listed below:-

Demerits

i. In comparison to other methods of collecting primary data, this method is quite costly as enumerators are generally paid persons.

ii. The success of the method depends largely upon the training imparted to the enumerators.

iii. Interviewing is a very skilled work and it requires experience and training, hut there is a tendency of statisticians to neglect this extremely important part of the data collecting process. Without good interviewing most of the information collected is of doubtful value.

iv. Interviewing is not only a skilled work but it also requires great degree of politeness and thus the way the enumerators conduct the interview would affect the data collected. When questions are asked by a number of different interviewers, it is possible that variations in the personalities of the interviewers will cause variation in the answers obtained. This variation will not be obvious. Hence every effort must be made to remove as much of variation as possible due to different interviewers.

Secondary Data:-As already stated earlier, secondary data are those data which have been already collected and analyzed by some earlier agency for its own use, and later the same data are used by a different agency. According to W.A.Neiswanger," A primary source is a publication in which the data are published by the same authority which gathered and analyzed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible."

Sources of secondary data:-The various sources of secondary data can be divided into two broad categories:

1. Published sources and,
2. Unpublished sources.

1. **Published Sources**: Various governmental, international and local agencies publish statistical data, and chief among them are explained below:

(a) International Publications: There are some international institutions and bodies like I.M.F, I.B.R.D, I.C.A.F.E, and the U.N.O etc. who publish regular and occasional reports on economic and statistical matters.

(b) Official publications of Central and State Governments: Several departments of the Central and State Governments regularly publish reports on a number of subjects. They gather additional information. Some of the important publications are: the Reserve Bank of India Bulletin, Census of India, Statistical Abstracts of States, Agricultural Statistics of India, Indian Trade Journal, etc.

(c) Semi- official publications: Semi-Government institutions like Municipal Corporations, District Boards, Panchayats, etc. publish reports relating to different matters of public concern.

(d) Publications of Research Institutions: Indian Statistical Institution (I.S.I), Indian Council of Agricultural Research (I.C.A.R), Indian Agricultural Statistics Research Institute (I.A.S.R.I), etc. publish the findings of their research programmes.

(e) Publications of various Commercial and Financial Institutions

(f) Reports of various Committees and Commissions appointed by the Government as the Raj Committee's Report on Agricultural Taxation, Wanchoo Committee's Report on Taxation and Black Money, etc. are also important sources of secondary data.

(g) Journals and News Papers: Journals and News Papers are very important and powerful source of secondary data. Current and important materials on statistics and socio-economic problems can be obtained from journals and newspapers like, Economic Times, Commerce, Capital, Indian Finance, Monthly Statistics of trade etc.

2. **Unpublished Sources:** Unpublished data can be obtained from many unpublished sources like records maintained by various government and private offices, the theses of the numerous research scholars in the universities or institutions etc.

Precautions in the use of Secondary Data: Since secondary data have already been obtained it is highly desirable that a proper scrutiny of such data is made before they are used by the investigator. In fact the user has to be extra-cautious while using secondary data. In this context Prof. Bowley rightly points out that "Secondary data should not be accepted at their face value." The reason being that data may be erroneous in many respects due to bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors etc. Even if there is no error such data may not be suitable and adequate for the purpose of the enquiry. Prof. Simon Kuznet's view in this regard is also of great importance. According to him, "The degree of reliability of secondary source is to be assessed from the source, the compiler and his capacity to produce correct statistics and the users also, for the most part, tend to accept a series particularly one issued by a government agency at its face value without enquiring its reliability".

Therefore, before using the secondary data the investigators should consider the following factors:

6. **The suitability of data**: The investigator must satisfy him self that the data available are suitable for the purpose of enquiry. It can be judged

by the nature and scope of the present enquiry with the original enquiry. For example, if the object of the present enquiry is to study the trend in retail prices, and if the data provide only wholesale prices, such data are unsuitable.

(a) Adequacy of data: If the data are suitable for the purpose of investigation then we must consider whether the data are useful or adequate for the present analysis. It can be studied by the geographical area covered by the original enquiry. The time for which data are available is very important element. In the above example, if our object is to study the retail price trend of India, and if the available data cover only the retail price trend in the State of Bihar, then it would not serve the purpose.

(b) Reliability of data: The reliability of data is must. Without which there is no meaning in research. The reliability of data can be tested by finding out the agency that collected such data. If the agency has used proper methods in collection data, statistics may be relied upon.

It is not enough to have baskets of data in hand. In fact data in a raw form are nothing but a handful of raw material waiting for proper processing so that they can become useful. Once data have been obtained from primary or secondary source, the next step in a statistical investigation is to edit the data i.e. to scrutinize the same. The chief objective of editing is to detect possible errors and irregularities. The task of editing is a highly specialized one and requires great care and attention. Negligence in this respect may render useless the findings of an otherwise valuable study. Editing data collected from internal records and published sources is relatively simple but the data collected from a survey need excessive editing.

While editing primary data following considerations should be born in mind:

1. The data should be complete in every respect
2. The data should be accurate

3. The data should be consistent, and
4. The data should be homogeneous.

Data to posses the above mentioned characteristics have to undergo the same type of editing which are discussed below:

7. **Editing for completeness**: While editing the editor should see that each schedule and questionnaire is complete in all respects. Answers to each and every question have been furnished. If some questions are not answered and they are of vital importance, the informants should be contacted again either personally or through correspondence. Even after all the efforts it may happen that a few questions remain unanswered. In such questions, the editor should mark 'No answer' in the space provided for answers and if the questions are of vital importance then the schedule or questionnaire should be dropped.

1. **Editing for consistency**: At the time of editing the data for consistency, the editor should see that the answers to questions are not contradictory in nature. If the are mutually contradictory answers, he should try to obtain the correct answers either by referring back the questionnaire or by contacting, wherever possible, the informant in person. For example, if amongst others, two questions in questionnaire are (a) Are you a student? (b) Which class do you study and the reply to the first question is 'no' and to the latter 'tenth' then there is contradiction and it should be clarified.

2. **Editing for accuracy**: The reliability of conclusions depends basically on the correctness of information. If the information supplied is wrong, conclusions can never be valid. It is, therefore, necessary for the editor to see that the information is accurate in all respects. If the inaccuracy is due to arithmetical errors, it can be easily detected and corrected. But if

the cause of inaccuracy is faulty information supplied, it may be difficult to verify it e.g. information relating to income, age etc.

3. **Editing for homogeneity:** Homogeneity means the condition in which all the questions have been understood in the same sense. The editor must check all the questions for uniform interpretation. For example, as to the question of income, if some informants have given monthly income, others annual income and still others weekly income or even daily income, no comparison can be made. Therefore, it becomes an essential duty of the editor to check up that the information supplied by the various people is homogeneous and uniform.

Choice between Primary and Secondary Data:-As we have already seen, there are lot of difference in the methods of collecting Primary and Secondary data. In the case of primary data which is to be collected originally, the entire scheme of the plan starting with the definitions of various terms used, units to he employed, type of enquiry to be conducted, extent of accuracy aimed at etc. is to he formulated whereas the collection of secondary data is in the form of mere compilation of the existing data. A proper choice between the type of data needed for any particular statistical investigation is to be made after taking into consideration the nature, objective and scope of the enquiry; the time and the finances at the disposal of the agency; the degree of precision aimed at and the status of the agency (whether government- state or central-or private institution of an individual).

In using the secondary data it is best to obtain the data from the primary source as far as possible. By doing so, we would at least save ourselves from the errors of transcription which might have inadvertently crept in the secondary source. Moreover, the primary source will also provide us with detailed discussion about the terminology used, statistical units employed, size of the sample and the

technique of sampling (if sampling method was used), methods of data collection and analysis of results and we can ascertain ourselves if these suit our purpose.

Now a days in a large number of statistical enquiries secondary data are generally used because fairly reliable published data on a large number of diverse fields are now available in publication of governments, private organizations and research institutions, agencies, periodicals and magazines etc. In fact primary data are collected only if there do not exist any secondary data suited to the investigation under study. In some of the investigations both primary as well as secondary data may be used.

SUMMARY

There are two types of data, Primary and secondary. Data which are collected first hand are called Primary data and data which have already been collected and used by some body or agency are called Secondary data. There are two methods of collecting data. They are- (a) Survey method or total enumeration method and (b) Sample method. When a researcher goes for investigating all the units of the subject, it is called as survey method and on the other hand when resorts to investigating only a few units of the subject and to give the result on the basis of that, it is known as sample survey method. There are different sources of collecting Primary and Secondary data. Some of the important sources of Primary data are—Direct Personal Interviews, Indirect Oral Interviews, Information from correspondents, Mailed questionnaire method, Schedules sent through enumerators. Though all these sources or methods of Primary data have their relative merits and demerits, a researcher should use a particular method with lot of care. There are basically two sources of collecting secondary data- (a) Published sources and (b) Un published sources. Published sources are like publications of different government and semi-government

departments, research institutions and agencies etc. whereas unpublished sources are like records maintained by different government departments and unpublished theses of different universities etc. Editing of secondary data is necessary for different purposes as – editing for completeness, editing for consistency, editing for accuracy and editing for homogeneity.

It is always a tough task for the researcher to choose between primary and secondary data. Though primary data are more authentic and accurate, time, money and labor involved in obtaining these more often prompt the researcher to go for the secondary data. There are certain amount of doubt about its authenticity and suitability, but after the arrival of many government and semi government agencies and some private institutions in the field of data collection, most of the apprehensions in the mind of the researcher have been removed.

SELF ASSESMENT QUESTIONS (SAQs)

1. Explain primary and secondary data and distinguish between them.
   (Refer to the introduction part of this lesson.)
   8. Explain different methods of collection primary data.
   (Explain direct personal, indirect oral interview, information received through agencies etc.)
3. Explain merits and demerits of different methods of collecting primary data.
   (Refer the methods of collecting primary data)
4. Explain the different sources of secondary data and precaution in using secondary data.
5. What is editing of secondary data? Why is it required?
6. What are the different types of editing of secondary data?

GLOSSARY OF TERMS

Primary Source: It is one that itself collects the data.

Secondary Source: It is one that makes available data collected by some other agency.

Collection of Statistics: Collection means the assembling for the purpose of particular investigation of entirely new data presumably not already available in published sources.

Questionnaire: A list of questions properly selected and arranged pertaining to the investigation.

Investigator: Investigator is a person who collects the information.

Respondent: A person who fills the questionnaire or supplies the required information.

***

UNIT II

LESSON **2** **QUESTIONNAIRE AND SAMPLING**

LESSON OUTLINE

- **Meaning of questionnaire.**
- **Drafting of questionnaire.**
- **Size of questions**
- **Clarity of questions**
- **Logical sequence of questions**
- **Simple meaning questions**
- **Other requirements of a good questionnaire**
- **Meaning and essentials of sampling.**

LEARNING OBJECTIVES:

- **After reading this lesson you**
- **should be able to**
- **Understand the meaning of questionnaire**
- **Different requirements and characteristics of a good questionnaire**
- **Meaning of sampling**
- **Essentials of sampling**

Introduction:

Nowadays questionnaire is widely used for data collection in social research. It is a reasonably fair tool for gathering data from large, diverse, varied and scattered social groups. The questionnaire is the media of communication between the investigator and the respondents. According to Bogardus a questionnaire is a list of questions sent to a number of persons for their answers and which obtains standardized results that can be tabulated and treated statistically. The Dictionary of Statistical Terms defines it as a " group of or sequence of questions designed to elicit information upon a subject or sequence of subjects from an information." A questionnaire should be designed or drafted with utmost care and caution so that all the relevant and essential information for the enquiry may be collected without any difficulty, ambiguity and vagueness. Drafting of a good questionnaire is a highly specialized hob and requires great care skill, wisdom, efficiency and experience. No hard and fast rule can be laid down for designing or framing a questionnaire. However, in this connection, the following general points may be borne in mind:

1. **Size of the questionnaire should be small:** A researcher should try his best to keep the number of the questions as small as possible, keeping in view the nature, objectives and scope of the enquiry. Respondent's time should not be wasted by asking irrelevant and unimportant questions. A large number of questions would involve more work for the investigator and thus result in delay on his part in collecting and submitting the information. A large number of unnecessary questions may annoy the respondent and he may refuse to cooperate. A reasonable questionnaire should contain from 15 to 25 questions at large. If a still larger number of questions is a must in any enquiry, then the questionnaire should be divided into various sections or parts.

2. The questions should be clear: The questions should be easier, brief, unambiguous, non-offending, courteous in tone, corroborative in nature and to the point so that much scope of guessing is left on the part of the respondents.

3. The questions should be arranged in a logical sequence: Logical arrangement of questions reduces lot of unnecessary work on the part of the researcher because it not only facilitates the tabulation work but does not leave any chance for omissions or commissions. For example, to find if a person owns a television the logical order of questions would be: Do you own a television? When did you buy it? What is its make? How much did it cost you? Is its performance satisfactory? Have you ever got it serviced?

4. Questions should be simple to understand: The vague words like good, bad, efficient, sufficient, prosperity, rarely, frequently, reasonable, poor, rich etc., should not be used since these may be interpreted differently by different persons and as such might give unreliable and misleading information. Similarly the use of words having double meaning like price, assets, capital income etc., should also be avoided.

5. Questions should be comprehensive and easily answerable: Questions should be so designed that they are readily comprehensible and easy to answer for the respondents. They should not be tedious nor should they tax the respondents' memory. At the same time questions involving mathematical calculations like percentages, ratios etc., should not be asked.

6. Questions of personal nature and sensitive should not be asked: There are some questions which disturb the respondents and he may be shy or irritated by hearing such questions. Therefore, every effort should be made to avoid such questions. For example, do you cook yourself or your wife cooks? Or do you drink? Such questions will certainly irk the respondents and thus be avoided at any cost. If unavoidable then highest amount of politeness should be used

7. Types of questions: Under this head, the questions in the questionnaire may be classified as follows:

(a) Shut questions: Shut questions are those where possible answers are suggested by the framers of the questionnaire and the respondent is required to tick one of them. Shut questions can further be subdivided into the following forms:

(i) Simple alternate questions: In this type of questions the respondent has to choose from the two clear cut alternatives like 'Yes' or 'No' 'Right or Wrong' etc. Such questions are also called *dichotomous questions.* This technique can be applied with elegance to situations where two clear cut alternatives exist.

(ii) Multiple choice questions: Many a times it becomes difficult to define a clear cut alternative and accordingly in such a situation either the first method is not used of additional answers between Yes and No like Do not know, No opinion, Occasionally, Casually, Seldom etc. are added. For example, in order to find if a person smokes or drinks, the following multiple choice answers may be used:

Do you smoke?

(a)Yes regularly [ ] (b) No never [ ]

(c) Occasionally [ ] (d) Seldom [ ]

Multiple choice questions are very easy and convenient for the respondents to answer. Such questions save time and also facilitate tabulation. This method should be used if only a selected few alternative answers exist to a particular question.

8. Leading questions should be avoided: Questions like 'Why do you use a particular type of car, say Maruti car' should preferably be framed into two questions-

(i) Which car do you use?

(ii) Why do you prefer it?

It gives smooth ride [ ]

It gives more mileage [ ]

It is cheaper [ ]

It is maintenance free [ ]

9 Cross Checks: The questionnaire should be so designed as to provide internal checks on the accuracy of the information supplied by the respondents by including some connected questions at least with respect to matters which are fundamental to the enquiry.

10 Pre testing the questionnaire: It would be practical in every sense to try out the questionnaire on a small scale before using it for the given enquiry on a large scale. This has been found extremely useful in practice. The given questionnaire can be improved or modified in the light of the drawbacks, shortcomings and problems faced by the investigator in the pre test.

11 A covering letter: A covering letter from the organizers of the enquiry should be enclosed along with the questionnaire for the purposes of – regarding definitions, units, concepts used in the questionnaire, for taking the respondent's confidence, self addressed envelop in case of mailed questionnaire, mention about award or incentives for the quick response, a promise to send a copy of the survey report etc.

SAMPLING

Though sampling is not new but the sampling theory has been developed recently. People knew or not but they have been using the sampling technique in their day to day life. For example a house wife tests a small quantity of rice to see whether it has been well-cooked and give the generalized result about the whole rice boiling in the vessel. The result arrived at is most of the times 100%

correct. In another example, when a doctor wants to examine the blood for any deficiency, takes only a few drops of blood of the patient and examines. The result arrived at is most of the times correct and represent the whole amount of blood available in the body of the patient. In all these cases, by inspecting a few, they simply believe that the samples give a correct idea about the population. Most of our decision are based on the examination of a few items only i.e. sample studies. In the words of Croxton and Cowdon," It may be too expensive or too time consuming to attempt either a complete or a nearly complete coverage in a statistical study. Further to arrive at valid conclusions, it may not be necessary to enumerate all or nearly all of a population. We may study a sample drawn from the large population and, if that sample is adequately representative of the population, we should be able to arrive at valid conclusions."

According to Rosander," The sample has many advantages over a census or complete enumeration. If carefully designed, the sample is not only considerably cheaper; but may give results which are just accurate and sometimes more accurate than those of a census. Hence a carefully designed sample may actually be better than a poorly planned and executed census."

Merits:

1. **It saves time:** Sampling method of data collection saves time because fewer items are collected and processed. When the results are urgently required, this method is very helpful.
2. **It reduces cost:** Since only a few and selected items are studied in sampling, so there is reduction in cost of money and reduction in terms of man hours.
3. **More reliable results can be obtained:** Through sampling more reliable results can be obtained because (a) there are fewer chances of sampling

statistical errors. If there is sampling error, it possible to estimate and control the results.(b) Highly experienced and trained persons can be employed for scientific processing and analyzing of relatively limited data and they can use their high technical knowledge and get more accurate and reliable results.

4. **It provides more detailed information:** As it saves time, money and labor, more detail information can be collected in a sample survey.
5. **Some times only method to depend upon:** Some times it so happens that one has to depend upon sampling method alone because if the population under study is finite, sampling method is the only method to be used. For example, if some ones blood has to be examined, it will become fatal to take all the blood out from the body and study depending upon the total enumeration method.
6. **Administrative convenience:** The organization and administration of sample survey are easy for the same time, money and labor reasons which have been discussed earlier.
7. **More scientific:** Since the methods used to collect data are based on scientific theory and results obtained can be tested, sampling is more scientific method to collect data.

It is not that sampling is free from demerits or shortcomings. There are certain **shortcomings of this method** which are discussed below:

1. **Illusory conclusion:** If a sample enquiry is not carefully planned and executed, the conclusions may be inaccurate and misleading.
2. **Sample not representative:** To make the sample representative is a difficult task. If a representative sample is taken from the universe, the result is applicable to the whole population. If the sample is not representative of the universe the result may be false and misleading.

3. **Lack of experts:** As there is lack of experts to plan and conduct a sample survey, its execution and analyze, the results of the sample survey are not satisfactory and trustworthy.
4. **Some times more difficult than census method:** Some times the sampling plan bay be complicated and requires more money, labor, time than a census method.
5. **Personal bias:** There may be personal biases and prejudices with regard to the choice of technique and drawing of sampling units.
6. **Choice of sample size:** If the size of the sample is not appropriate then it may lead to untrue characteristics of the population.
7. **Conditions of complete coverage:** If the information is required for each and every item of the universe, then a complete enumeration survey is better.

Essentials of sampling: In order to reach to a clear conclusion, the sampling should possess the following essentials:

1. **It must be representative:** The sample selected should possess the similar characteristics of the original universe from which it has been drawn.
2. **Homogeneity:** Selected samples from the universe should have similar nature and should mot have any difference when compared with the universe.
3. **Adequate samples:** In order to have a more reliable and representative result, a good number of items are to be included in the sample.
4. **Optimization:** All efforts should be made to get maximum results both in terms of cost as well as efficiency. If size of the sample is larger, there is better efficiency and at the same time the cost is more. A proper size

of sample is maintained in order to have optimized results in terms of cost and efficiency.

STATISTICAL LAWS

One of the basic reasons for undertaking a sample survey is to predict and generalize the results for the population as a whole. The logical process of drawing general conclusions from a study of representative items is called induction. In statistics induction is a generalization of facts on the assumption that the results provided by an adequate sample may be taken as applicable to the whole. The fact that the characteristics of the sample provide a fairly good idea about the population characteristics is borne out by the theory of probability. Sampling is based on two fundamental principles of statistics theory viz, (i) the Law of Statistical Regularity and (ii) the Law of Inertia of Large Numbers.

THE LAW OF STATISTICAL REGULARITY

The Law of Statistical Regularity is derived from the mathematical theory of probability. According to W.I.King, "The Law of Statistical Regularity formulated in the mathematical theory of probability lays down that a moderately large number of items chosen at random from a very large group are almost sure on the to have the characteristics of the large group." For example, if we want to find out the average income of 10,000 people, we take a sample of 100 people and find the average. Suppose that another person takes another sample of 100 people from the same population and finds the average. The average income found out by both the persons will have the least difference. On

the other hand if the average income of the same 10,000 people is found out by the census method, the result will be more or less same.

Characteristics

1. The item selected will represent the universe and the result is generalized to universe as a whole.
2. Since sample size is large, it is representative of the universe.
3. There is a very remote chance of bias.

LAW OF INERTIA OF LARGE NUMBERS

The Law of inertia of Large Numbers is an immediate deduction from the Principle of Statistical Regularity .Law of Inertia of Large Numbers states," Other things being equal, as the sample size increases, the results tend to be more reliable and accurate." This is based on the fact that the behavior or a phenomenon en masse. i.e., on a large scale is generally stable. It implies that the total change is likely to be very small, when a large number or items are taken in a sample .The law will be true on an average. If sufficient large samples are taken from the patent population, the reverse movements of different parts in the same will offset by the corresponding movements of some other parts.

Sampling Errors: In a sample survey, since only a small portion of the population is studied its results are bound to differ from the census results and thus, have a certain amount of error. In statistics the word error is used to denote the difference between the true value and the estimated or approximated value. This error would always be there no matter that the sample is drawn at random and that it is highly representative. This error is attributed to fluctuations of sampling and is called sampling error. Sampling error is due to the fact that only a sub set of the population has been used to estimate the population parameters

and draw inferences about the population. Thus, sampling error is present only in a sample survey and is completely absent in census method.

Sampling errors occur primarily due to the following reasons:

1. **Faulty selection of the sample:** Some of the bias is introduced by the use of defective sampling technique for the selection of a sample e.g. purposive or judgment sampling in which the investigator deliberately selects a representative sample to obtain certain results. This bias can be easily overcome by adopting the technique of simple random sampling.
2. **Substitution:** When difficulties arise in enumerating a particular sampling unit included in the random sample, the investigators usually substitute a convenient member of the population. This obviously leads to some bias since the characteristics possessed by the substituted unit will usually be different from those possessed by the unit originally included in the sample.
3. **Faulty demarcation of sampling units:** Bias due to defective demarcation of sampling units is particularly significant in area surveys such as agricultural experiments in the field of crop cutting surveys etc. In such surveys, while dealing with border line cases, it depends more or less on the discretion of the investigator whether to include them in the sample or not.
4. **Error due to bias in the estimation method:** Sampling method consists in estimating the parameters of the population by appropriate statistics computed from the sample. Improper choice of the estimation techniques might introduce the error.
5. **Variability of the population:** Sampling error also depends of the variability or heterogeneity of the population to be sampled.

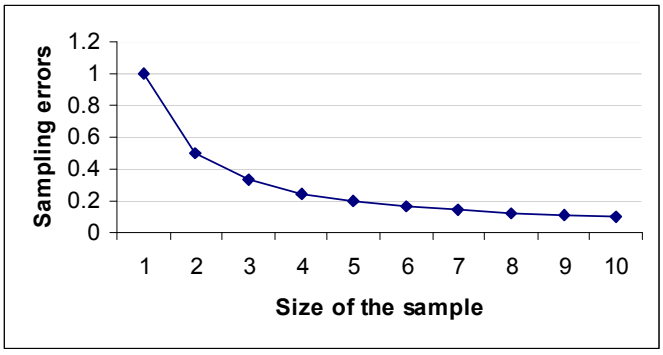
Sampling errors are of two types- Biased Errors and Unbiased Errors

Biased Errors: The errors that occur due to a bias of prejudice on the part of the informant or enumerator in selecting, estimating measuring instruments are called biased errors. Suppose for example, the enumerator used the deliberate sampling method in the place of simple random sampling method; then it is called biased errors. These errors are cumulative in nature and increase when the sample size also increases. These errors arise due to defect in the methods of collection of data, defect in the method of organization of data and defect in the method of analysis of data.

Unbiased errors: Errors which occur in the normal course of investigation or enumeration on account of chance are called unbiased errors. They may arise accidentally without any bias or prejudice. These errors occur due to faulty planning of statistical investigation.

To avoid these errors, the statistician must take proper precaution and care in using the correct measuring instrument. He must see that the enumerators are also not biased. Unbiased errors can be removed with the proper planning of statistical investigations. Both of these errors should be avoided by the statisticians.

Reducing Sampling Errors: Errors in sampling can be reduced, if the size of sample is increased. This is shown in the following diagram.



From the above diagram it is clear that when the size of the sample increases, sampling error decreases. And by this process samples can be made more representatives to the population.

Testing of Hypothesis: As a part of investigation, samples are drawn from the population and results are drawn which helps take the decision. But such decisions involve an element of uncertainty causing wrong decisions. Hypothesis is an assumption which may or may not be true about a population parameter. For example, if we toss a coin 200 times, we may get 110 heads and 90 tails. At this instance we are interested in testing whether the coin is unbiased or not.

Therefore, we may conduct a test to judge significance whether the difference is due to sampling of otherwise. To carry out a test of significance following procedure has to be followed:

1. Framing the Hypothesis: To verify the assumption, which is based on sample study, we collect data and find out the difference between the sample value and the population value. If there is no difference found or the difference is very small then the hypothetical value is correct. Generally two hypotheses are constructed, and if one is found correct the other is rejected.

(a) Null Hypothesis: The random selection of the samples from the given population makes the tests of significance valid for us. For applying any test of significance we first set up a hypothesis- a definite statement about the population parameter/s. Such a statistical hypothesis, which is under test, is usually a hypothesis of no difference and hence is called *Null hypothesis.* It is usually denoted by Ho. In the words of Prof. R.A.Fisher **"Null hypothesis is the hypothesis which**

is tested for possible rejection under the assumption that it is true."

(b) Alternative Hypothesis. Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is usually denoted by H_1 .It is very important to explicitly state the alternative hypothesis in respect of any null hypothesis H_0 because the acceptance or rejection of H_o is meaningful only if it is being tested against a rival hypothesis. For example, if we want to test the null hypothesis that the population has a specified mean **μ_0(say), i.e.,**

$$\mathbf{H_0: \mu = \mu}$$

Then the alternative hypothesis could be:

(i) $\mathbf{H_1: \mu \neq \mu_0}$ **(i.e.** $\mathbf{\mu > \mu_0}$ **or** $\mathbf{\mu < \mu_0}$**)**

(ii) $\mathbf{H_1: \mu > \mu_0}$ **(iii)** $\mathbf{H_1: \mu < \mu_0}$

The alternative hypothesis (i) is known as a two – tailed alternative and the alternatives in (ii) and (iii) are known as right – tailed and left- tailed alternatives. Accordingly, the corresponding tests of significance are called two-tailed, tight-tailed and left-tailed tests respectively.

The null hypothesis consists of only a single parameter value and is usually simple while alternative hypothesis is usually composite.

Types of Errors in Testing of Hypothesis: As stated earlier, the inductive inference consists in arriving at a decision to accept or reject a null hypothesis (Ho) after inspecting only a sample from it. As such an element of risk – the risk of taking wrong decision is involved. In any test procedure, the four possible mutually disjoint and exhaustive decisions are:

(i) Reject Ho when actually it is mot true, i.e., when Ho id false.

(ii) Accept Ho when it is true.

(iii) Reject Ho when it is true.

(iv) Accept Ho when it is false.

The decision in (i) and (ii) are correct decisions while the decisions (iii) and (iv) are wrong decisions. These decisions may be expressed in the following dichotomous table:

		Decision from sample	
True State		Reject Ho	Accept Ho
	Ho True	Wrong Type I Error	Correct
	Ho False (H_1 True)	Correct	Wrong Type II Error.

Thus, in testing of hypothesis we are likely to commit two types of errors. The error of rejecting Ho when Ho is true is known as Type I error and the error of accepting Ho when Ho is false is known as Type II Error.

For example, in the Industrial Quality Control, while inspecting the quality of a manufactured lot, the Inspector commits Type I Error when he rejects a good lot and he commits Type II Error when he accepts a bad lot.

SUMMARY

Nowadays questionnaire method of data collection has become very popular. It is a very powerful tool to collect required data in shortest period of time and with little expense. It is scientific too. But drafting of questionnaire is a very skilled and careful work. Therefore, there are certain requirements and essentials which should be followed at the time of framing the questionnaire. They include- size of the questionnaire should be small, questions should be very

clear in understanding, questions should be put in a logical order, questions should have simple meaning etc. Apart from this, multiple choice questions should be asked. Questionnaire should be pre tested before going for final data collection. Information supplied should be cross checked for any false or insufficient information. After all these formalities have been completed, a covering note should accompany the questionnaire explaining various purposes, designs, units and incentives.

There are two ways of survey- Census survey and Sample survey through which data can be collected. Census survey means total enumeration i.e. collecting data from each and every unit of the universe whereas sample survey concentrates on collecting data from few units of the universe selected scientifically for the purpose. Since census method is more time taking, expensive and labor intensive, it becomes impractical to depend on it. Therefore, sample survey is preferred which is scientific, less expensive, less time taking and less labor intensive too.

But there are merits and demerits of this method which are detailed below:

Merits- it reduces cost, it is more reliable, it saves time; it provides more detailed information, some times only method to depend upon, administrative convenience, more scientific etc.

Demerits- it may give illusory conclusions, sometimes samples may not be representative, there is lack of experts, some times it is more difficult than census method, personal bias, determining the size of the sample very difficult etc.

Apart from these, there are some essentials of sampling which must be followed. They are – Samples must be representative, samples must be homogeneous and the number of samples must be adequate. When the researcher resorts to sampling, he intends to collect some data which help him to

draw results and finally take a decision. When he takes a decision on the basis of hypothesis which is precisely assumption and is prone to two types of errors- Type I Error and Type II Error. When a researcher rejects a correct hypothesis, he commits type I error and when he accepts a wrong hypothesis he commits type II error. The researcher should try to avoid both types of errors but committing type II error is more harmful than type I error.

SELF ASSESMENT QUESTIONS (SEQs)

1. Explain questionnaire and examine its main characteristics.
   (Refer to the introduction part of the questionnaire section)
2. Explain main requirements of a good questionnaire.
   (Refer to the sub points from 1 to 11)
3. What is sampling? Explain its main merits and demerits.
   (Refer to the introduction and the following part of the lesson)
4 What are null and alternative hypothesis? Explain.
   (Refer the point Framing the Hypothesis)
6. What are Type I error and Type II error? (Refer to types of error in hypothesis)

***

UNIT II

LESSON 3

EXPERIMENTS

LESSON OUTLINE

- Procedures adopted in experiments
- Meaning of Experiments
- Research design in case of hypothesis testing research studies
- Basic principles in experimental designs
- Prominent experimental designs

LEARNING OBJECTIVES

- After reading this lesson you should be able to
- Nature and meaning of Experiments
- Kinds of experiments

Introduction

The meaning of experiment lies in the process of examining the truth of a statistical hypothesis relating to some research problem. For example, a researcher can conduct an experiment to examine the newly developed medicine. Experiment is of two types- absolute experiment and comparative experiment. When a researcher wants to determine the impact of a fertilizer on the yield of a crop it is a case of absolute experiment. On the other hand if he wants to determine the impact of one fertilizer as compared to the impact of some other fertilizer, the experiment will then be called as a comparative experiment. Normally a researcher conducts a comparative experiment when he talks of designs of experiments.

Research design can be of three types-

(a) Research design in case of descriptive and diagnostic research studies,

(b) Research design in case of exploratory research studies and,

(c) Research design in case of hypothesis testing research studies.

Here we are mainly concerned with the third one which is Research design in case of hypothesis testing research studies.

Research design in case of hypothesis testing research studies: Hypothesis testing research studies is generally known as experimental studies. This is a study where a researcher tests the hypothesis of causal relationships between variables. This type of study requires some procedures which will not only reduce bias and increase reliability, but will permit drawing inferences about causality. Most of the times, experiments meet these requirements. Prof. Fisher is considered as the pioneer of this type of studies (experimental studies). He did pioneering work when he was working at Rothamsted Experimental Station in England which was a centre for Agricultural Research. While working there Prof. Fisher found that by dividing plots into different blocks and then by

conducting experiments in each of these blocks , whatever information is collected and inferences drawn from them, happens to be more reliable. This was where he was inspired to develop certain experimental designs for testing hypotheses concerning scientific investigations. Nowadays the experimental design is used in researches relating to almost every disciplines of knowledge. Now let us see the basic principles of experimental designs which are discussed below:

Prof. Fisher has laid three principles of experimental designs:

(1) The Principle of Replication

(2) The Principle of Randomization and

(3) The Principle of Local Control.

(1) The Principle of Replication: According to this principle the experiment should be repeated more thon once. Thus, each treatment is applied in many experimental units instead of one. This way the statistical accuracy of the experiments is increased. For example, suppose we are going to examine the effect of two varieties of wheat. Accordingly we divide the field into two parts and grow one variety in one part and the other variety in the other. Then we compare the yield of the two parts and draw conclusion on that basis. But if we are to apply the principle of replication to this experiment, then we first divide the field into several parts, grow one variety in half of these parts and the other variety in the remaining parts. Then we collect the data of yield of the two varieties and draw conclusion by comparing the same. The result so obtained will be more reliable in comparison to the conclusion we draw without applying the principle of replication. The entire experiment can be repeated several times for the better results.

(2) The Principle of Randomization: When we conduct an experiment the principle of randomization provides us a protection against the effects of

extraneous factors by randomization. This means that, this principle indicates that the researcher should design or plan the experiment in such a way that the variations caused by extraneous factors can all be combined under the general heading of 'chance'. For example, when a researcher grow one variety of wheat , say , in the first half of the parts of a field and the other variety he grows in the other half, then it is just possible that the soil fertility may be different in the first half in comparison to the other half. If this is so the researcher's result is not realistic. In this situation, he may assign the variety of wheat to e grown in different parts of the field on the basis of some random sampling technique, i.e., he may apply randomization principle and protect himself against the effects of the extraneous factors. Therefore, by using the principle of randomization, he can draw a better estimate of the experimental error.

(3). The Principle of Local Control: This is another important principle of experimental designs. Under this principle, the extraneous factor, the known source of variability, is made to vary deliberately over as wide a range as necessary and this need to be done in such a way that the variability it causes can be measured and hence eliminated from the experimental error. The experiment should be planned in such a way that the researcher can perform a two-way analysis of variance , in which the total variahili8ty of the data is divided into three components attributed to treatments (varieties of wheat in this case) the extraneous factor(soil fertility in this case) and experimental error. In short, through the principle of local control we can eliminate the variability due to extraneous factors from the experimental error.

Kinds of experimental Designs and control

Experimental designs refer to the framework of structure of an experiment and as such there are several experimental designs. Generally experimental designs are classified into two broad categories: informal experimental designs and

formal experimental designs. Informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental designs offer relatively more control and use precise statistical procedures for analysis. Important experimental designs are discussed below:

(1) Informal experimental designs:

(i) Before and after without control design

(ii) After only with control design

(iii) Before and after with control design

(2) Formal experimental designs:

(i) Completely randomized design (generally called C.R design)

(ii) Randomized block design (generally called R.B design)

(iii) Latin square design (generally called L.S design)

(iv) Factorial designs.

(1)Before and after without control design: In this design a single test group or area is selected and the dependent variable is measured before the introduction of the treatment. Then the treatment is introduced and the dependent variable is measured again after the treatment has been introduced. The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level of the phenomenon before the treatment. Thus the design can be presented in the following manner:

Test area	Level of phenomenon Before treatment(X)	Treatment introduced ⟶	Level of phenomenon after treatment(Y)
	Treatment Effect =(Y)-(X)		

The main difficulty of such a design is that with the passage of time considerable extraneous variations may be there in its treatment effect.

(2) After-only with control design:- Two groups or areas are selected in this design and the treatment is introduced into the test area only. Then the dependent variable is measured in both the areas at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control area from its value in the test area. The design can be presented in the following manner:

Test area:	Treatment introduced	Level of phenomenon after Treatment (Y)
Control area:	⟶	Level of phenomenon Without treatment (Z)
	Treatment Effect = (Y)-(Z)	

The basic assumption in this type of design is that the two areas are identical with respect to their behavior towards the phenomenon considered. If this assumption is not true, there is the possibility of extraneous variation entering into the treatment effect.

(3) Before and after with control design:- In this design two areas are selected and the dependent variable is measured in both the areas for an identical time-period before the treatment. Thereafter, the treatment is introduced into the test area only, and the dependent variable id measured in both for and identical time –period after the introduction of the treatment. The effect of the treatment is determined by subtracting the change in the dependent variable in the control area from the change in the dependent variable in test area. This design can be shown in the following way:

	Time Period I		Time Period II
Test area:	Level of phenomenon	Treatment	Level of phenomenon
	Before treatment (X)	introduced	after treatment (Y)
Control area:	Level of phenomenon		Level of phenomenon
	Without treatment		without treatment
	(A)		(Z)

Treatment Effect = (Y-X)-(Z-A)

This design is superior to the previous two designs because it avoids extraneous variation resulting both from the passage of time and from non-comparability of the rest and control areas. But at times, due to lack of historical data time or a comparable control area, we should prefer to select one of the first two informal designs stated above.

(2) Formal Experimental Design

(i) Completely randomized design:- This design involves only two principles i.e., the principle of replication and the principle of randomization of experimental designs. Among all other designs this is the simpler and easier because it's procedure and analysis are simple. The important characteristic of this design is that the subjects are randomly assigned to experimental treatments. For example, if the researcher has 20 subjects and if he wishes to test 10 under treatment A and 10 under treatment B, the randomization process gives every possible group of 10 subjects selected from a set of 20 an equal opportunity of being assigned to treatment A and treatment B. One way analysis of variance (one way ANOVA) is used to analyze such a design.

(ii) Randomized block design:- R. B. design is an improvement over the C.R. design. In the R .B. design the principle of local control can be applied along with the other two principles of experimental designs. In the R.B. design, subjects are first divided into groups, known as blocks, such that within each group the subjects are relatively homogenous in respect to some selected variable. The number of subjects in a given block would be randomly assigned to each treatment. Blocks are the levels at which we hold the extraneous factor fixed, so that its contribution to the total variability of data can he measured. The main feature of the R.B. design is that in this each treatment appears the same number of times in each block. This design is analyzed by the two-way analysis of variance (two-way ANOVA) technique.

(3) Latin squares design:- The Latin squares design (L.S design) is an experimental design which very frequently used in agricultural research. Because agriculture to a large extent depends upon nature, therefore, the condition of research and investigation in agriculture is different than the other studies. For example, an experiment has to be made through which the effects of fertilizers on the yield of a certain crop, say wheat, is to he judged. In this situation, the varying fertility of the soil in different blocks in which the experiment has to be performed must be taken into consideration; otherwise the results obtained may not be very dependable because the output happens to be the effects of not only of fertilizers, but also be the effect of fertility of soil. Similarly there may be the impact of varying seeds of the yield. In order to overcome such difficulties, the L.S. design is used when there are two major extraneous factors such as the varying soil fertility and varying seeds. The Latin square design is such in which each fertilizer will appear five times but will be used only once in each row and in each column of the design. In other words, in this design, the treatment is so allocated among the plots that no treatment

occurs more than once in any one row or any one column. This experiment can be shown with the help or the following diagram:

FERTILITY LEVEL

	I	II	III	IV	V
X_1	A	B	C	D	E
X_2	B	C	D	E	A
X_3	C	D	E	A	B
X_4	D	E	A	B	C
X_5	E	A	B	C	D

From the above diagram it is clear that in L.S. design the field is divided into as many blocks as there are varieties of fertilizers and then each block is again divided into as many parts as there are varieties of fertilizers in such a way that each of the fertilizer variety is used in each of the block only once. The analysis of L.S. design is very similar to the two-way ANOVA technique.

4. Factorial design:- Factorial designs are used in experiments where the effects of varying more than one factor are to be determined. These designs are more used in economic and social matters where usually a large number of factors affect a particular problem. Factorial designs are usually of two types:

(i) Simple factorial designs and (ii) complex factorial design.

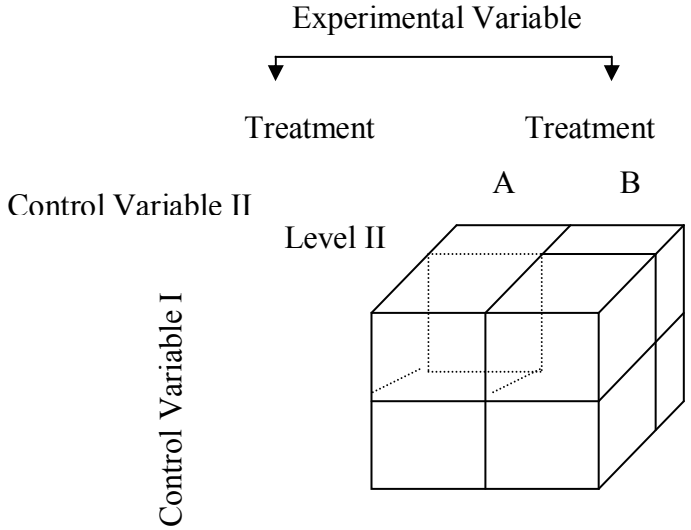
(i) Simple factorial design: In simple factorial design, the effects of varying two factors on the dependent variable is considered but when an experiment is done with more than two factors, complex factorial designs are used. Simple factorial design is also termed as a 'two-factor-factorial design,' whereas complex factorial design is known as 'multi-factor-factorial design.

(ii) Complex factorial designs:- When the experiments with more than two factors at a time are conducted, it involves the use of complex factorial designs. A design which considers three or more independent variables simultaneously is called a complex factorial design .In case of three factors with one experimental variable having two treatments and two levels, will be termed 2x2x2 complex factorial design which will contain a total of eight cells can be soon through the following diagram

2x2x2 COMPLEX FACTORIAL DESIGN

	Experimental Variable			
	Treatment A		Treatment B	
	Control Variable 2 Level I	Control Variable 2 Level II	Control Variable 2 Level I	Control Variable 2 Level II
Control Variable1 Level I	Cell 1	Cell 3	Cell 5	Cell 7
Control Variable1 Level II	Cell 2	Cell 4	Cell 6	Cell 8

A pictorial presentation is given of the design shown above in the following



Level I

Level I

Level II

The dotted line cell in this diagram corresponds to cell I of the above stated 2x2x2 design and is for treatment A, level I of the control variable 1, and level I of the control variable 2. From this design it is possible to determine the main effects for three variables i.e., one experimental and true control variables. The researcher can also determine the interaction between each possible pair of variables (such interactions are called 'first order interactions') and interaction between variable taken in triplets (such interactions are called second order interactions). In case of a 2x2x2 design, the further given first order interactions are possible:

Experimental variable with control variable 1 (or EV x CV 1);

Experimental variable with control variable 2 (or EV x CV 2);

Control variable 1 with control variable 2 (or CV 1 x CV 2);

There will be one second order interaction as well in the given design (it is between all the three variables i.e., EV x CV 1 x CV 2).

To determine the main effect for the experimental variable the researcher must necessarily compare the combined mean of data in cells 1, 2,3 and 4 for Treatment A with the combine mean of data in cells 5,6,7 and 8 for Treatment B. In this way the main effect experimental variable, independent of control variable 1 and variable 2, is obtained. Similarly, the main effect for control variable 1, independent experimental variable and control variable 2, is obtained if we compare the combined mean of data in cells 1, 3, 5 and 7 with the combined mean of data in cells 2, 4, 6 and 8 of our 2x2x2 factorial design. On

similar lines, one can determine the effect of control variable 2 independent of experimental variable and control variable 1, if the combined mean of data in cells 1,2,5 and 6 are compare with the combined mean of data in cells 3,4,7 and 8.

To obtain the first order interaction, say, for EV x CV 1 in the above stated design, the researcher must necessarily ignore control variable 2 for which purpose he may develop 2x2 design from the 2x2x2 design by combining the data of the relevant cells of the latter design as has been shown on next page:

		Experimental Variable	
		Treatment A	Treatment B
Control Variable 1	Level I	Cells 1.3	Cells 5,7
	Level II	Cells 2,4	Cells 6,8

Similarly, the researcher can determine other first order interactions. The analysis of the first order interaction in the manner described above, is essentially a simple factorial analysis as only two variables are considered at a time and the remaining one is ignored. But the analysis of the second order interaction would not ignore one of the three independent variables in case of a 2x2x2 design. The analysis would be termed as a complex factorial analysis.

It may, however, be remembered that the complex factorial design need not necessarily be of 2x2x2 type design, but can be generalized to any number and combination of experimental and control independent variables. Of course, the greater the number of independent variables included in a complex factorial design, the higher the order of the interaction analysis possible. But the overall task goes on becoming more and more complicated with the inclusion of more and more independent variables in our design.

Factorial designs are used mainly because of the two advantages. (i) They provide equivalent accuracy (as happens in the case of experiments with only one factor) with less labour and as such are source of economy. Using factorial designs, we can determine the effects of two (in simple factorial design) or more (in case of complex factorial design) factors (or variables) in one single experiment. (ii) They permit various other comparisons of interest. For example, they give information about such effects which cannot be obtained by treating one single factor at a time. The determination of interaction effects is possible in case of factorial designs.

Conclusion

There are several research designs and the researcher must decide in advance of collection and analysis of data as two which design would true to be more appropriate for his research project. He must give due weight to various points such as type universe and it's nature, the objective of the study, the source list or the sampling frame, desired standard accuracy and the like when taking a decision in respect of the design for his research project.

SUMMARY

Experiment is the process of examining the truth of a statistical hypothesis relating to some research problem. There are two types of experiment- absolute and comparative. There are three types of research designs- research design for descriptive and diagnostic research, research design for exploratory research studies and research design for hypothesis testing. Prof. Fisher has laid three principles of experimental design. They are—Principle of Replication, Principle of Randomization and Principle of Local control. There are different kinds of experimental design. Some of them are –Informal experimental design, After only with control design, Formal experimental design, Completely randomized design, Randomized block design, Latin square design and Factorial design.

SELF ASSESMENT QUESTIONS (SEQs)

1. Explain the meaning and types of experiment.
   (Ref. introduction and types of research design next to introduction)
2. Explain informal designs.
   (Ref. i,ii,iii in informal experiment design portion.)
3. Explain formal experimental design and control.
   (Ref. i,ii,iii,iv in formal experiment design section.)
4. Explain complex factorial design.

***

UNIT II

LESSON **4**

OBSERVATION

LESSON OUTLINE:

- **Meaning and Characteristics of observation**
- **Types of observation**
- **Stages of observation**
- **Steps in observation**
- **Problems and**
- **Merits and Demerits**

After reading this lesson you will be able to know

- **Meaning and types of observation**
- **Stages through which observation**

Passes

- **Steps followed and the problems coming in observation**
- **Merits and Demerits**

Introduction

Observation is a method that employs vision as its main means of data collection. It implies the use of eyes rather than of ears and the voice. It is accurate watching and noting of phenomena as they occur with regard to the cause and effect or mutual relations. It is watching other persons' behavior as it actually happens without controlling it. For example, watching bonded laborer's life, or treatment of widows and their drudgery at home, provide graphic description of their social life and sufferings. Observation is also defined as "a planned methodical watching that involves constraints to improve accuracy".

CHARACTERISTICS OF OBSERVATION

Scientific observation differs from other methods of data collection specifically in four ways: (i) observation is always direct while other methods could be direct or indirect; (ii) field observation takes place in a natural setting; (iii)

observation tend to be less structured; and (iv) it makes only the qualitative(and not the quantitative) study which aims at discovering subjects' experiences and how subjects make sense of them(phenomenology) or how subjects understand their life(interpretivism).

Lofland(1955:101-113) has said that this method is more appropriate for studying lifestyles or sub-culture, practices, episodes, encounters, relationships, groups, organizations, settlements and roles, etc. Black and Champion (1976:330) have given the following characteristics of observation:

- Behavior is observed in natural surroundings.
- It enables understanding significant events affecting social relations of the participants.
- It determines reality from the perspective of observed person himself.
- It identifies regularities and recurrences in social life by comparing data in our study with those in other studies.
  Besides, four other characteristics are.
- Observation involves some controls pertaining to the observation and to the means he uses to record data. However, such controls do not exist for the setting or the subject population.
- It is focused on hypotheses-free inquiry.
- It avoids manipulations in the independent variable i.e., one that is supposed to cause other variable(s) and is not caused by them.
- Recording is not selective.

Since, at times, observation technique is indistinguishable from experiment technique, it is necessary to distinguish the two. O*ne*, that observation involves few controls than the experiment technique. *Two*, the behavior observed in observation is natural whereas in experiment it is not always so. *Three*, behavior observed in experiment in more molecular (of a smaller unit) while one in observation is molar. *Four,* in observation, fewer subjects are watched for long periods of time in more varied circumstances than in experiment. *Five,* training required in observation study is directed more

towards sensitizing the observer to the flow of events whereas training in experiments serves to sharpen the judgment of the subject, *Lastly,* in observational study, the behavior observed is more diffused. Observational methods differ from one another along several variables or dimensions.

***

UNIT – III

STATISTICAL ANALYSIS

CONTENTS

1. Probability
2. Probability distribution
   - 2.1 Binomial distribution
   - 2.2 Poisson distribution
   - 2.3 Normal distribution
3. Testing of Hypothesis
   - 3.1 Small sample
   - 3.2 Large sample test
4. χ^2 test
5. Index Number
6. Analysis of Time Series

OBJECTIVES:

The objectives of the present chapter are:

i) To examine the utility of various statistical tool in decision making.

ii) To inquire about the testing of a hypothesis

1. PROBABILITY

If an experiment is repeated under essentially homogeneous and similar conditions, we will arrive at two types of conclusions. They are: - the results are unique and the outcome can be predictable and result is not unique but may be one of the several possible outcomes. In this context, it is better to understand various terms pertaining to probability before examining the probability theory. The main terms are explained as follows:-

(i) Random experiment:

An experiment which can be repeated under the same conditions and the outcome cannot be predicted under any circumstances is known as random experiment. For example: An unbiased coin is tossed. Here we are not in a position to predict head or tail is going to occur. Hence, this type of experiment is known as random experiment.

(ii) Sample Space

A set of possible outcomes of a random experiment is known as sample space. For example in the case of tossing an unbiased coin twice, the possible outcomes are HH, HT, TH and TT. This can be represented in a sample space as S= (HH, HT, TH, TT).

(iii) An event

Any possible outcomes of an experiment are known as an event. In the case of tossing of an unbiased coin twice, HH is an event. An even can be classified into two. They are: (a) Simple events, and (ii) compound event. Simple event is an event which has only one sample point in the sample space. Compound event is an event which has more than one sample point in the sample space. In the case of tossing of an unbiased coin twice HH is a simple event and TH and TT are the compound events.

(iv) Complementary event

A and A' are the complementary event if A' consists of all those sample point which is not included in A. For instance, an unbiased dice is thrown once. The probability of an odd number turns up are complementary to an even number turns up. Here, it is worth mentioning that the probability of sample space is always is equal to one. Hence, the P (A') = 1 - P (A).

(v) Mutually exclusive events

A and B are the two mutually exclusive events if the occurrence of A precludes the occurrence of B. For example, in the case of tossing of an unbiased coin once, the occurrence of head precludes the occurrence of tail. Hence, head and tail are the mutually exclusive event in the case of tossing of an unbiased coin once. If A and B are mutually exclusive events, then the probability of occurrence of A or B is equal to sum of their individual probabilities. Symbolically, it can be presented as:-

$$P (A \cup B) = P (A) + P (B)$$

If A and B is joint sets, then the addition theorem of probability can be stated as:

$$P (A \cup B) = P(A) + P(B) - P(AB)$$

(vi) Independent event

A and B are the two independent event if the occurrence of A does not influence the occurrence of B. In the case of tossing of an unbiased coin twice, the occurrence of head in the first toss does not influence the occurrence of head or tail in the toss. Hence, these two events are called independent events. In the case of independent event, the multiplication theorem can be stated as the probability of A and B is the product of their individual probabilities. Symbolically, it can be presented as:-

P (A B) = P (A) * P (B)

Addition theorem of Probability

Let A and B are the two mutually exclusive events then the probability of A or B is equal to sum of their individual probabilities. (For detail refer mutually exclusive events)

Multiplication theorem of Probability

Let A and B are the two independent events, then the probability of A and B is equal to the product of their individual probabilities. (For details refer independent events)

Example: The odds that person X speaks the truth are 4:1 and the odds that Y speaks the truth are 3:1. Find the probability that:-

(i) both of them speak the truth,
(ii) any one of them speak the truth, and
(iii) truth may not be told.

Solution: The probability of X speaks the truth = 1/5

The probability that X speaks lie = 4/5
The probability that Y speaks the truth = 1/4
The probability that Y speaks lie = ¼

(i) Both of them speak truth = P(X) * P(Y) = 1/5 * 1/4 = 1/20
(independent event)

(ii) any one of them speak truth = P(X) + P(Y) - P(X*Y)
= 1/5 + 1/4 - 1/5*1/4 = 8/20 = 2/5 (not mutually exclusive events)

(iii) Truth may not be told
= 1 – P(any one of them speak truth)(complementary event)
= 1 – 2/5 = 3/5.

2. PROBABILITY DISTRIBUTION

Let X is discrete random variable which takes the values of x_1, x_2,x_3….. x_n and the corresponding probabilities will be p_1, p_2, …….p_n. Then, X follows the

probability distribution. The two main properties of probability distribution are :- (i) P(Xi) is always greater than or equal to zero and less than or equal to one, and (ii) the summation of probability distribution is always equal to one. For example, tossing of an unbiased coin twice.

Then the probability distribution is:

X (probability of obtaining head): 0 1 2
P(Xi) : ¼ ½ ¼

Expectation of probability

Let X is discrete random variable which takes the value of $x_1, x_2, \ldots\ldots x_n$ then the respective probability is $p_1, p_2, \ldots\ldots\ldots p_n$, Then the expectation of probability distribution is $p_1x_1 + p_2x_2 + \ldots\ldots\ldots\ldots + p_nx_n$. In the above example, the expectation of probability distribution is (0* ¼ +1*1/2+2*¼)=1.

2.1 BINOMIAL DISTRIBUTION

The binomial distribution also known as 'Bernoulli Distribution' and it is associated with the name of a Swiss mathematician James Bernoulli also known as Jacques or Jakon (1654 – 1705). Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives. It can be explained as follows:

(i) Let an experiment is repeated under the same conditions for a fixed number of trials, say, n.

(ii) In each trial, there are only two possible outcomes of the experiment. Let us define it as a "success" or "failure". Then the sample space of possible outcomes of an each experiment is:-

S = [failure, success]

(iii) The probability of a success denoted by p remains constant from trial to trial and the probability of a failure denoted by q which is equal to (1 – p).

(iv) The trials are independent in nature i.e., the outcomes of any trial or sequence of trials do not affect the outcomes of subsequent trials. Hence, the Multiplication theorem of probability can be applied for the occurrence of success and failure. Thus, the probability of success or failure is p.q.

(v) Let us assume that we conduct an experiment in, n times. Out of which x times be the success and failure is (n-x) times. The occurrence of success or failure in successive trials is mutually exclusive events. Hence, we can apply addition theorem of probability.

(v) Based on the above two theorem the probability of a success or failure is

$$P(X) = {}^{n}C_{x}p^{x}q^{n-x}$$

$$\frac{n!}{x!\,(n-x)!} \cdot p^{x}\,q^{n-x}$$

Where, P = Probability of success in a single trail, q = 1 – p, n = Number of trials and x = no. of successes in n trials.

Thus for an event A with probability of occurrence p and non-occurrence q, if n trials are made probability distribution of the number of occurrences of A will be as set. If we want to obtain the probable frequencies of the various outcomes in N sets of n trials, the following expression shall be used: $N(p + q)^n$

$$N(p + q)^n = Np^n + {}^{n}C_1p^{n-1}q + {}^{n}C_2p^{n-2}q^2 + \ldots\ldots + {}^{n}C_rp^{n-r}q^r + \ldots\ldots q^n.$$

The frequencies obtained by the above expansion are known as expected or theoretical frequencies. On the other hand, the frequencies actually obtained by making experiments are called actual or observed frequencies. Generally, there is some difference between the observed and expected frequencies but the difference becomes smaller and smaller as N increases.

Obtaining Coefficient of the Binomial Distribution:

The following rules may be considered for obtaining coefficients from the binomial expansion.

(i) The first term is $q^{n.}$,
(ii) The second term is ${}^{n}C_1 q^{n-1} p$,
(iii) In each succeeding term the power of q is reduced by 1 and the power of p is increased by 1.
(iv) The coefficient of any term is found by multiplying the coefficient of the preceding term by the power of q in that preceding term, and dividing the products so obtained by one more than the power of p in that proceeding term.
Thus, when we expand $(q + p)^n$, we will obtain the following:-

$$(p + q)^n = p^n + {}^{n}C_1 p^{n-1} q + {}^{n}C_2 p^{n-2} q^2 + \ldots\ldots + {}^{n}C_r p^{n-r} q^r + \ldots\ldots q^n.$$

Where, 1, ${}^{n}C_1$, ${}^{n}C_2$ ……. are called the binomial coefficient. Thus in the expansion of $(p + q)^4$ we will have $(p + q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4p^1q^3 + q^4$ and the coefficients will be 1, 4, 6, 4, 1.

From the above binomial expansion, the following general relationships should be noted:

(i) The number of terms in a binomial expansion is always n + 1,
(ii) The exponents of p and q, for any single term, when added together, always sum to n.
(iii) The exponents of p are n, (n – 1), (n – 2),…….1, 0, respectively and the exponents of q are 0, 1, 2, ……(n – 1), n, respectively.
(iv) The coefficients for the n + 1 terms of the distribution are always symmetrical in nature.

Properties of Binomial Distribution

The main properties of Binomial Distribution are:-

(i) The shape and location of binomial distribution changes as p changes for a given n or as n changes for a given p. As p increase for a fixed n, the binomial distribution shifts to the right.

(ii) The mode of the binomial distribution is equal to the value of x which has the largest probability. The mean and mode are equal if np is an integer.

(iii) As n increases for a fixed p, the binomial distribution moves to the right, flattens, and spreads out.

(iv) The mean of the binomial distribution, np and it increases as n increases with p held constant. For larger n there are more possible outcomes of a binomial experiment and the probability associated with any particular outcome becomes smaller.

(v) If n is larger and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by $z = (X - np) / \sqrt{npq}$.

(vi) The various constants of binomial distribution are:

Mean	=	np
Standard Deviation	=	$\sqrt{npq}$
μ_1	=	0
μ_2	=	npq
μ_3	=	npq(q – p)
μ_4	=	$3n^2p^2q^2 + npq(1 - 6pq)$.
Skewness	=	$\frac{(q-p)^2}{npq}$
Kurtosis	=	$3 + \frac{1-6pq}{npq}$

Illustrations: A coin is tossed four times. What is the probability of obtaining two or more heads?

Solution: When a coin is tossed the probabilities of head and tail in case of an unbiased coin are equal, i.e., p = q = ½

They various possibilities for all the events are the terms of the expansion $(q+p)^4$

$$(p - q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4p^1q^3 + q^4$$

Therefore, the probability of obtaining 2 heads is

$6p^2q^2 = 6 \times (½)^2(½)^2 = 3/8$

The probability of obtaining 3 heads is $6p^3q^1 = 4 \times (½)^3(½)^1 = 1/4$

The probability of obtaining 4 heads is $(q)^4 = (½)^4 = 1/16$

Therefore, the probability of obtaining 2 or more heads is

$$\frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{11}{16}$$

Illustration: Assuming that half the population is vegetarian so that the chance of an individual being a vegetarian is ½ and assuming that 100 investigations can take sample of 10 individuals to verify whether they are vegetarians, how many investigation would you expect to report that three people or less were vegetarians?

Solution:

n = 10, p, i.e., probability of an individual being vegetarian = ½.q =1 – p= ½

Using binomial distribution, we have $P(r) = {}^nc_r\, q^{n-r}p^r$

Putting the various values, we have

$$10c_r(½)^r(½)^{10-r} = 10cr = (½)^{10} = \frac{1}{1024}{}^{10}c_r$$

The probability that in a sample of 10, three or less people are vegetarian shall be given by: P(0) + P(1) + P(2) + P(3)

$$= \frac{1}{1024}[{}^{10}c_0 + {}^{10}c_1 + {}^{10}c_2 + {}^{10}c_3]$$

$$= \frac{1}{1024}[1 + 10 + 45 + 120] = \frac{176}{1024} = \frac{11}{64}$$

Hence out of 1000 investigators, the number of investigators who will

11

report 3 or less vegetarians in a sample of 10 is 1000 x $\frac{}{64}$ = 172.

2.2 POISSON DISTRIBUTION

Poisson distribution was derived in 1837 by a French mathematician Simeon D Poisson (1731 – 1840). In binomial distribution, the values of p and q and n are given. There is a certainty of the total number of events. But there are cases where p is very small and n is very large, such case is normally related to Poisson distribution. For example, Persons killed in road accidents, the number of defective articles produced by a quality machine. Poisson distribution may be obtained as a limiting case of binomial probability distribution, under the following condition.

(i) p, successes, approach zero ($p \rightarrow 0$)
(ii) $np = m$ is finite.

The Poisson distribution of the probabilities of occurrence of various rare events (successes) 0,1,2,…. Given below:

Number of success (X)	Probabilities p(X)
0	e^{-m}
1	me^{-m}
2	$\frac{m^2e^{-m}}{2!}$
r	$\frac{m^re^{-m}}{r!}$
n	$\frac{m^ne^{-m}}{n!}$

Where, e = 2.718, and m = average number of occurrence of given distribution.

The Poisson distribution is a discrete distribution with a parameter m. the various constants are:

(i) Mean = $m = p$
(ii) Standard Deviation = $\sqrt{m}$
(iii) Skewness β_1 = $1/m$
(iv) Kurtosis, β_2 = $3 + 1/m$
(v) Variance = m

Illustration: A book contains 100 misprints distributed randomly throughout its 100 pages. What is the probability that a page observed at random contains at least two misprints. Assume Poisson Distribution.

Solution:

$$m = \frac{\text{Total Number of misprints}}{\text{Total number of page}} = \frac{100}{100} = 1$$

Probability that a page contains at least two misprints:

$$p(r \geq 2) = 1 - [p(0) + p(1)]$$

$$p(r) = \frac{m^r e^{-m}}{r!}$$

$$p(0) = \frac{1^0 e^{-1}}{0!} = e^{-1} = \frac{1}{e} = \frac{1}{2.7183}$$

$$p(1) = \frac{1^1 e^{-1}}{1!} = e^{-1} = \frac{1}{e} = \frac{1}{2.7183}$$

$$p(0) + p(1) = \frac{1}{2.718} + \frac{1}{2.718} = 0.736$$

$$P(r \geq 2) = 1 - [p(0) + p(1)] = 1 - 0.736 = \mathbf{0.264}$$

Illustration: If the mean of a Poisson distribution is 16, find (1) S.D.(2) B_1 (3) B_2 (4) μ_3 (5) μ_4

Solution: m = 16

1. S.D. = $\sqrt{m} = \sqrt{16} = 4$
2. $\beta_1 = 1/m = 1/16 = 0.625$
3. $\beta_2 = 3 + 1/m = 3 + 0.625 = 3.0625$
4. $\mu_3 = m = 16$
5. $\mu_4 = m + 3m^2 = 16 + 3(16)^2 = 784$

2.3 NORMAL DISTRIBUTION

The normal distribution was first described by Abraham Demoivre (1667-1754) as the limiting form of binomial model in 1733. Normal distribution was rediscovered by Gauss in 1809 and by Laplace in 1812. Both Gauss and Laplace were led to the distribution by their work on the theory of errors of observations arising in physical measuring processes particularly in astronomy.

The probability function of a Normal Distribution is defined as:

$$P(X) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$

Where, X = Values of the continuous random variable, μ = Mean of the normal random variable, e = 2.7183, Π = 3.1416

Relation between Binomial, Poisson and Normal Distributions

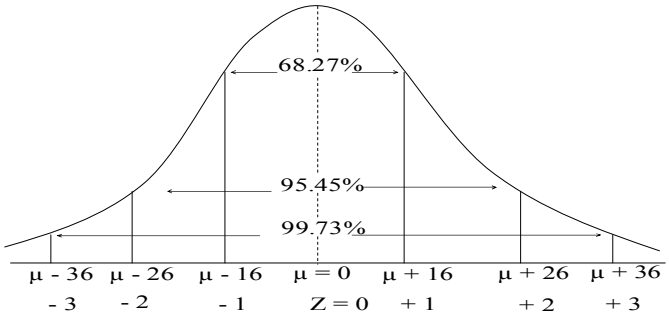
Binomial, Poisson and Normal distribution are closely related to each other. When N is large while the probability P of the occurrence of an event is close to zero so that q = (1-p) the binomial distribution is very closely approximated by the Poisson distribution with m = np.

The Poisson distribution approaches a normal distribution with statdardised variable $(x - m)/ \sqrt{m}$ as m increases to infinity.

Normal Distribution and its properties

The important properties of the normal distribution are:-

1. The normal curve is "bell shaped" and symmetrical in nature. The distribution of the frequencies on either side of the maximum ordinate of the curve is similar with each other.
2. The maximum ordinate of the normal curve is at $x = \mu$. Hence the mean, median and mode of the normal distribution coincide.
3. It ranges between $-\infty$ to $+\infty$
4. The value of the maximum ordinate is $1/\sigma\sqrt{2\Pi}$.
5. The points where the curve change from convex to concave or vice versa is at $X = \mu \pm \sigma$.
6. The first and third quartiles are equidistant from median.
7. The area under the normal curve distribution are:
   a) $\mu \pm 1\sigma$ covers 68.27% area;
   b) $\mu \pm 2\sigma$ covers 95.45% area.
   c) $\mu \pm 3\sigma$ covers 99.73% area.



8. When $\mu = 0$ and $\sigma = 1$, then the normal distribution will be a standard normal curve. The probability function of standard normal curve is

$$P(X) = \frac{1}{\sqrt{2\Pi}} e^{-x^2/2}$$

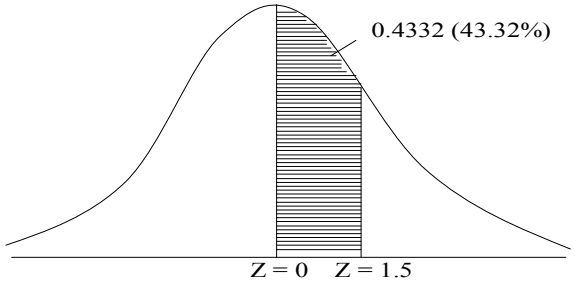
The following table gives the area under the normal probability curve for some important value of Z.

Distance from the mean ordinate in	Area under the curve

Terms of ± σ	
Z = ± 0.6745	0.50
Z = ± 1.0	0.6826
Z = ± 1.96	0.95
Z = ± 2.00	0.9544
Z = ± 2.58	0.99
Z = ± 3.0	0.9973

9. All odd moments are equal to zero.
10. Skewness = 0 and Kurtosis = 3 in normal distribution.

Illustration: Find the probability that the standard normal value lies between 0 to 1.5



As the mean, Z = 0.
To find the area between Z = 0 and Z = 1.5, look the area between 0 to 1.5, from the table. It is 0.4332 (shaded area)

Illustration: The results of a particular examination are given below in a summary form:

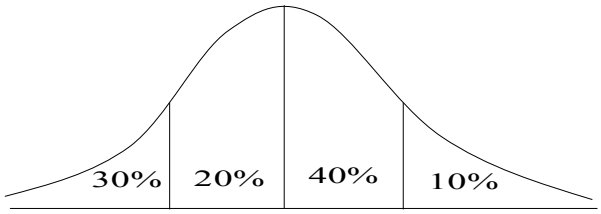
Result	Percentage of candidates
Passed with distinction	10
Passed	60
Failed	30

It is known that a candidate gets plucked if he obtains less than 40 marks, out of 100 while he must obtain at least 75 marks in order to pass with distinction. Determine the mean and standard deviation of the distribution of marks assuming this to be normal.

Solution:

30% students get marks less than 40.

$$Z = \frac{40 - \overline{X}}{\sigma} = -0.52 \text{ (from the table)}$$



$40 - \overline{X} = -0.52\sigma$ ---------- (i)

10% students get more than 75

40% area $= 75 - \overline{X} = 1.28$ ---------- (ii)

$= 75 - \overline{X} = 1.28\sigma$

Subtract (ii) from (i)

$40 - \overline{X} = -0.52\ \sigma$

$75 - \overline{X} = 1.28\ \sigma$

$-35 = -1.8\ \sigma$

$35 = 1.8\ \sigma$

$1.80\ \sigma = 35$

$$\sigma = \frac{35}{1.80} = 19.4$$

Mean $40 - \overline{X} = -0.52 \times (19.4)$

$-\overline{X} = -40 - 10.09 = 50.09$

Illustration: The scores made by candidate in a certain test are normally distributed with mean 1000 and standard deviation 200. what per cent of candidates receive scores (i) less than 800, (ii) between 800 and 1200? (the area under the curve between Z = 0 and Z = 1 is 0.34134).

Solution:

$\overline{X} = 1000$; $\sigma = 200$

$$Z = \frac{X - \overline{X}}{\sigma}$$

(i) For X = 800

$$Z = \frac{800 - 1000}{200} = -1$$

Area between Z = -1 and Z = 0 is 0.34134

Area for Z = -1 = 0.5 – 0.34134 = 0.15866

Therefore, the percentage = 0.15866 x 100 = 15.86%

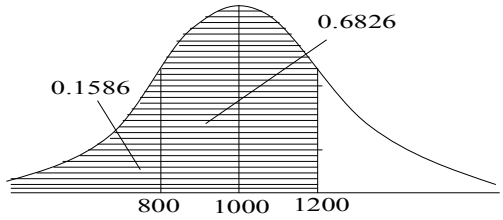
(ii) When, X = 1200,

$$Z = \frac{1200 - 1000}{200} = 1$$

Area between Z = 0 and Z = 1 is 0.34134

Area between X = 400 to X = 600

i.e., Z = -1 and Z = 1 is 0.34134 + 0.34134 = 0.6826 = 68.26%



3. TESTING OF HYPOTHESIS

3.1 Test of Significance for Large Samples

The test of significance for the large samples can be explained by these following assumptions:

(i) The random sampling distribution of statistics is approximately normal.

(ii) Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of estimate.

1. **The standard error of mean.**

In the case of large samples, when we are testing the significance of statistic, the concept of standard error is used. It measures only sampling errors. Sampling errors are involved in estimating a population parameter from a sample, instead of including all the essential information in the population.

(i) when standard deviation of the population is known, the formula is

$$S.E.\ \bar{X} = \frac{\sigma_p}{\sqrt{n}}$$

Where,
$S.E.\bar{X}$ = The standard error of the mean, σ_p = Standard deviation of the population, and n = Number of observations in the sample.

(ii) When standard deviation of population is not known, we have to use the standard deviation of the sample in calculating standard error of mean. The formula is

$$S.E.\ \bar{X} = \frac{\sigma\ (Sample)}{\sqrt{n}}$$

Where, σ = standard deviation of the sample, and n = sample size

Illustration: A sample of 100 students from Pondicherry University was taken and their average was found to be 116 lbs with a standard deviation of 20 lbs. Could the mean weight of students in the population be 125 pounds?

Solution:

Let us take the hypothesis that there is no significance difference between the sample mean and the hypothetical population mean.

$$\text{S.E.}\ \overline{X} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

$$\frac{\text{Difference}}{\text{S.E.}\overline{X}} = \frac{125 - 116}{2} = \frac{9}{2} = 4.5$$

Since, the difference is more than 2.58 S.E.(1% level) it could not have arisen due to fluctuations of sampling. Hence the mean weight of students in the population could not be 125 lbs.

3.2 Test of Significance for Small Samples

If the sample size is less than 30, then those samples may be regarded as small samples. As a rule, the methods and the theory of large samples are not applicable to the small samples. The small samples are used in testing a given hypothesis, to find out the observed values, which could have arisen by sampling fluctuations from some values given in advance. In a small sample, the investigator's estimate will vary widely from sample to sample. An inference drawn from a smaller sample result is less precise than the inference drawn from a large sample result.

t-distribution will be employed, when the sample size is 30 or less and the population standard deviation is unknown.

The formula is

$$t = \frac{(\overline{X} - \mu)}{\sigma} \times \sqrt{n}$$

Where, $\sigma = \sqrt{\Sigma(X - X)^2/n - 1}$

Illustration: the following results are obtained from a sample of 20 boxes of mangoes:
Mean weight of contents = 490gms,

Standard deviation of the weight = 9 gms.
Could the sample come from a population having a mean of 500 gms.

Solution:

Let us take the hypothesis that $\mu = 510$ gms.

$$t = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$$

$\bar{X} = 500$; $\mu = 510$; $\sigma = 10$; $n = 20$.

$$t = \frac{500 - 510}{10} \times \sqrt{20}$$

Df = 20 – 1 = 19 = (10/9) $\sqrt{20}$ = (10/9) x 4.47 = 44.7/9 = 4.96
Df = 19, $t_{0.01} = 3.25$

The computed value is less than the table value. Hence, our null hypothesis is accepted.

4. CHI-SQUARE TEST

F, t and Z tests were based on the assumption that the samples were drawn from normally distributed populations. The testing procedure requires assumption about the type of population or parameters, and these tests are known as 'parametric tests'.

There are many situations in which it is not possible to make any rigid assumption about the distribution of the population from which samples are being drawn. This limitation has led to the development of a group of alternative techniques known as non-parametric tests. Chi-square test of independence and goodness of fit is a prominent example of the use of non-parametric tests.

Though non-parametric theory developed as early as the middle of the nineteenth century, it was only after 1945 that non-parametric test came to be

used widely in sociological and psychological research. The main reasons for the increasing use of non-parametric tests in business research are:-

(i) These statistical tests are distribution-free

(ii) They are usually computationally easier to handle and understand than parametric tests; and

(iii) They can be used with type of measurements that prohibit the use of parametric tests.

The χ^2 test is one of the simplest and most widely used non-parametric tests in statistical work. It is defined as:

$$\chi^2 = \frac{\sum(O - E)^2}{E}$$

Where O = the observed frequencies, and E = the expected frequencies.

Steps: The steps required to determine the value of χ^2 are:

(i) Calculate the expected frequencies. In general the expected frequency for any cell can be calculated from the following equation:

$$E = \frac{R \times C}{N}$$

Where, E = Expected frequency, R = row's total of the respective cell, C = column's total of the respective cell and N = the total number of observations.

(ii) Take the difference between observed and expected frequencies and obtain the squares of these differences. Symbolically, it can be represented as $(O - E)^2$

(iii) Divide the values of $(O - E)^2$ obtained in step (ii) by the respective expected frequency and obtain the total, which can be symbolically represented by $\sum[(O - E)^2/E]$. This gives the value of χ^2 which can range from zero to infinity. If χ^2 is zero it means that the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater shall be the value of χ^2.

The computed value of χ^2 is compared with the table value of χ^2 for given degrees of freedom at a certain specified level of significance. If at the

stated level, the calculated value of χ^2 is less than the table value, the difference between theory and observation is not considered as significant.

The following observation may be made with regard to the χ^2 distribution:-

(i) The sum of the observed and expected frequencies is always zero.
Symbolically, $\sum(O - E) = \sum O - \sum E = N - N = 0$

(ii) The χ^2 test depends only on the set of observed and expected frequencies and on degrees of freedom v. It is a non-parametric test.

(iii) χ^2 distribution is a limiting approximation of the multinomial distribution.

(iv) Even though χ^2 distribution is essentially a continuous distribution it can be applied to discrete random variables whose frequencies can be counted and tabulated with or without grouping.

The Chi-Square Distribution

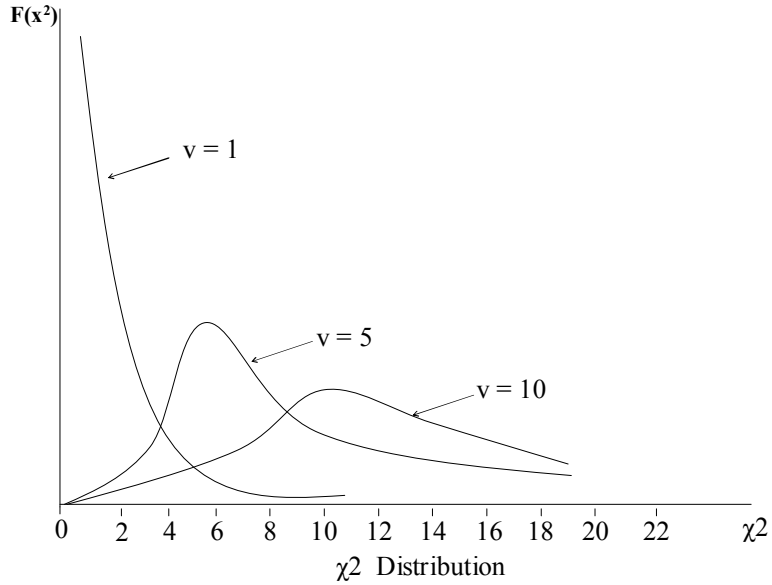
For large sample sizes, the sampling distribution of χ^2 can be closely approximated by a continuous curve known as the Chi-square distribution. The probability function of χ^2 distribution is:

$$F(\chi^2) = C\,(\chi^2)^{(v/2-1)}{}_e - x^2/2$$

Where e = 2.71828, v = number of degrees of freedom, C = a constant depending only on v.

The χ^2 distribution has only one parameter, v, the number of degrees of freedom. As in case of t-distribution there is a distribution for each different number of degrees of freedom. For very small number of degrees of freedom, the Chi-square distribution is severely skewed to the right. As the number of degrees of freedom increases, the curve rapidly becomes more symmetrical. For large values of v the Chi-square distribution, it is closely approximated by the normal curve.

The following diagram gives χ^2 distribution for 1, 5 and 10 degrees of freedom:



χ2 Distribution

It is clear from the given diagram that as the degrees of freedom increase, the curve becomes more and more symmetric. The Chi-square distribution is a probability distribution and the total area under the curve in each chi-square distribution is unity.

Properties of χ^2 distribution

The main Properties **of** χ^2 distribution are:-

(i) the mean of the χ^2 distribution is equal to the number of degrees of freedom, i.e., $\bar{X} = v$

(ii) the variance of the χ^2 distribution is twice the degrees of freedom, Variance = 2v

(iii) $\mu_1 = 0$,

(iv) $\mu_2 = 2v$,

(v) $\mu_3 = 8v$,

(vi) $\mu_4 = 48v + 12v^2$.

(vii) $\beta_1 = \frac{\mu_3^2}{\mu_2^2} = \frac{64v^2}{8v^3} = \frac{8}{v}$

$$(v) \qquad \beta_1\mu_3 = \frac{\mu_4}{\mu_2^2} = \frac{48v + 12v^2}{4v^2} = 3 + \frac{12}{v}$$

The table values of χ^2 are available only up to 30 degrees of freedom. For degrees of freedom greater than 30, the distribution of $\sqrt{2\chi^2}$ approximates the normal distribution. For degrees of freedom greater than 30, the approximation is acceptable close. The mean of the distribution $\sqrt{2\chi^2}$ is $\sqrt{2v} - 1$, and the standard deviation is equal to 1. Thus the application of the test is simple, for deviation of $\sqrt{2\chi^2}$ from $\sqrt{2v} - 1$ may be interpreted as a normal deviate with units standard deviation. That is,

$$Z = \sqrt{2\chi^2} - \sqrt{2v} - 1$$

Alternative Method of Obtaining the Value of χ^2

In a 2x2 table where the cell frequencies and marginal totals are as below:

a	b	(a+b)
c	d	(c+d)
(a+c)	(b+d)	N

N is the total frequency and ad the larger cross-product, the value of χ^2 can easily be obtained by the following formula:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(c + d)(a + b)} \text{ or}$$

With Yate's corrections

$$\chi^2 = \frac{N(ab - bc - \frac{1}{2}N)^2}{(a + c)(b + d)(c + d)(a + b)}$$

Conditions for applying χ^2 test:

The main conditions considered for employing the χ^2 test are:

(i) N must be to ensure the similarity between theoretically correct distribution and our sampling distribution of χ^2.

(ii) No theoretical cell frequency should be small when the expected frequencies are too small. If it is so, then the value of χ^2 will be overestimated and will result in too many rejections of the null hypothesis. To avoid making incorrect inferences, a general rule is followed that expected frequency of less than 5 in one cell of a contingency table is too small to use. When the table contains more than one cell with an expected frequency of less than 5 then add with the preceding or succeeding frequency so that the resulting sum is 5 or more. However, in doing so, we reduce the number of categories of data and will gain less information from contingency table.

(iii) The constraints on the cell frequencies if any should be linear, i.e., they should not involve square and higher powers of the frequencies such as $\sum O = \sum E = N$.

Uses of χ^2 test:

The main uses of $\chi 2$ test are:-

(i) **χ^2 test as a test of independence.** With the help of χ^2 test we can find out whether two or more attributes are associated or not. Suppose we have N observations classified according to some attributes. We may ask whether the attributes are related or independent. Thus, we can find out whether there is any association between skin colour of husband and wife. To examine the attributes are associated we formulate the null hypothesis that there is no association against an alternative hypothesis that there is an association between the attributes under study. If the calculated value of χ^2 is less than the table value at a certain level of significance, we say that the result of the experiment provide no evidence for doubting the hypothesis. On the other hand, if the calculated value of χ^2 is greater than the table value at a certain level of significance, the results of the experiment do not support the hypothesis.

(ii) **χ^2 test as a test of goodness of fit.** This is due to the fact that it enables us to ascertain how appropriately the theoretical distributions such as binomial, Poisson, Normal, etc., fit empirical distributions. When an ideal frequency curve whether normal or some other type is fitted to the data, we are interested in finding out how well this curve fits with the observed facts. A test of the concordance of the two can be made just by inspection, but such a test is obviously inadequate. Precision can be secured by applying the χ^2 test.

(iii) **χ^2 test as a test of Homogeneity.** The χ^2 test of homogeneity is an extension of the chi-square test of independence. Tests of homogeneity are designed to determine whether two or more independent random samples are drawn from the same population or from different populations. Instead of one sample as we use with independence problem we shall now have 2 or more samples. For example, we may be interested in finding out whether or not university students of various levels, i.e., middle and richer poor income groups are homogeneous in performance in the examination.

Illustration: In an anti diabetes campaign in a certain area, a particular medicine, say x was administered to 812 persons out of a total population of 3248. The number of diabetes castes is shown below:

Treatment	**Diabetes**	**No Diabetes**	**Total**
Medicine x	20	792	812
No Medicine x	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of medicine x in checking malaria.

Solution: Let us take the hypothesis that quinine is not effective in checking diabetes. Applying χ^2 test :

$$\text{Expectation of (AB)} = \frac{\text{(A) X (B)}}{\text{------------}} = \frac{240 \times 812}{\text{------------}} = 60$$

N 3248

Or E_1, i.e., expected frequency corresponding to first row and first column is 60. the bale of expected frequencies shall be:

60	752	812
180	2256	2436
240	3008	3248

O	E	$(O - E)^2$	$(O - E)^2/E$
20	60	1600	26.667
220	180	1600	8.889
792	752	1600	2.218
2216	2256	1600	0.709
			$[\sum(O - E)^2/E] = 38.593$

$\chi^2 = [\sum(O - E)^2/E] = 38.593$
$v = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$
for $v = 1$, $\chi^2_{0.05} = 3.84$

The calculated value of χ^2 is greater than the table value. The hypothesis is rejected. Hence medicine x is useful in checking malaria.

Illustration: In an experiment on immunization of cattle from tuberculosis the following results were obtained:

	Affected	Not affected
Inoculated	10	20
Not inoculated	15	5

Calculate χ^2 and discuss the effect of vaccine in controlling susceptibility to tuberculosis (5% value of χ^2 for one degree of freedom = 3.84).

Solution: Let us take the hypothesis that the vaccine is not effective in controlling susceptibility to tuberculosis. Applying χ^2 test:

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{50\,(11x5 - 20x15)^2}{30x20x25x25} = 8.3$$

Since the calculated value of χ^2 is greater than the table value the hypothesis is not true. We, therefore, conclude the vaccine is effective in controlling susceptibility to tuberculosis.

5. INDEX NUMBERS

An Index Number is used to measure the level of a certain phenomenon as compared to the level of the same phenomenon at some standard period. An Index Number is a statistical device for comparing the general level of magnitude of a group of related variables in two or more situations. If we want to compare the price level of 2004 with what it was in 2000, we may have to look into a group of variables – prices of rice, wheat, vegetables clothes, etc. Hence, we will have one figure to indicate the changes of different commodities as a whole and it is called an Index Number.

Utility of Index Number:

The main uses of index numbers are:

(i) Index Numbers are particularly useful in measuring relative changes. Example Changes in level of price, production, etc.
(ii) Index numbers are economic barometers. Various index numbers computed for different purposes, like employment, trade, agriculture are of immense value in dealing with different economic problems.
(iii) Index numbers are useful to compute the standard of living. Index numbers may measure the cost of living of different classes and comparison across groups becomes easier.
(iv) They help in formulating policies. For instance increase or decrease in wages required to study the cost of living index numbers.

Steps of construction of Index Numbers:

The main steps involved in the construction of index numbers are:

(i) **Purpose.** The researcher must clearly define the purpose for which the index numbers are to be constructed. For example, cost of living index numbers of workers in an industrial area and those of the workers of an agricultural area are different in respect of requirement. So, it is very essential to define the purpose of the index numbers.

(ii) **Selection of Base.** The base period is important for the construction of index numbers. When we select a base year, the year must be recent and normal. A normal year is one which is free from economic and natural, social and economic disturbance. Besides, when we selecting base period one of the following criteria should be considered (a) Fixed base, (b) Average base, (b) Chain Base.

(iii) **Selection of commodities.** We should include important commodities and they are representative of the defined purpose. For the purpose of finding the cost of living index number for low income groups, the selected items should be mostly consumed by that group.

(iv) **Sources of data.** The price relating to the thing to be measured must be collected. If we want to study the changes in industrial production, we must collect the prices relating to the production of various goods of factories.

(v) **Weighting.** All commodities are not equally important because different groups of people will have different preferences on different commodities. For instance, when the price of rice is doubled than the price of ice-cream, then the people suffer much, due to hike in price of rice which is essential. Therefore, a relative weight should be given for each commodity based on its importance.

(vi) **Choice of Formulae.** The index number computed based on different formulas usually produce different results. Hence, the problem is perhaps of greater theoretical than practical importance. In general, choice of the formula to be used depends upon the availability of data and the nature, propose and scope of the study.

The various methods of construction of index number are:

1. Unweighted
   - (a) Simple Aggregate
   - (b) Simple average of price relative

2. Weighted
   - (a) Weighted Aggregate
   - (b) Weighted average of price relative.

1. **Unweighted**

(a) **Simple Aggregate method.**

The price of the different commodities of the current year is added and the total and it is divided by the sum of the prices of the base year commodity and multiplied by 100: symbolically,

$$P_{01} = \frac{\Sigma P_1 \times 100}{\Sigma P_0}$$

Where,
P_{01} = Price index number for the current year with reference to the base year.
ΣP_1 = Aggregate of prices for the current year, and
ΣP_0 = Aggregate of prices for the base year.

(b) **Simple average of price relative method.**

Under this method, the price relative of each item is calculated separately and then averaged. A price relative is the price of the current year expressed as a percentage of the price of the base year:

$$P_{01} = \frac{\Sigma \left[\frac{P_1 \times 100}{P_0} \right]}{N} = \frac{\Sigma P}{N}$$

Where, N = Number of items, $P = P_1 \times 100 / P_0$

If we employ geometric mean in the place of the arithmetic mean then the formula is

$$P_{01} = \text{antilog} \frac{\Sigma \log \left[\frac{P_1 \times 100}{\Sigma P_0} \right]}{N} = \text{antilog} \frac{\Sigma \log P}{N}$$

Illustration: Compute a price index for the following by a (a) simple aggregate and (b) average of price relative method by using both arithmetic mean and geometric mean:

Commodity	A	B	C	D	E	F
Price in 2000 (Rs.)	20	30	10	25	40	50
Price in 2005 (Rs.)	25	30	15	35	45	55

Solution: Calculation for Price Index

Commodity	Price in 2000 P_0	Price in 2005 P_1	Price relative $P= P_1/P_0 \times 100$	log P
A	20	25	125	2.0969
B	30	30	100	2.0000
C	10	15	150	2.1761
D	25	35	140	2.1461
E	40	45	112.5	2.0511
F	50	55	110	2.0414
	175	205	737.5	12.5116

$$\text{(a) Simple Aggregative Index} = \frac{\Sigma P_1 \times 100}{\Sigma P_0}$$

$\Sigma P_0 = 175$, $\Sigma P_1 = 205$

$$= \frac{205}{175} \times 100 = 117.143$$

(b) (i) Arithmetic mean of Price Relatives $= \Sigma P / N$

$\Sigma P = 737.5$, $N = 6$

$= 737.5 / 6 = 122.92$

(ii) Geometric Mean of Price

$$\text{Relative Index} = \text{Antilog} \frac{\Sigma \log P}{N}$$

$$= \text{Antilog} \frac{12.5116}{6} = \text{Antilog } 2.0853 = 121.7$$

Weighted Index Numbers

Under this method, prices themselves are weighted by quantities, i.e., p*q. Thus physical quantities are used as weights. The different methods of assigning weights are:

(a) Laspeyre's method,
(b) Paasche's method,
(c) Bowley Dorfish method,
(d) Fisher's Ideal method,
(e) Marshall Edgworth method,
(f) Kelley's Method

(a) **Laspeyre's method.**

Under this method, the base year quantities are taken as weights: symbolically,

$$P_{01(La)} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

(b) **Paasche's method.**

The current year quantities are taken as weights under Paasche's method: symbolically,

$$P_{01(Pa)} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

(c) **Bowley Dorfish method.**

This is an index number got by the arithmetic mean of Laspeyre's and Paasche's methods; symbolically

$$P_{01(B)} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100 = \frac{L + P}{2}$$

Where, L = Laspeyre's method & P = Paasche's method.

(d) **Fisher's Ideal method.**

Fisher's price index number is given by the geometric mean of Laspeyre's and Paasche's Index; symbolically,

$$P01(F) = \sqrt{L \times P} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

(e) **Marshall Edgeworth Method**

$$P_{01(Ma)} = \frac{\Sigma p_1 (q_1 + q_0)}{\Sigma p_0 (q_0 + q_1)}$$

By removal of brackets,

$$P_{01(Ma)} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

(f) **Kelley's method.**

$$P_{01(K)} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

$$q = q_0 + q_1 / 2$$

Illustrations:

Calculate various weighted index number from the following data:

	Base year		Current year	
	Kilo	Rate (Rs.)	Kilo	Rate (Rs.)
Bread	10	3	10	4
Meat	20	15	16	20
Tea	2	20	3	30

Solution:

	Base year	Current year	p_1q_0	p_0q_0	p_1q_1	p_0q_1

	Kilo	Rate (Rs.)	Kilo	Rate (Rs.)				
	Q_0	p_0	Q_1	p_1				
Bread	10	3.00	10	4.00	40.00	30.00	40.00	30.00
Meat	20	15.00	16	20.00	400.00	300.00	320.00	240.00
Tea	2	20.00	3	30.00	60.00	40.00	90.00	60.00
					500.00	370.00	450.00	330.00
		Total						

(a) Laspeyre's method

$$P_{01(La)} = \frac{\Sigma p_1 q_0 \times 100}{\Sigma p_0 q_0} = \frac{500 \times 100}{370.00} = 135.1$$

(b) Paasche's method

$$P_{01(Pa)} = \frac{\Sigma p_1 q_1 \times 100}{\Sigma p_0 q_1} = \frac{450 \times 100}{370} = 136.4$$

(c) Bowley's Method

$$P_{01(B)} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100 = \frac{L + P}{2}$$

$$= \frac{L + P}{2} = \frac{135.1 + 136.1}{2} = 135.8$$

(d) Fisher's ideal formula

$$P01(F) = \sqrt{L \times P} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

$$= \sqrt{L \times P} = \sqrt{(135.1 \times 136.1)} = 135.7$$

(e) Marshall Edgeworth method

$$P_{01(Ma)} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{500 + 450}{500 + 330} \times 100$$

$$= \frac{950 \times 100}{830} = 1.14 \times 100 = 114$$

6. ANALYSIS OF TIME SERIES

An arrangement of statistical data in accordance with time of occurrence or in a chronological order is called a time series. The numerical data which we get at different points of time is known as time series. It plays an important role in economics, statistics and commerce. For example, if we observe agricultural production, sales, national income etc., over a period of time, say the last 3 or 5 years, the set of observations is called time series. The analysis of time series is done mainly for the purpose of forecasts and for evaluating the past performances.

Utility of Time series.

The main uses of time series are:

(i) It helps in understanding past behaviour and it will help in estimating the future behaviour.

(ii) It helps in planning and forecasting and it is very essential for the business and economics to prepare plans for the future.

(iii) Comparison between data of one period with that of another period is possible.

(iv) We can evaluate the progress in any field of economic and business activity with the help of time series data,.

(v) Seasonal, cyclical, secular trend of data is useful not only to economists but also to the businessmen.

Components of time series:

There are four basic types of variations and these are called the components or elements of time series. They are:

1. Secular Trend,
2. Seasonal variation,
3. Cyclical fluctuations, and
4. irregular or random fluctuations.

1. **Secular trend**

The general tendency of the time series data to increase or decrease or stagnate during a long period of time is called the secular trend, also known as long-term trend. This phenomenon is usually observed inmost of the series relating to Economics and Business, for instance, an upward tendency is usually observed in time series relating to population, production, prices, income, money in circulation etc. while a downward tendency is noticed in the time series relating to deaths, epidemics etc. due to an advancement in medical technology, improved medical facilities, better sanitation, etc. in a long term trend there are two types of trend. They are:

(i) Linear – Straight Line Trend, and
(ii) Non-Linear or Curvilinear Trend..

(i) **Linear or Straight Line Trend.** When the value of time series are plotted on a graph, then it is called the straight line trend or linear trend and if we obtain straight line.

(ii) **Non-linear or Curvilinear Trend**. When we plot the timer series values on a graph and if it forms a curve or a non-linear one, then it is called Non-linear or Curvilinear Trend.

2. **Seasonal Variation**

A variation which occurs weekly, monthly or quarterly is known as Seasonal Variation. The seasonal variation may occur due to the following reasons:

(i) **Climate and natural forces:**

The result of natural forces like climate is causing seasonal variation. For example, umbrellas are sold more in rainy season, in winter season.

(ii) **Customs and habits:**

Man-made conventions are the customs habits, fashion, etc. there is a custom of wearing new clothes, preparing sweets for Deepavali, Chritmas etc. At that time, there is more demand for cloth, sweet, etc.

3. **Cyclical Variation:**

According to Lincoln L. Chou, "Up and down movements are different from seasonal fluctuations, in that they extend over longer period of time-usually two or more years". Most of economic and business time series are influenced by the wave-like changes of prosperity and depression. There is periodic up and down movement. This movement is known as cyclical variation. There are four phases in a business cycle. They are a) Prosperity (boom), b) recession, c) depression, and d) recovery.

4. **Irregular variation:**

Irregular variations arise owing to unforeseen and unpredictable forces at random and affect the data. These variations are not a regular ones. These are caused by war, flood, strike etc.

In the classical time series model, the elements of trend, cyclical and seasonal variations are viewed resulting from systematic influences leading to either gradual growth, decline or recurrent movements, irregular movement are considered to be erratic movement. Therefore, the residual that remains after the elimination of systematic components is taken as representing irregular fluctuations.

Measurement of Secular Trend

The time series analysis is absolutely essential for planning. It guides the planners to achieve better results. The study of trend enables the planner to project the plan in a better direction. The following are the four methods which can be used for determining the trend.

(i) Free-hand or Graphic Method,
(ii) Semi-average Method,
(iii) Moving Average Method, and
(iv) Method of Least Squares.

(i) **Graphic or Free-hand Fitting Method:**

This is the easiest, simplest and the most flexible method of estimating secular trend. In this method we must plot the original data on the graph. Draw a smooth curve carefully which will show the direction of the trend, where time is shown on the horizontal axis and the value of the variables is shown on the vertical axis.

For fitting a trend line by the free-hand method, the following points should be taken into consider, they are:

a) the curve should be smooth.
b) Approximately there must be equal number of points above and below the curve
c) The total deviations of the data above the trend line must be the same as the vertical deviations below the line.
d) The sum of the squares of the vertical deviations from the trend should be as small as possible.

(ii) **Semi-average Method:**

In this method, the original data is divided into two equal parts and averages are calculated for both the parts. These averages are called semi-averages. For example, we can divide the 10 years 1993 to 2002 into two equal parts; from

1993 to 1997 and 1998 to 2002. if period is odd number of years, the value of the middle year is omitted.

We can draw the line by a straight line by joining the two points of average. By extending the line downward or upward, we can predict the future values.

(iii) **Moving Average Method:**

In the moving average method, the average value for a number of periods is consider and placing it at the centre of the time-span. It is calculated from overlapping groups of successive time series data. It simplifies the analysis and removes periodic variations; and the influence of the fluctuations is also reduced. The formula for calculating 3 yearly moving averages is:

$$\frac{a+b+c}{3}, \quad \frac{b+c+d}{3}, \quad \frac{c+d+e}{3}$$

Steps for calculating odd number of years (3, 5, 7, 9)
If we want to calculate the three-yearly moving average, then:

(i) Compute the value of first three years (1, 2, 3) and place the three year total against the middle year
(ii) Leave the first year's value and add up the values of the next three years and place the three-year total against the middle year.
(iii) this process must be continued until the last year's value is taken for calculating moving average.
(iv) the three-yearly total must be divided by 3 and placed in the next column. This is the trend value of moving average.

Even period of moving average:

If the period of moving average is 4,6,8, it is an even number. The four-yearly total cannot be placed against any year as the median 2.5 is between the second

and the third year. So the total should be placed in between the 2nd and 3rd years.

(iv) **Method of least square:**

By the method of least square, a straight line trend can be fitted, to the given time series data. With this method economic and business time series data can be fitted and can derive the results for the forecasting and prediction. The trend line is called the line of best fit. The straight line trend or the first degree parabola is represented by the mathematical equation.

$$Y = a + bX$$

Where, Y = required trend value, X = unit of time
a and b are constants
the value of the unknown or constants can be calculated by the following two normal equation.

$\Sigma Y = Na + b\Sigma X$

$\Sigma YX = a\Sigma X + b\Sigma X^2$

Where, N = the number of period
By solving the above two equation obtain the parameters of a and b.

Illustration: Calculation of Trend Values by the Method of Least Square

Year	Sales Y	Deviation from 1988 X	XY	X^2
2000	100	-2	-200	4
2001	110	-1	-110	1
2002	130	0	0	0
2003	140	+1	+140	1
2004	150	+2	+300	4
N = 5	$\Sigma Y = 630$	$\Sigma X = 0$	$\Sigma xy = 130$	$\Sigma X^2 = 10$

Since $\Sigma X = 0$

$$a = \frac{\Sigma Y}{N} = \frac{400}{5} = 80$$

$$\frac{\Sigma XY}{} \quad 52$$

$$b = \frac{}{X^2} = \frac{}{10} = 5.2$$

Hence, Y = 126 + 13X
The forecasted value for 2005 is Y = 126 + 13(3) = 165

Questions:

1. Define probability and explain various concepts of probability
2. State and explain the addition and multiplication theorem of probability with an example.
3. Define Binomial distribution Explain its properties.
4. What are the properties of Poisson distribution?
5. What are the salient features of Normal distribution?
6. Explain the utility of normal distribution in statistical analysis?
7. Explain how Poisson, binomial and normal distribution are related?
8. Distinguish between null and alternative hypothesis.
9. How you will conduct test pertaining to comparison between sample mean and population mean.
10. What are the properties of χ^2 distribution?
11. What are the uses of χ^2 test?
12. Define Index Number. Explain its uses?
13. What are the steps involved in the construction of index number.
14. Explain any four weighted index number.
15. What are the components of time series?
16. What do you mean by time series? State its utility?
17. The probability of defective needle is 0.3 in a box, find (a) the mean and standard deviation for the distribution of defective needles in a total of 1000 box, and (b) the moment coefficient of skewness and kurtosis of the distribution
18. The incidence of a certain disease is such that on the average 10% of workers suffer from it. If 10 workers are selected at random. Find the probability that (i) Exactly 4 workers suffer from the disease, (ii) not more than 2 workers suffer from the disease.
19. Out of 1000 families with 4 children each, what percentage would be expected to have (a) 2 boys and 2 girls, (b) at least one boy, (c) no girls, and (d) at the most 2 girls. Assume equal probabilities for boys and girls.
20. A multiple-choice test consists of 8 questions with 3 answers to each question (of which only one is correct). A student answer each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he get 3 or 4 and the third answer if he gets 5 or 6. To get

a distinction, the student must secure at least 75% correct answers. If there is no negative marking, what is the probability that the student secures a distinctions?

21. One fifth per cent of the blades produced by a blade manufacturing factory turn out to be defective. The blades are supplied in packets of 10. use Poisson distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 100,000 packets.
22. It is known from past experience that in a certain plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 4 accidents. Assume Poisson distribution.
23. Calculate Laspeyre's, Paasche's, Bowley's, Fisher's, Marshall Edgeworth index number from the following data:

	Base year		Current year	
	Price	Value	Price	Value
A	6	50	6	75
B	8	90	12	80
C	12	80	15	100
D	5	20	8	30
E	10	60	12	75

24. From the data given below about the treatment of 500 patients suffering from a disease, state whether the new treatment is superior to the conventional treatment:

Treatment	No. of Patients		
	Favourable	Not favourable	Total
New	250	40	290
Conventional	160	50	210
Total	410	90	500

(Given for degrees of freedom = 1, chi-square 5 per cent = 3.84)

25. 300 digits are chosen at random from a set of tables. The frequencies of the digits are as follows:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	28	29	36	31	20	35	35	30	31	25

Use χ^2 test to assets the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which they were chosen.
(Given for degrees of freedom = 1, chi-square 5 per cent = 3.84)

26. The number of defects per unit in a sample of 165 units of a manufactured product was found as follows:

Number of defects:	0	1	2	3	4
Number of units :	107	46	10	1	1

Fit a Poisson distribution to the data and test for goodness

27. Assume the mean height of soldiers to be 68 inches with a variance of 9 inches. How many solders in a regiment of 1,000 would you expect is be over 70 inches tall?
28. The weekly wages of 5,000 workmen are normally distributed around a mean of Rs.70 and with a standard deviation of Rs. 5. Estimate the number of workers whose weekly wages will be:

(a) between Rs. 70 and Rs. 72,
(b) between Rs. 69 and Rs. 72,
(c) more than Rs. 75,
(d) less than Rs. 63, and
(e) more than Rs. 80.

29. In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution?
30. Find the mean and standard deviation of a normal distribution of marks in an examination where 58 percent of the candidates obtained marks below 75, four per cent got above 80 and the rest between 75 and 80.
31. A sample of 1600 male students is found to have a mean height of 170 cms. Can it be reasonably regarded as a sample from a large population with mean height 173 cms and standard deviation 3.50 cms.
32. Fit a trend line to the following data by the free-hand method, semi-average method and moving average method.

Year	1995	1996	1997	1988	1999	2000	2001
Sales	65	95	85	115	110	120	130

33. The following table gives the sterling assets of the R.B.I. in crores of rupees:

(a) Represent the data graphically,
(b) Fit a straight line trend
(c) Show the trend on the graph

Year:	1996-97	1997-98	1998-99	1999-00	2000-01	2001-02
Assets:	83	92	71	90	169	191

Also estimate the figures for 1996-97.

***

UNIT – IV

STATISTICAL APPLICATIONS

A BRIEF INTRODUCTION TO STATISTICAL APPLICATIONS

A manager in a business organization – whether in the top level, or the middle level, or the bottom level - has to perform an important role of decision making. For solving any organizational problem – which most of the times happens to be complex in nature -, he has to identify a set of alternatives, evaluate them and choose the best alternative. The experience, expertise, rationality and wisdom gained by the manager over a period of time will definitely stand in good stead in the evaluation of the alternatives available at his disposal. He has to consider several factors, sometimes singly and sometimes jointly, during the process of decision making. He has to deal with the data of not only his organization but also of other competing organizations. It would be a challenging situation for a manager when he has to face so many variables operating simultaneously, something internal and something external. Among them, he has to identify the important variables or the dominating factors and he should be able to distinguish one factor from the other. He should be able to find which factors have similar characteristics and which factors stand apart. He should be able to know which factors have an inter play with each other and which factors remain independent. It would be advantageous to him to know whether there is any clear pattern followed by the variables under consideration. At times he may be required to have a good idea of the values that the variables would assume in future occasions. The task of a manager becomes all the more difficult in view of the risks and uncertainties surrounding the future events. It is imperative on the part of a manager to understand the impact of various policies and programmes on the development of the organization as well as the environment.

Also he should be able to understand the impact of several of the environmental factors on his organization. Sometimes a manager has to take a single stage decision and at times he is called for to take a multistage decision on the basis of various factors operating in a situation.

Statistical analysis is a tool for a manager in the process of decision making by means of the data on hand. There is hardly any managerial activity that does not involve an analysis of data. Statistical approach would enable a manager to have a scientific guess of the future events also. Statistical methods are systematic and built by several experts on firmly established theories and consequently they would enable a manager to overcome the uncertainties associated with future occasions. However, statistical tools have their shortcomings too. The limitations do not reflect on the subject. Rather they shall be traced to the methods of data collection and recording of data. Even with highly sophisticated statistical methods, one may not arrive at valid conclusions if the data collected are devoid of representative character. In any practical problem, one has to see whether the assumptions are reasonable or not, whether the data represents a wide spectrum, whether the data is adequate, whether all the conditions for the statistical tests have been fulfilled, etc. If one takes care of these aspects, it would be possible to arrive at better alternatives and more reliable solutions, thereby avoiding future shocks. While it is true that a statistical analysis, by itself, cannot solve all the problems faced by an organization, it will definitely enable a manager to comprehend the ground realities of the situation and provide a foresight in the identification of the crucial variables and the key areas so that he can locate a set of possible solutions within his ambit. A manager has to have a proper blend of the statistical theories and practical wisdom and he shall always strive for a holistic approach to solve any organizational problem. A manager has to provide some

safe-guarding measures against the limitations of the statistical tools. In the process he will be able to draw valid inferences thereby providing a clue as to the direction in which the organization shall move in future. He will be ably guided by the statistical results in the formulation of appropriate strategies for the organization. Further, he can prepare the organization to face the possible problems of business fluctuations in future and minimize the risks with the help of the early warning signals indicated by the relevant statistical tools.

A marketing manager of a company or a manager in a service organization will have occasions to come across the general public and consumers with several attendant social and psychological variables which are difficult to be measured and quantified.

Depending on the situation and the requirement, a manager may have to deal with the data of just one variable (univariate data), or data on two variables (bivariate data) or data concerning several simultaneous variables (multivariate data).

The unit on hand addresses itself to the role of a manager as a decision maker with the help of data available with him. Different statistical techniques which are suitable for different requirements are presented in this unit in a simple style. A manager shall know the strengths and weaknesses of various statistical tools. He shall know which statistical tool would be the most appropriate in a particular context so that the organization will derive the maximum benefit out of it.

The interpretation of the results from statistical analysis occupies an important place. Statistics is concerned with the aggregates and not just the individual data items or isolated measurements of certain variables. Therefore the conclusions from a statistical study will be valid for a majority of the objects and normal situations only. There are always extreme cases in any problem and

they have to be dealt with separately. Statistical tools will enable a manager to identify such outliers (abnormal cases or extreme variables) in a problem. A manager has to evaluate the statistical inferences, interpret them in the proper context and apply them in appropriate situations.

While in an actual research problem, one has to handle a large quantum of data, it is not possible to treat such voluminous data by a beginner in the subject. Keeping this point in mind, any numerical example in the present unit is based on a few data items only. It would be worthwhile to the budding managers to make a start in solving statistical problems by practicing the ones furnished in this unit.

The candidates are suggested to use hand calculators for solving statistical problems. There will be frequent occasions to use Statistical Tables of F-values furnished in this unit. The candidates are suggested to have with them a copy of the tables for easy, ready reference. The books and articles listed under the references may be consulted for further study or applications of statistical techniques in relevant research areas.

LESSON 1

CORRELATION AND REGRESSION ANALYSIS

LESSON OUTLINE

- The concept of correlation
- Determination of simple correlation coefficient
- Properties of correlation coefficient
- The concept of rank correlation
- Determination of rank correlation coefficient
- The concept of regression
- The principle of least squares
- Normal equations
- Determination of regression equations

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the concept of correlation
- calculate simple correlation coefficient
- understand the properties of correlation coefficient
- understand the concept of rank correlation
- calculate rank correlation coefficient
- resolve ties in ranks
- understand the concept of regression
- determine regression equations
- understand the managerial applications of correlation and regression

SIMPLE CORRELATION

Correlation

Correlation means the average relationship between two or more variables.

When changes in the values of a variable affect the values of another variable, we say that there is a correlation between the two variables. The two variables may move in the same direction or in opposite directions. Simply because of the presence of correlation between two variables, we cannot jump to the conclusion that there is a cause-effect relationship between them. Sometimes, it may be due to chance also.

Simple correlation

We say that the correlation is simple if the comparison involves two variables only.

TYPES OF CORRELATION

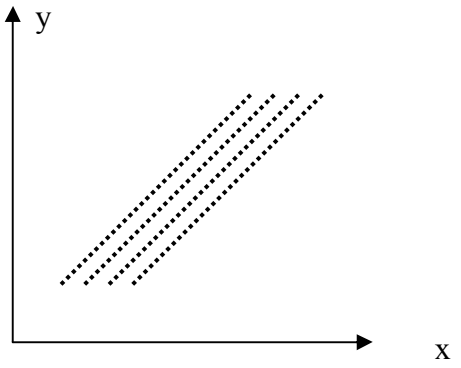
Positive correlation

If two variables x and y move in the same direction, we say that there is a positive correlation between them. In this case, when the value of one variable increases, the value of the other variable also increases and when the value of one variable decreases, the value of the other variable also decreases. Eg. The age and height of a child.

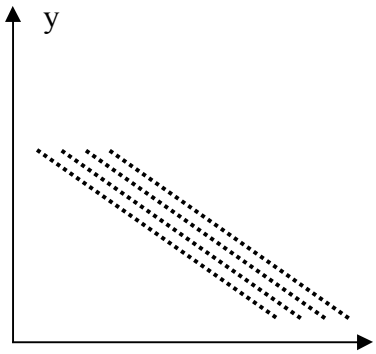
Negative correlation

If two variables x and y move in opposite directions, we say that there is a negative correlation between them. i.e., when the value of one variable increases, the value of the other variable decreases and vice versa. Eg. The price and demand of a normal good.

The following diagrams illustrate positive and negative correlations between x and y.



Positive Correlation



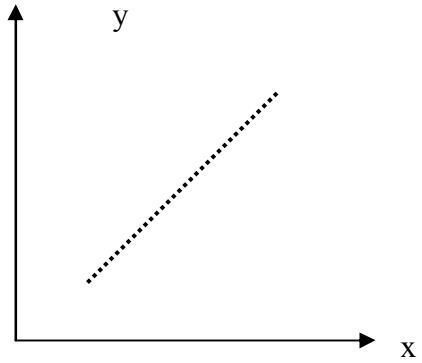
Negative Correlation

Perfect positive correlation

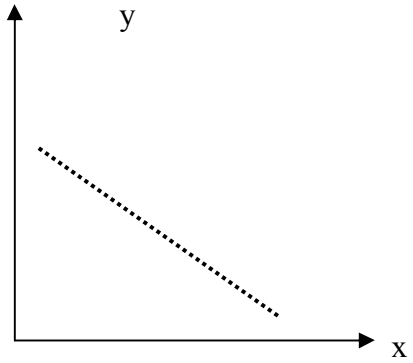
If changes in two variables are in the same direction and the changes are in equal proportion, we say that there is a perfect positive correlation between them.

Perfect negative correlation

If changes in two variables are in opposite directions and the absolute values of changes are in equal proportion, we say that there is a perfect negative correlation between them.



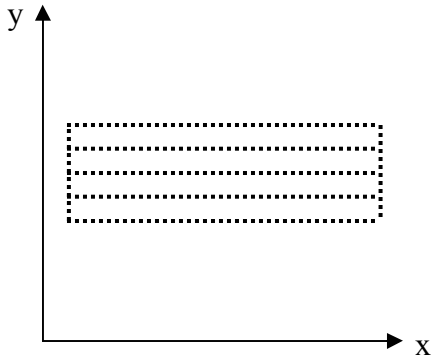
Perfect Positive Correlation



Perfect Negative Correlation

Zero correlation

If there is no relationship between the two variables, then the variables are said to be independent. In this case the correlation between the two variables is zero.



Zero correlation

Linear correlation

If the quantum of change in one variable always bears a constant ratio to the quantum of change in the other variable, we say that the two variables have a linear correlation between them.

Coefficient of correlation

The coefficient of correlation between two variables X, Y is a measure of the degree of association (i.e., strength of relationship) between them. The coefficient of correlation is usually denoted by 'r'.

Karl Pearson's Coefficient of Simple Correlation:

Let N denote the number of pairs of observations of two variables X and Y.

The correlation coefficient r between X and Y is defined by

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

This formula is suitable for solving problems with hand calculators. To apply this formula, we have to calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$.

Properties of Correlation Coefficient

Let r denote the correlation coefficient between two variables. r is interpreted using the following properties.

1. The value of r ranges from – 1 to 0 or from 0 to 1.
2. A value of r = 1 indicates that there exists perfect, positive correlation between the two variables.
3. A value of r = - 1 indicates that there exists perfect, negative correlation between the two variables.
4. A value r = 0 indicates zero correlation. i.e., It shows that there is no correlation at all between the two variables.
5. A positive value of r shows a positive correlation between the two variables.
6. A negative value of r shows a negative correlation between the two variables.
7. A value of r = 0.9 and above indicates a very high degree of positive correlation between the two variables.
8. A value of - $0.9 \geq r > -1.0$ shows a very high degree of negative correlation between the two variables.
9. For a reasonably high degree of positive correlation, we require r to be from 0.75 to 0.9
10. A value of r from 0.6 to 0.75 may be taken as a moderate degree of positive correlation.

Problem 1

The following are data on Advertising Expenditure (in Rs. Thousand) and Sales (Rs. In lakhs) in a company.

Advertising Expenditure	: 18	19	20	21	22	23
Sales	: 17	17	18	19	19	19

Determine the correlation coefficient between them and interpret the result.

Solution: We have N = 6**.** Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$ as follows:

X	Y	XY	X^2	Y^2
18	17	306	324	289
19	17	323	361	289
20	18	360	400	324
21	19	399	441	361
22	19	418	484	361
23	19	437	529	361
Total :123	109	2243	2539	1985

The correlation coefficient r between the two variables is calculated as follows:

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

$$r = \frac{6\times 2243 - 123\times 109}{\sqrt{6\times 2539 - (123)^2}\sqrt{6\times 1985 - (109)^2}}$$

= (13458 – 13407) / {√(15234- 15129) √(11910- 11881)}

=51/{√105 √29} = 51/ (10.247 X 5.365) = 51/ 54.975 = 0.9277

Interpretation

The value of r is 0.92. It shows that there is a high, positive correlation between the two variables 'Advertising Expenditure' and 'Sales'. This provides a basis to consider some functional relationship between them.

Problem 2

Consider the following data on two variables X and Y.

X : 12 14 18 23 24 27

Y : 18 13 12 30 25 10

Determine the correlation coefficient between the two variables and interpret the result.

Solution: We have N = 6**.** Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$ as follows:

	X	Y	XY	X^2	Y^2
	12	18	216	144	324
	14	13	182	196	169
	18	12	216	324	144
	23	30	690	529	900
	24	25	600	576	625
	27	10	270	729	100
Total :	118	108	2174	2498	2262

The correlation coefficient between the two variables is r =

$\{6 \text{ X } 2174 – (118 \text{ X } 108)\} / \{ \sqrt{(6 \text{ X } 2498 - 118^2)} \sqrt{(6 \text{ X } 2262 - 108^2)} \}$

= (13044 – 12744) / {√(14988- 13924) √(13572- 11664)}

=300 / {√1064 √1908} = 300 / (32.62 X 43.68) = 300 / 1424.84 = 0.2105

Interpretation

The value of r is 0.21. Even though it is positive, the value of r is very less. Hence we conclude that there is no correlation between the two variables X and Y. Consequently we cannot construct any functional relational relationship between them.

Problem 3

Consider the following data on supply and price. Determine the correlation coefficient between the two variables and interpret the result.

Supply : 11 13 17 18 22 24 26 28

Price : 25 32 26 25 20 17 11 10

Determine the correlation coefficient between the two variables and interpret the result.

Solution:

We have N = 8**.** Take X = Supply and Y = Price**.**

Calculate $\sum X, \sum Y, \sum XY, \sum X^2, \sum Y^2$ as follows:

	X	Y	XY	X^2	Y^2

11	25	275	121	625
13	32	416	169	1024
17	26	442	289	676
18	25	450	324	625
22	20	440	484	400
24	17	408	576	289
26	11	286	676	121
28	10	280	784	100
Total: 159	166	2997	3423	3860

The correlation coefficient between the two variables is r =

$\{8 \text{ X } 2997 – (159 \text{ X } 166)\} / \{ \sqrt{(8 \text{ X } 3423 - 159^2)} \sqrt{(8 \text{ X } 3860 - 166^2)} \}$

= (23976 – 26394) / {√(27384- 25281) √(30880- 27566)}

= - 2418 / {√2103 √3314} = - 2418 / (45.86 X 57.57)

= - 2418 / 2640.16 = - 0.9159

Interpretation

The value of r is - 0.92. The negative sign in r shows that the two variables move in opposite directions. The absolute value of r is 0.92 which is very high. Therefore we conclude that there is high negative correlation between the two variables 'Supply' and 'Price'.

Problem 4

Consider the following data on income and savings in Rs. thousand.

Income : 50 51 52 55 56 58 60 62 65 66

Savings : 10 11 13 14 15 15 16 16 17 17

Determine the correlation coefficient between the two variables and interpret the result.

Solution:

We have N = 10**.** Take X = Income and Y = Savings**.**

Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$ as follows:

X	Y	XY	X^2	Y^2
50	10	500	2500	100
51	11	561	2601	121
52	13	676	2704	169
55	14	770	3025	196
56	15	840	3136	225
58	15	870	3364	225
60	16	960	3600	256
62	16	992	3844	256
65	17	1105	4225	289
66	17	1122	4356	289
Total: 575	144	8396	33355	2126

The correlation coefficient between the two variables is r =

$\{10 \text{ X } 8396 - (575 \text{ X } 144)\} / \{\sqrt{(10 \text{ X } 33355 - 575^2)}\ \sqrt{(10 \text{ X } 2126 - 144^2)}\}$

= (83960 – 82800) / {√(333550- 330625) √(21260- 20736)}

= 1160 / {√2925 √524} = 1160 / (54.08 X 22.89)

= 1160 / 1237.89 = 0.9371

Interpretation

The value of r is 0.93. The positive sign in r shows that the two variables move in the same direction. The value of r is very high. Therefore we conclude that there is high positive correlation between the two variables 'Income' and 'Savings'. As a result, we can construct a functional relationship between them.

RANK CORRELATION

Spearman's Rank Correlation Coefficient

If ranks can be assigned to pairs of observations for two variables X and Y, then the correlation between the ranks is called the **rank correlation coefficient**. It is usually denoted by the **symbol** ρ (rho). It is given by the formula

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

where D = difference between the corresponding ranks of X and Y

$= R_X - R_Y$

and N is the total number of pairs of observations of X and Y.

Problem 5

Alpha Recruiting Agency short listed 10 candidates for final selection. They were examined in written and oral communication skills. They were ranked as follows:

Candidate's Serial No.	1	2	3	4	5	6	7	8	9	10
Rank in written communication	8	7	2	10	3	5	1	9	6	4
Rank in oral communication	10	7	2	6	5	4	1	9	8	3

Find out whether there is any correlation between the written and oral communication skills of the short listed candidates.

Solution:

Take X = Written communication skill and Y = Oral communication skill.

RANK OF X: R_1	RANK OF Y: R_2	$D=R_1 - R_2$	D^2
8	10	- 2	4
7	7	0	0
2	2	0	0

10	6	4	16
3	5	- 2	4
5	4	1	1
1	1	0	0
9	9	0	0
6	8	- 2	4
4	3	1	1

Total: 30

We have N = 10. The rank correlation coefficient is

$\rho = 1 - \{6 \Sigma D^2 / (N^3 - N)\} = 1 - \{6 \times 30 / (1000 - 10)\} = 1 - (180 / 990)$

$= 1 - 0.18 = 0.82$

Inference:

From the value of r, it is inferred that there is a high, positive rank correlation between the written and oral communication skills of the short listed candidates.

Problem 6

The following are the ranks obtained by 10 workers in ABC Company on the basis of their length of service and efficiency.

Ranking as per service	1	2	3	4	5	6	7	8	9	10
Rank as per efficiency	2	3	6	5	1	10	7	9	8	4

Find out whether there is any correlation between the ranks obtained by the workers as per the two criteria.

Solution:

Take X = Length of service and Y = Efficiency.

Rank of X: R_1	Rank of Y: R_2	D= R_1- R_2	D^2
1	2	- 1	1
2	3	- 1	1
3	6	- 3	9
4	5	- 1	1
5	1	4	16

6	10	- 4	16
7	7	0	0
8	9	- 1	1
9	8	1	1
10	4	6	36
		Total	82

We have N = 10. The rank correlation coefficient is

$\rho = 1 - \{ 6 \Sigma D^2 / (N^3 - N)\} = 1 - \{ 6 \times 82 / (1000 - 10) \} = 1 - (492 / 990)$

$= 1 - 0.497 = 0.503$

Inference:

The rank correlation coefficient is not high.

Problem 7 (Conversion of scores into ranks)

Calculate the rank correlation to determine the relationship between equity shares and preference shares given by the following data on their price.

Equity share	90.0	92.4	98.5	98.3	95.4	91.3	98.0	92.0
Preference share	76.0	74.2	75.0	77.4	78.3	78.8	73.2	76.5

Solution:

From the given data on share price, we have to find out the ranks for equity shares and preference shares.

Step 1. First, consider the equity shares and arrange them in descending order of their price as 1,2,…,8. We have the following ranks.

Equity share	98.5	98.3	98.0	95.4	92.4	92.0	91.3	90.0
Rank	1	2	3	4	5	6	7	8

Step 2. Next, take the preference shares and arrange them in descending order of their price as 1,2,…,8. We obtain the following ranks.

Preference share	78.8	78.3	77.4	76.5	76.0	75.0	74.2	73.2
Rank	1	2	3	4	5	6	7	8

Step 3. Calculation of D^2:

Fit the given data with the correct rank. Take X = Equity share and Y = Preference share. We have the following table.

X	Y	Rank of X: R_1	Rank of Y: R_2	$D=R_1- R_2$	D^2
90.0	76.0	8	5	3	9
92.4	74.2	5	7	- 2	4
98.5	75.0	1	6	- 5	25
98.3	77.4	2	3	- 1	1
95.4	78.3	4	2	2	4
91.3	78.8	7	1	6	36
98.0	73.2	3	8	- 5	25
92.0	76.5	6	4	2	4
				Total	108

Step 4. Calculation of ρ:

We have N = 8. The rank correlation coefficient is

$\rho = 1 - \{ 6 \Sigma D^2 / (N^3 - N)\} = 1 - \{ 6 \times 108 / (512 - 8) \} = 1 - (648 / 504)$

$= 1 - 1.29 = - 0.29$

Inference:

From the value of ρ, it is inferred that the equity shares and preference shares under consideration are negatively correlated. However, the absolute value of ρ is 0.29 which is not even moderate.

Problem 8

Three managers evaluate the performance of 10 sales persons in an organization and award ranks to them as follows:

Sales Person	1	2	3	4	5	6	7	8	9	10
Rank awarded by Manager I	8	7	6	1	5	9	10	2	3	4
Rank awarded by Manager II	7	8	4	6	5	10	9	3	2	1
Rank awarded by	4	5	1	8	9	10	6	7	3	2

Manager III										

Determine which two managers have the nearest approach in the evaluation of the performance of the sales persons.

Solution:

Sales Person	Manager I Rank: R_1	Manager II Rank: R_2	Manager III Rank: R_3	$(R_1 - R_2)^2$	$(R_1 - R_3)^2$	$(R_2 - R_3)^2$
1	8	7	4	1	16	9
2	7	8	5	1	4	9
3	6	4	1	4	25	9
4	1	6	8	25	49	4
5	5	5	9	0	16	16
6	9	10	10	1	1	0
7	10	9	6	1	16	9
8	2	3	7	1	25	16
9	3	2	3	1	0	1
10	4	1	2	9	4	1
			Total	44	156	74

We have N = 10. The rank correlation coefficient between mangers I and II is

$\rho = 1 - \{ 6 \Sigma D^2 / (N^3 - N)\} = 1 - \{ 6 \times 44 / (1000 - 10) \} = 1 - (264 / 990)$

$= 1 - 0.27 = 0.73$

The rank correlation coefficient between mangers I and III is

$1 - \{ 6 \times 156 / (1000 - 10) \} = 1 - (936 / 990) = 1 - 0.95 = 0.05$

The rank correlation coefficient between mangers II and III is

$1 - \{ 6 \times 74 / (1000 - 10) \} = 1 - (444 / 990) = 1 - 0.44 = 0.56$

Inference:

Comparing the 3 values of ρ, it is inferred that Mangers I and II have the nearest approach in the evaluation of the performance of the sales persons.

Repeated values: Resolving ties in ranks

When ranks awarded to candidates, it is possible that certain candidates obtain equal ranks. For example two, or three, or four candidates secure equal ranks. A procedure to resole the ties is described below.

We follow the **Average Rank Method.** If there are n items, arrange them in ascending order or descending order and give ranks 1, 2, 3, …, n. Then look at those items which have equal values. For such items, take the average ranks.

If there are two items with equal values, their ranks will be two consecutive integers, say s and s + 1. Their average is { s + (s+1)} / 2. Assign this rank to both items. Note that we allow ranks to be fractions also.

If there are three items with equal values, their ranks will be three consecutive integers, say s, s + 1 and s + 2. Their average is { s + (s+1) + (s+2) } / 3 = (3s + 3) / 3 = s + 1. Assign this rank to all three items. A similar procedure is followed if four or more number of items have equal values.

Correction term for ρ when ranks are tied

Consider the formula for rank correlation coefficient. We have

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

If there is a tie involving m items, we have to add

$$\frac{m^3 - m}{12}$$

to the term D^2 in ρ. We have to add as many terms like $(m^3 – m) / 12$ as there are ties.

Let us calculate the correction terms for certain values of m. These are provided in the following table.

m	m^3	$m^3 - m$	Correction term = $\frac{m^3 - m}{12}$
2	8	6	0.5
3	27	24	2
4	64	60	5
5	125	120	10

Illustrative examples:

If there is a tie involving 2 items, then the correction term is 0.5
If there are 2 ties involving 2 items each, then the correction term is
0.5 + 0.5 = 1
If there are 3 ties with 2 items each, then the correction term is
0.5 + 0.5 + 0.5 = 1.5
If there is a tie involving 3 items, then the correction term is 2
If there are 2 ties involving 3 items each, then the correction term is 2 + 2 = 4
If there is a tie with 2 items and another tie with 3 items, then the correction term is 0.5 + 2 = 2.5
If there are 2 ties with 2 items each and another tie with 3 items, then the correction term is 0.5 + 0.5 + 2 = 3

Problem 9 : Resolving ties in ranks

The following are the details of ratings scored by two popular insurance schemes. Determine the rank correlation coefficient between them.

Scheme I	**80**	**80**	**83**	**84**	**87**	**87**	**89**	**90**
Scheme II	**55**	**56**	**57**	**57**	**57**	**58**	**59**	**60**

Solution:

From the given values, we have to determine the ranks.
Step 1. Arrange the scores for Insurance Scheme I in descending order and rank them as 1,2,3,…,8.

Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3	4	5	6	7	8

The score 87 appears twice. The corresponding ranks are 3, 4. Their average is (3 + 4) / 2 = 3.5. Assign this rank to the two equal scores in Scheme I.

The score 80 appears twice. The corresponding ranks are 7, 8. Their average is (7 + 8) / 2 = 7.5. Assign this rank to the two equal scores in Scheme I.

The revised ranks for Insurance Scheme I are as follows:

Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3.5	3.5	5	6	7.5	7.5

Step 2. Arrange the scores for Insurance Scheme II in descending order and rank them as 1,2,3,…,8.

Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	4	5	6	7	8

The score 57 appears thrice. The corresponding ranks are 4, 5, 6.
Their average is (4 + 5 + 6) / 3 = 15 / 3 = 5. Assign this rank to the three equal scores in Scheme II.

The revised ranks for Insurance Scheme II are as follows:

Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	5	5	5	7	8

Step 3. Calculation of D^2:

Assign the revised ranks to the given pairs of values and calculate D^2 as follows:

Scheme I Score	Scheme II Score	Scheme I Rank: R_1	Scheme II Rank: R_2	$D = R_1 - R_2$	D^2
80	55	7.5	8	- 0.5	0.25
80	56	7.5	7	0.5	0.25
83	57	6	5	1	1
84	57	5	5	0	0
87	57	3.5	5	- 1.5	2.25

87	58	3.5	3	0.5	0.25
89	59	2	2	0	0
90	60	1	1	0	0
				Total	4

Step 4. Calculation of ρ:

We have N = 8.
Since there are 2 ties with 2 items each and another tie with 3 items, the correction term is 0.5 + 0.5 + 2 .
The rank correlation coefficient is

$\rho = 1 - [\{ 6 \Sigma D^2 + (1/2) + (1/2) + 2 \} / (N^3 - N)\}]$

$= 1 - \{ 6 (4.+0.5+0.5+2) / (512 - 8) \} = 1 - (6 \times 7 / 504) = 1 - (42/504)$

$= 1 - 0.083 = 0.917$

Inference:

It is inferred that the two insurance schemes are highly, positively correlated.

REGRESSION

In the pairs of observations, if there is a cause and effect relationship between the variables X and Y, then the average relationship between these two variables is called regression.

Regression means "stepping back" or "return to the average". The linear relationship giving the best mean value of a variable corresponding to the other variable is called a **regression line** or **the line of the best fit**. The regression of X on Y is different from the regression of Y on X. Thus there are one two equations of regression and the two regression lines are given as follows:

Regression of Y on X: $Y - \bar{Y} = b_{yx}(X - \bar{X})$

Regression of X on Y: $X - \bar{X} = b_{xy}(Y - \bar{Y})$

where $\bar{X}$, $\bar{Y}$ are the means of X, Y respectively.

Result:

Let σ_x, σ_y denote the standard deviations of x, y respectively. We have the following result.

$$b_{yx} = r\frac{\sigma_Y}{\sigma_X} \quad and \quad b_{xy} = r\frac{\sigma_X}{\sigma_Y}$$
$$\therefore r^2 = b_{yx}b_{xy} \quad and \ so \quad r = \sqrt{b_{yx}b_{xy}}$$

Result:

The coefficient of correlation r between X and Y is the square root of the product of the b values in the two regression equations. We can find r by in this way also.

Application

The method of regression is very much useful for business forecasting.

PRINCIPLE OF LEAST SQUARES

Let x, y be two variables under consideration. Out of them, let x be an independent variable and let y be a dependent variable, depending on x. We desire to build a functional relationship between them. For this purpose, the first and foremost requirement is that x, y have a high degree of correlation. If the correlation coefficient between x and y is moderate or less, we shall not go ahead with the task of fitting a functional relationship between them.

Suppose there is a high degree of correlation (positive or negative) between x and y. Suppose it is required to build a linear relationship between them. i.e., We want a regression of y on x.

Geometrically speaking, if we plot the corresponding values of x and y in a 2-dimensional plane and join such points, we shall obtain a straight line. However, hardly we can expect all the pairs (x, y) to lie on a straight line. We can consider several straight lines which are, to some extent, near all the points

(x, y). Consider one line. An observation (x_1, y_1) may be either above the line of consideration or below the line. Project this point on the x-axis. It will meet the straight line at the point (x_1, y_{1e}). Here the theoretical value (or the expected value) of the variable is y_{1e} while the observed value is y_1. When there is a difference between the expected and observed values, there appears an error. This error is $E_1 = y_1 - \hat{y}_1$. This is positive if (x_1, y_1) is a point above the line and negative if (x_1, y_1) is a point below the line. For the n pairs of observations, we have the following n quantities of error:

$E_1 = y_1 - \hat{y}_1$,
$E_2 = y_2 - \hat{y}_2$,
.
.
.

$E_n = y_n - \hat{y}_n$.

Some of these quantities are positive while the remaining ones are negative. However, the squares of all these quantities are positive.

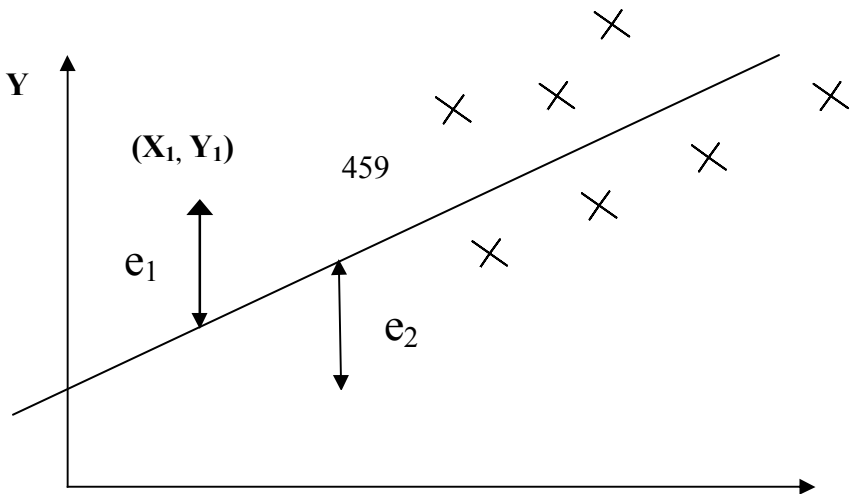
i.e., $E^2_1 = (y_1 - \hat{y}_1)^2 \geq 0$, $E^2_2 = (y_2 - \hat{y}_2)^2 \geq 0$, …, $E^2_n = (y_n - \hat{y}_n)^2 \geq 0$.

Hence the sum of squares of errors (SSE) $= E^2_1 + E^2_2 + \ldots + E^2_n$

$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 \geq 0$.

Among all those straight lines which are somewhat near to the given observations

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we consider that straight line as the ideal one for which the SSE is the least. Since the ideal straight line giving regression of y on x is based on this concept, we call this principle as the **principle of least squares**.



Normal equations

Suppose we have to fit a straight line to the n pairs of observations (x_1, y_1), (x_2, y_2), …,

(x_n, y_n). Suppose the equation of straight line finally comes as

$$Y = a + b X \qquad (1)$$

where a, b are constants to be determined. Mathematically speaking, when we require to find the equation of a straight line, two distinct points on the straight line are sufficient. However, a different approach is followed here. We want to include all the observations in our attempt to build a straight line. Then all the n observed points (x, y) are required to satisfy the relation (1). Consider the summation of all such terms. We get

$$\sum y = \sum (a + b x) = \sum (a .1 + b x) = (\sum a.1) + (\sum b x) = a (\sum 1) + b (\sum x).$$

$$\text{i.e. } \sum y = an + b (\sum x) \qquad (2)$$

To find two quantities a and b, we require two equations. We have obtained one equation i.e., (2). We need one more equation. For this purpose, multiply both sides of (1) by x. We obtain

$$x y = ax + bx^2 .$$

Consider the summation of all such terms. We get

$\sum x\,y = \sum (ax + bx^2) = (\sum a\,x) + (\sum bx^2)$

i.e., $\sum x\,y = a(\sum x) + b(\sum x^2)$ …….. (3)

Equations (2) and (3) are referred to as the normal equations associated with the regression of y on x. Solving these two equations, we obtain

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - \left(\sum X\right)^2}$$

$$\text{and} \quad b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

Note. For calculating the correlation coefficient,

we require $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$.

For calculating the regression of y on x, we require $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$. Thus the tabular column is the same in both cases, with the difference that $\sum Y^2$ is also required for the correlation coefficient.

Next, if we consider the regression line of x on y, we get the equation $X = a + b\,Y$. The expressions for the coefficients can be got by interchanging the roles of X and Y in the previous discussion. Thus we obtain

$$a = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{n\sum Y^2 - \left(\sum Y\right)^2}$$

$$\text{and} \quad b = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - \left(\sum Y\right)^2}$$

Problem 10

Consider the following data on sales and profit.

X	5	6	7	8	9	10	11

Y	2	4	5	5	3	8	7

Determine the regression of profit on sales.

Solution:

We have N = 7**.** Take X = Sales**,** Y = Profit.

Calculate $\sum X, \sum Y, \sum XY, \sum X^2$ as follows:

X	Y	XY	X^2
5	2	10	25
6	4	24	36
7	5	35	49
8	5	40	64
9	3	27	81
10	8	80	100
11	7	77	121
Total: 56	34	293	476

$a = \{(\sum x^2)(\sum y) - (\sum x)(\sum x y)\} / \{n(\sum x^2) - (\sum x)^2\}$

$= (476 \times 34 - 56 \times 293) / (7 \times 476 - 56^2) = (16184 - 16408) / (3332 - 3136)$

$= -224 / 196 = -1.1429$

$b = \{n(\sum x y) - (\sum x)(\sum y)\} / \{n(\sum x^2) - (\sum x)^2\}$

$= (7 \times 293 - 56 \times 34) / 196 = (2051 - 1904) / 196 = 147 / 196 = 0.75$

The regression of Y on X is given by the equation

$Y = a + b X$

i.e., $Y = -1.14 + 0.75 X$

Problem 11

The following are the details of income and expenditure of 10 households.

Income	40	70	50	60	80	50	90	40	60	60
Expenditure	25	60	45	50	45	20	55	30	35	30

Determine the regression of expenditure on income and estimate the expenditure when the income is 65.

Solution:
We have N = 10**.** Take X = Income**,** Y = Expenditure
Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ as follows:

X	Y	XY	X^2
40	25	1000	1600
70	60	4200	4900
50	45	2250	2500
60	50	3000	3600
80	45	3600	6400
50	20	1000	2500
90	55	4950	8100
40	30	1200	1600
60	35	2100	3600
60	30	1800	3600
Total: 600	395	25100	38400

$a = \{(\sum x^2)(\sum y) - (\sum x)(\sum x y)\} / \{n(\sum x^2) - (\sum x)^2\}$

$= (38400 \times 395 - 600 \times 25100) / (10 \times 38400 - 600^2)$

$= (15168000 - 15060000) / (384000 - 360000) = 108000 / 24000 = 4.5$

$b = \{n(\sum x y) - (\sum x)(\sum y)\} / \{n(\sum x^2) - (\sum x)^2\}$

$= (10 \times 25100 - 600 \times 395) / 24000 = (251000 - 237000) / 24000$

$= 14000 / 24000 = 0.58$

The regression of Y on X is given by the equation

$Y = a + b X$

i.e., $Y = 4.5 + 0.583 X$

To estimate the expenditure when income is 65:

Take X = 65 in the above equation. Then we get
Y = 4.5 + 0.583 x 65 = 4.5 + 37.895 = 42.395 = 42 (approximately).

Problem 12

Consider the following data on occupancy rate and profit of a hotel.

Occupancy rate	40	45	70	60	70	75	70	80	95	90
Profit	50	55	65	70	90	95	105	110	120	125

Determine the regressions of (i) profit on occupancy rate and

(ii) occupancy rate on profit.

Solution:

We have N = 10**.** Take X = Occupancy rate**,** Y = Profit.
Note that in Problems 10 and 11, we wanted only one regression line and so we did not take $\sum Y^2$. Now we require two regression lines. Therefore, calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$.

S

X	Y	XY	X^2	Y^2
40	50	2000	1600	2500
45	55	2475	2025	3025
70	65	4550	4900	4225
60	70	4200	3600	4900
70	90	6300	4900	8100
75	95	7125	5625	9025
70	105	7350	4900	11025
80	110	8800	6400	12100
95	120	11400	9025	14400
90	125	11250	8100	15625
Total: 695	885	65450	51075	84925

The regression line of Y on X:

$Y = a + b X$

where $a = \{(\sum x^2)(\sum y) - (\sum x)(\sum x y)\} / \{n(\sum x^2) - (\sum x)^2\}$

and $b = \{n(\sum x y) - (\sum x)(\sum y)\} / \{n(\sum x^2) - (\sum x)^2\}$

We obtain

a = (51075 x 885 – 695 x 65450) / (10x51075 - 695²)

= (45201375 – 45487750)/ (510750 – 483025)

= - 286375 / 27725 = - 10.329

b = (10 x 65450 – 695 x 885) / 27725

= (654500 – 615075) / 27725 = 39425 / 27725 = 1.422

So, the regression equation is Y = - 10.329 + 1.422 X

Next, if we consider **the regression line of X on Y**,

we get the equation $X = a + b Y$ where

$a = \{(\sum y^2)(\sum x) - (\sum y)(\sum x y)\} / \{n(\sum y^2) - (\sum y)^2\}$

and $b = \{n(\sum x y) - (\sum x)(\sum y)\} / \{n(\sum y^2) - (\sum y)^2\}$.

We get

a = (84925 x 695 – 885 x 65450) / (10 x 84925 – 885²)

= (59022875 – 57923250) / (849250 – 783225) = 1099625 / 66025 = 16.655,

b = (10 x 65450 – 695 x 885) / 66025 = (654500 – 615075) / 66025

= 39425 / 66025 = 0.597

So, the regression equation is X = 16.655 + 0.597 Y

Note: For the data given in this problem**,** if we use the formula for r, we get

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

= (10 x 65450 – 695 x 885) / { √ (10 x 51075 - 695²) √ (10 x 84925 - 885²) }

= (654500 – 615075) / (√ 27725 √ 66025) = 39425 / 166.508 x 256.95

= 39425 / 42784.23 = 0.9214

However, once we know the two b values, we can find the coefficient of correlation r between X and Y as the square root of the product of the two b values.

Thus we obtain

$r = \sqrt{(1.422 \times 0.597)} = \sqrt{0.848934} = 0.9214.$

Note that this agrees with the above value of r.

QUESTIONS

1. Explain the aim of 'Correlation Analysis'.
2. Distinguish between positive and negative correlation.
3. State the formula for simple correlation coefficient.
4. State the properties of the correlation coefficient.
5. What is 'rank correlation'? Explain.
6. State the formula for rank correlation coefficient.
7. Explain how to resolve ties while calculating ranks.
8. Explain the concept of regression.
9. What is the principle of least squares? Explain.
10. Explain normal equations in the context of regression analysis.
11. State the formulae for the constant term and coefficient in the regression equation.
12. State the relationship between the regression coefficient and correlation coefficient.
13. Explain the managerial uses of Correlation Analysis and Regression Analysis.

UNIT IV

LESSON 2 ANALYSIS OF VARIANCE

LESSON OUTLINE

- Definition of ANOVA
- Assumptions of ANOVA
- Classification of linear models
- ANOVA for one-way classified data
- ANOVA table for one-way classified data
- Null and Alternative Hypotheses
- Type I Error
- Level of significance
- SS, MSS and Variance ratio
- Calculation of F value
- Table value of F
- Coding Method
- Inference from ANOVA table
- Managerial applications of ANOVA

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the concept of ANOVA
- formulate Null and Alternative Hypotheses
- construct ANOVA table for one-way classified data
- calculate T, N and CF
- calculate SS, df and MSS
- calculate F value
- find the table value of F
- draw inference from ANOVA
- apply coding method
- understand the managerial applications of ANOVA

ANALYSIS OF VARIANCE (ANOVA)

Introduction

For managerial decision making, sometimes one has to carry out tests of significance. The analysis of variance is an effective tool for this purpose. The objective of the analysis of variance is to test the homogeneity of the means of different samples.

Definition

The following definition was given by R.A. Fisher: "Analysis of variance is the separation of variance ascribable to one group of causes from the variance ascribable to other groups".

Assumptions of ANOVA

The technique of ANOVA is mainly used for the analysis and interpretation of data obtained from experiments. This technique is based on three important assumptions namely,

1. The parent population is normal.
2. The error component is distributed normally with zero mean and constant variance.
3. The various effects are additive in nature.

The technique of ANOVA essentially consists of partitioning the total variation in an experiment into components of different sources of variation. These sources of variations are due to controlled factors and uncontrolled factors. Since the variation in the sample data is characterized by means of many components of variation, it can be symbolically represented in the mathematical form called a linear model for the sample data.S

Classification of models

Such linear models for the sample data may broadly be classified into three types as follows:

1. Random effect model
2. Fixed effect model
3. Mixed effect model

In any variance components model, the error component has always random effects, since it occurs purely in a random manner. All other components may be either mixed or random.

Random effect model

A model in which each of the factors has random effect (including error effect) is called a random effect model or simply a random model.

Fixed effect model

A model in which each of the factors has fixed effects, buy only the error effect is random is called a fixed effect model or simply a fixed model.

Mixed effect model

A model in which some of the factors have fixed effects and some others have random effects is called a mixed effect model or simply a mixed model.

In what follows, we shall restrict ourselves to a fixed effect model.

In a fixed effect model, the main objective is to estimate the effects and find the measure of variability among each of the factors and finally to find the variability among the error effects.

The ANOVA technique is mainly based on the linear model which depends on the types of data used in the linear model. There are several types of data in ANOVA, depending on the number of sources of variation namely,

One-way classified data,

Two-way classified data,
…
m-way classified data.

One-way classified data

When the set of observations is distributed over different levels of a single factor, then it gives one-way classified data.

ANOVA for One-way classified data

Let y_{ij} denote the jth observation corresponding to the ith level of factor A and Y_{ij} the corresponding random variate.

Define the linear model for the sample data obtained from the experiment by the equation

$$y_{ij} = \mu + a_i + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

where μ represents the general mean effect which is fixed and which represents the general condition of the experimental units, a_i denotes the fixed effect due to ith level of the factor A (i=1,2,…,k) and hence the variation due to a_i (i=1,2,…,k) is said to be control.

The last component of the model e_{ij} is the random variable. It is called the error component and it makes the Y_{ij} a random variate. The variation in e_{ij} is due to all the uncontrolled factors and e_{ij} is independently, identically and normally distributed with mean zero and constant variance σ^2.

For the realization of the random variate $Y_{ij,}$ consider y_{ij} defined by

$$y_{ij} = \mu + a_i + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

The expected value of the general observation y_{ij} in the experimental units is given by $E(y_{ij}) = \mu_i \ \ for\ all\ \ i = 1,2,...,k$

with $y_{ij} = \mu_i + e_{ij}$, where e_{ij} is the random error effect due to uncontrolled factors (i.e., due to chance only).

Here we may expect $\mu_i = \mu \ for\ all\ \ i = 1,2,...,k$, if there is no variation due to control factors. If it is not the case, we have

$$\mu_i \neq \mu \ for\ all\ i = 1,2,...,k$$
$$i.e., \mu_i - \mu \neq 0 \ for\ all\ i = 1,2,...,k$$
$$Suppose\ \ \mu_i - \mu \neq a_i.$$
$$Then\ we\ have\ \ \mu_i \neq \mu + a_i \ for\ all\ i = 1,2,...,k$$

On substitution for μ_i in the above equation, the linear model reduces to

$$y_{ij} = \mu + a_i + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix} \qquad (1)$$

The objective of ANOVA is to test the null hypothesis $H_o : \mu_i = \mu \ for\ all\ i = 1,2,...,k$ or $H_o : a_i = 0 \ for\ all\ i = 1,2,...,k$. For carrying out this test, we need to estimate the unknown parameters $\mu, a_i \ for\ all\ i = 1,2,...,k$ by the principle of least squares. This can be done by minimizing the residual sum of squares defined by

$$E = \sum_{ij} e^2_{ij}$$
$$= \sum_{ij} (y_{ij} - \mu - a_i)^2,$$

using (1). The normal equations can be obtained by partially differentiating E with respect to μ and $a_i \ for\ all\ i = 1,2,...,k$ and equating the results to zero. We obtain

$$G = N\mu + \sum_i n_i a_i \qquad (2)$$

and $T_i = n_i \mu + n_i a_i,\ i = 1,2,\ldots,k \qquad (3)$

where N = nk. We see that the number of variables (k+1) is more than the number of independent equations (k). So, by the theorem on a system of liner equations, it follows that unique solution for this system is not possible.

However, by making the assumption that $\sum_i n_i a_i = 0$, we can get a unique solution for μ and a_i (i = 1,2,…,k). Using this condition in equation (2), we get

$$G = N\mu$$

$$i.e.\ \mu = \frac{G}{N}$$

Therefore the estimate of μ is given by $\hat{\mu} = \frac{G}{N}$ (4)

Again from equation (2), we have

$$\frac{T_i}{n_i} = \mu + a_i$$

$$Hence,\ a_i = \frac{T_i}{n_i} - \mu$$

Therefore, the estimate of a_i is given by

$$\hat{a}_i = \frac{T_i}{n_i} - \hat{\mu}$$

$$\text{i.e.,} \qquad \hat{a}_i = \frac{T_i}{n_i} - \frac{G}{N} \tag{5}$$

Substituting the least square estimates of $\hat{\mu}$ and $\hat{a}_i$ in the residual sum of squares, we get

$$E = \sum_{ij} (y_{ij} - \hat{\mu} - \hat{a}_i)^2$$

After carrying out some calculations and using the normal equations (2) and (3) we obtain

$$E = \left(\sum_{ij} y_{ij}^2 - \frac{G^2}{N} \right) - \left(\sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N} \right) \tag{6}$$

The first term in the RHS of equation (6) is called the **corrected total sum of squares** while $\sum_{ij} y_{ij}^2$ is called the **uncorrected total sum of squares**.

For measuring the variation due to treatment (controlled factor), we consider the null hypothesis that all the treatment effects are equal. i.e.,

$$H_o : \mu_1 = \mu_2 = ... = \mu_k = \mu$$
$$i.e.,\ H_o : \mu_i = \mu \ \ for\ all\ i = 1,2,...,k$$
$$i.e.,\ H_o : \mu_i - \mu = 0 \ \ for\ all\ i = 1,2,...,k$$
$$i.e.,\ H_o : a_i = 0$$

Under H_o, the linear model reduces to

$$y_{ij} = \mu + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

Proceeding as before, we get the residual sum of squares for this hypothetical model as

$$E_1 = \left(\sum_{ij} y_{ij}^2 \right) - \frac{G^2}{N} \tag{7}$$

Actually, E_1 contains the variation due to both treatment and error. Therefore a measure of variation due to treatment can be obtained by "$E_1 - E$". Using (6) and (7), we get

$$E_1 - E = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{G^2}{N} \tag{8}$$

The expression in (8) is usually called the **corrected treatment sum of squares** while the term $\sum_{i=1}^{k} \frac{T_i^2}{n_i}$ is called **uncorrected treatment sum of squares**. Here it may be noted that $\frac{G^2}{N}$ is a correction factor (Also called a correction term). Since E is based on N-k free observations, it has N - k degrees of freedom (df). Similarly, since E_1 is based on N -1 free observation, E_1 has N -1 degrees of freedom. So $E_1 - E$ has k -1 degrees of freedom.

When actually the null hypothesis is true, if we reject it on the basis of the estimated value in our statistical analysis, we will be committing **Type – I error**. The probability for committing this error is referred to as the **level of significance**, denoted by α. The testing of the null hypothesis H_o may be carried out by F test. For given α, we have

$$F = \frac{TrMSS}{EMSS} = \frac{Trss/dF}{Ess/dF} \sim F_{k-1,N-k}.$$

i.e., It follows F distribution with degrees of freedom k-1 and N-k.

All these values are represented in the form of a table called ANOVA table, furnished below.

ANOVA Table for one-way classified data

Source of Variation	Degrees of freedom	Sum of Squares (SS)	Mean Squares (MS)	Variance ratio F
Between the level of the factor (Treatment)	k-1	$E_1 - E = Q_T$ $\sum_i^k \frac{T_i^2}{n_i} - \frac{G^2}{N}$	$M_T = \frac{Q_T}{k-1}$	$F_T = \frac{M_T}{M_E} \sim F_{k-1,N-k}$
Within the level of factor (Error)	N-k	Q_E : By subtraction	$M_E = \frac{Q_E}{N-k}$	-
Total	N-1	$Q = \sum_{ij} y_{ij} - \frac{G^2}{N}$	-	-

Variance ratio

The variance ratio is the ratio of the greater variance to the smaller variance. It is also called the F-coefficient. We have

F = Greater variance / Smaller variance.

We refer to the table of F values at a desired level of significance α. In general, α is taken to be 5 %. The table value is referred to as the theoretical value or the expected value. The calculated value is referred to as the observed value.

Inference

If the observed value of F is less than the expected value of F (i.e., $F_o < F_e$) for the given level of significance α, then the null hypothesis H_o is accepted. In this case, we conclude that there is no significant difference between the treatment effects.

On the other hand, if the observed value of F is greater than the expected value of F (i.e., $F_o > F_e$) for the given level of significance α, then the null hypothesis H_o is rejected. In this case, we conclude that all the treatment effects are not equal.

Note: If the calculated value of F and the table value of F are equal, we can try some other value of α.

Problem 1

The following are the details of sales effected by three sales persons in three door-to-door campaigns.

Sales person	Sales in door – to – door campaign			
A	8	9	5	10
B	7	6	6	9
C	6	6	7	5

Construct an ANOVA table and find out whether there is any significant difference in the performance of the sales persons.

Solution:

Method I (Direct method) :

$$\sum A = 8+9+5+10 = 32$$
$$\sum B = 7+6+6+9 = 28$$
$$\sum C = 6+6+7+5 = 24$$

Sample mean for A : $\bar{A} = \frac{32}{4} = 8$

Sample mean for B : $\bar{B} = \frac{28}{4} = 7$

Sample mean for C : $\bar{C} = \frac{24}{4} = 6$

Total number of sample items = No. of items for A + No. of items for B + No. of items for C

$$= 4 + 4 + 4 = 12$$

Mean of all the samples $\bar{X} = \frac{32+28+24}{12} = \frac{84}{12} = 7$

Sum of squares of deviations for A:

A	$A - \bar{A} = A - 8$	$\left(A - \bar{A}\right)^2$
8	0	0
9	1	1
5	-3	9
10	2	4
		14

Sum of squares of deviations for B:

B	$B - \bar{B} = B - 7$	$\left(B - \bar{B}\right)^2$
7	0	0
6	-1	1
6	-1	1

9	2	4
		6

Sum of squares of deviations for C:

C	$C-\bar{C}=C-6$	$\left(C-\bar{C}\right)^2$
6	0	0
6	0	0
7	1	1
5	-1	1
		2

Sum of squares of deviations within

varieties = $\sum\left(A-\bar{A}\right)^2+\sum\left(B-\bar{B}\right)^2+\sum\left(C-\bar{C}\right)^2$

= 14 + 6 + 2

= 22

Sum of squares of deviations for total variance:

Sales person	Sales	Sales - $\bar{X}$ = Sales – 7	$\left(Sales-7\right)^2$

A A A A B B B B C C C C	8 9 5 10 7 6 6 9 6 6 7 5	1 2 - 2 3 0 - 1 - 1 2 - 1 - 1 0 2	1 4 4 9 0 1 1 4 1 1 0 4
			30

ANOVA table

Source of variation	Degrees of freedom	Sum of squares of deviations	Variance
Between varieties	3 – 1 = 2	8	$\frac{8}{2}=4$
Within varieties	12 – 3 = 9	22	$\frac{22}{9}=2.44$
Total	12 – 1 = 11	30	

Calculation of F value:

$$F = \frac{Greater\ Variance}{Smaller\ Variance} = \frac{4.00}{2.44} = 1.6393$$

Degrees of freedom for greater variance $(df_1) = 2$

Degrees of freedom for smaller variance $(df_2) = 9$

Let us take the level of significance as 5%

The table value of F = 4.26

Inference:

The calculated value of F is less than the table value of F. Therefore, the null hypothesis is accepted. It is concluded that there is no significant difference in the performance of the sales persons, at 5% level of significance.

Method II (Short cut method) :

$\sum$ A = 32, $\sum$ B = 28, $\sum$ C = 24.

T= Sum of all the sample items

$$= \sum A + \sum B + \sum C$$
$$= 32 + 28 + 24$$
$$= 84$$

N = Total number of items in all the samples = 4 + 4 + 4 =12

Correction Factor = $\frac{T^2}{N} = \frac{84^2}{12} = 588$

Calculate the sum of squares of the observed values as follows:

Sales Person	X	X^2
A	8	64
A	9	81
A	5	25
A	10	100
B	7	49
B	6	36
B	6	36
B	9	81
C	6	36
C	6	36
C	7	49
C	5	25
		618

Sum of squares of deviations for total variance = $\sum X^2$ - correction factor

= 618 – 588 = 30.

Sum of squares of deviations for variance between samples

$$= \frac{\left(\sum A\right)^2}{N_1} + \frac{\left(\sum B\right)^2}{N_2} + \frac{\left(\sum C\right)^2}{N_3} - CF$$

$$= \frac{32^2}{4} + \frac{28^2}{4} + \frac{24^2}{4} - 588$$

$$= \frac{1024}{4} + \frac{784}{4} + \frac{576}{4} - 588$$

$$= 256 + 196 + 144 - 588$$

$$= 8$$

ANOVA Table

Source of variation	Degrees of Freedom	Sum of squares of deviations	Variance
Between varieties	3-1 = 2	8	$\frac{8}{2} = 4$
Within varieties	12 – 3 = 9	22	$\frac{22}{9} = 2.44$
Total	12 – 1 = 11	30	

It is to be noted that the ANOVA tables in the methods I and II are one and the same. For the further steps of calculation of F value and drawing inference, refer to method I.

Problem 2

The following are the details of plinth areas of ownership apartment flats offered by 3 housing companies A,B,C. Use analysis of variance to determine whether there is any significant difference in the plinth areas of the apartment flats.

Housing Company	Plinth area of apartment flats			
A	1500	1430	1550	1450
B	1450	1550	1600	1480

C	1550	1420	1450	1430

Use analysis of variance to determine whether there is any significant difference in the plinth areas of the apartment's flats.

Note: As the given figures are large, working with them will be difficult.

Therefore, we use the following facts:

i. Variance ratio is independent of the change of origin.

ii. Variance ratio is independent of the change of scale.

In the problem under consideration, the numbers vary from 1420 to 1600. So we follow a method called the **coding method**. First, let us subtract 1400 from each item. We get the following transformed data:

Company	Transformed measurement			
A	100	30	150	50
B	50	150	100	80
C	150	20	50	30

Next, divide each entry by 10.

The transformed data are given below.

Company	Transformed measurement			
A	10	3	15	5
B	5	15	10	8
C	15	2	5	3

We work with these transformed data. We have

$$\sum A = 10+3+15+5=33$$
$$\sum B = 5+15+10+8=38$$
$$\sum C = 15+2+5+3=25$$
$$\sum T = \sum A + \sum B + \sum C$$
$$= 33 + 38 + 25$$
$$= 96$$

N = Total number of items in all the samples = 4 + 4 + 4 = 12

Correction Factor = $\frac{T^2}{N} = \frac{96^2}{12} = 768$

Calculate the sum of squares of the observed values as follows:

Company	X	X^2
A	10	100
A	3	9
A	15	225
A	5	25
B	5	25
B	15	225
B	10	100
B	8	64
C	15	225
C	2	4
C	5	25
C	3	9
		1036

Sum of squares of deviations for total variance = $\sum X^2$ - correction factor

= 1036 – 768 = 268

Sum of squares of deviations for variance between samples

$$= \frac{\left(\sum A\right)^2}{N_1} + \frac{\left(\sum B\right)^2}{N_2} + \frac{\left(\sum C\right)^2}{N_3} - CF$$

$$= \frac{33^2}{4} + \frac{38^2}{4} + \frac{25^2}{4} - 768$$

$$= \frac{1089}{4} + \frac{1444}{4} + \frac{625}{4} - 768$$

$$= 272.25 + 361 + 156.25 - 768$$

$$= 789.5 - 768$$

$$= 21.5$$

ANOVA Table

Source of variation	Degrees of Freedom	Sum of squares of deviations	Variance
Between varieties	3-1 = 2	21.5	$\frac{21.5}{2} = 10.75$
Within varieties	12 – 3 = 9	264.5	$\frac{24.65}{9} = 27.38$
Total	12 – 1 = 11	268	

Calculation of F value:

$$\text{F} = \frac{\textit{Greater Variance}}{\textit{Smaller Variance}} = \frac{27.38}{10.75} = 2.5470$$

Degrees of freedom for greater variance $\left(df_1\right) = 9$

Degrees of freedom for smaller variance $\left(df_2\right) = 2$

The table value of F at 5% level of significance = 19.38

Inference:

Since the calculated value of F is less than the table value of F, the null hypothesis is accepted and it is concluded that there is no significant difference in the plinth areas of ownership apartment flats offered by the three companies, at 5% level of significance.

Problem 3

A finance manager has collected the following information on the performance of three financial schemes.

Source of variation	Degrees of Freedom	Sum of squares of deviations
Treatments	5	15
Residual	2	25
Total (corrected)	7	40

Interpret the information obtained by him.

Note: 'Treatments' means 'Between varieties'.

'Residual' means 'Within varieties' or 'Error'.

Solution:

Number of schemes = 3 (since 3 – 1 = 2)

Total number of sample items = 8 (since 8 – 1 = 7)

Let us calculate the variance.

Variance between varieties = $\frac{15}{2} = 7.5$

Variance between varieties = $\frac{25}{5} = 5$

$$F = \frac{Greater\ Variance}{Smaller\ Variance} = \frac{7.5}{5} = 1.5$$

Degrees of freedom for greater variance (df_1) = 2

Degrees of freedom for smaller variance $(df_2) = 5$

The total value of F at 5% level of significance = 5.79

Inference:

Since the calculated value of F is less than the table value of F, we accept the null-hypothesis and conclude that there is no significant difference in the performance of the three financial schemes.

QUESTIONS

1. Define analysis of variance.
2. State the assumptions in analysis of variance.
3. Explain the classification of linear models for the sample data.
4. Explain ANOVA table.
5. Explain how inference is drawn from ANOVA table.
6. Explain the managerial application of analysis of variance.

LESSON 3 DESIGN OF EXPERIMENTS

LESSON OUTLINE

- Definition of design of experiments
- Key concepts in the design of experiments
- Steps in the design of experiments
- Replication, Randomization and Blocking
- Lay out of an experimental design
- Data Allocation Table
- Completely Randomized Design
- ANOVA table for CRD
- Working rule for an example
- Randomized Block Design
- ANOVA table for RBD
- Latin Square Design
- ANOVA table for LSD
- Managerial applications of experimental designs

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the definition of design of experiments
- understand the key concepts in the design of experiments
- understand the steps in the design of experiments
- understand the lay out of an experimental design
- understand a data allocation table
- construct ANOVA table for CRD
- draw inference from ANOVA table for CRD
- construct ANOVA table for RBD
- draw inference from ANOVA table for RBD
- construct ANOVA table for LSD
- draw inference from ANOVA table for LS
- understand the working rules for solving problems
- understand the managerial applications of experimental designs

DESIGN OF EXPERIMENTS

I. FUNDAMENTALS OF DESIGNS

Introduction

The theory of design of experiments was originally developed for agriculture. For example, to determine which fertilizer would give more yield of a certain crop, from among a set of fertilizers. Nowadays the design of experiments finds its application in the area of management also.

While carrying out research for managerial decision making, one may go for descriptive research or experimental research. The advantage of experimental research is that it can be used to establish the cause-effect relationship between the variables under consideration. Such a relationship is called a **causal relationship**. An experiment may be carried out with a control group or without a control group, depending on the resources available and the nature of the subjects involved in the experiment. The researcher has to select different subjects, put them into several groups and administer treatments to the subjects within each group. It would be advisable to include a control group wherever possible so as to increase the level of validity of the inference drawn from the experiment.

Definition of design of experiments

The design of experiments is the logical construction of the experiment with a well-defined level of uncertainty involved in the inference drawn.

Key concepts in the design of experiments

The design of experiments centers around the following three key concepts:
(1) Treatments
(2) Factors
(3) Levels of a treatment factor

Types of experiments

There are two types of experiments, namely absolute experiment and comparative experiment. In an absolute experiment, one takes into account the absolute value of a certain characteristic. As distinct from this, a comparative experiment seeks to compare the effect of two or more objects on some characteristic of the population under examination. For example, one may think of the following situations:

* comparison of the effect of different fertilizers on a certain crop
* comparison of the effect of different medicines on a disease
* comparison of different marketing strategies for the promotion of a product
* comparison of different machines in the production of a certain product
* comparison of different methods of resource mobilization

Steps in the Design of Experiments

The design of experiments consists of the following steps:

1. Statement of the objectives
2. Formulation of the statistical hypotheses
3. Choice of the treatments
4. Choice of the experimental sites
5. Replication and levels of variation
6. Choice of the experimental blocks, if necessary
7. Characteristics of the plots undertaken for the experiments
8. Assignment of treatments to various units
9. Recording of data
10. Statistical analysis of data

Basic designs

The following are the basic designs in statistical analysis:

1. Completely Randomized Design (CRD)
2. Randomized Block Design (RBD)
3. Latin Square Design (LSD)

Other designs also can be used for drawing inferences from experiments. However, they are quite complex and we shall confine ourselves to the above three designs.

Basic principles

The design of experiments is mainly based on the following three basic principles:

1. Replication
2. Randomization
3. Blocking or Local Control.

Replication means the repetition of each treatment a certain number of times. This will help in reducing the effect due to a possible extreme situation (outlier) arising out of a single treatment. Thus replication will reduce the experimental error. Homogeneity is possible only within a replication.

Randomization means allocation of the treatments to different units in a random way. i.e., all the units will have equal chance of allotment of treatments. But, what treatment is actually allotted to a unit will depend on pure chance only.

The basic design is Completely Randomized Design (CRD). In this design, the first two principles namely replication and randomization are used. There is no necessity of blocking in CRD, because the entire area of experiment is assumed to be homogeneous. If it is not so, then it becomes necessary to subdivide the non-homogeneous experimental area into homogeneous sub-groups such that each subgroup has almost the same level of attribute. The technique of subdividing the experimental area into groups is called as **blocking or local control** and such subgroups are called as **Blocks**.

The RBD and LSD are bock designs. However, it should be remembered that CRD is not a bock design.

II. Completely Randomized Design (CRD)

This design is useful to compare several treatments in an experiment. For example, suppose that three are three training institutes each offering a distinct training programme to sales persons and a manager wants to know which of the three training programmes would be highly rewarding for his business organization. One option for him would be the comparison of the means of the samples taken two at a time. However, comparison of the sample means may not yield accurate results when more than two samples are involved in the experiment. Because of this reason, the manager may opt for a completely randomized design. In this design, all the samples are taken for simultaneous consideration and they are examined by means of a single statistical test.

For the application of this design, the first and foremost condition is that the experimental area should be homogeneous in the particular attribute about which the experiment is carried out. For the purpose of illustration, we consider an example with 3 treatments denoted by A, B, C. A **lay out** is a pictorial representation of assignment of treatments to various experimental areas. The example design has the following lay out.

Experimental area

B	A	B
A	A	C
C	B	A

Data on treatments

Suppose there are 3 treatments A, B, C and each treatment is used a certain number of times as illustrated in the following example:

TREATMENT	NO. OF TIMES THE TREATMENT IS APPLIED
A	4
B	3
C	2

Collect the results on the data arising out of the application of these treatments. Suppose the results on the attribute pertaining to treatment A are 38, 36, 35 and 40. Suppose the results pertaining to treatment B are 26, 30 and 28. Suppose the results pertaining to treatment C are 30 and 28. Using these values, a '**Data Allocation Table**' is constructed as follows:

Treatment Data Allocation

A	38	36	35	30

B	26	30	28

C	30	28

The sum of the values for the 3 treatments are denoted by T_1, T_2 and T_3, respectively. For the above example data, we obtain

$T_1 = 38 + 36 + 35 + 30 = 139$,
$T_2 = 26 + 30 + 28 = 84$ and
$T_3 = 30 + 28 = 58$.

Statistical Analysis of CRD

As already mentioned, the experimental units in a CRD are taken in a single group with the condition that the units forming the group must be homogeneous as far as possible. Suppose there are k treatments in an experiment. Let the i^{th} treatment be replicated n_i times. Then the total number of experimental units in the design is $n_1 + n_2 + ... + n_i + ... + n_k = \sum_{i=1}^{k} n_i = N$.

The treatments are allocated at random to all the units in the experimental area. This design provides a one-way classified data with different levels of a single factor called treatments. The linear model for CRD is defined by the relation

$$y_{ij} = \mu + a_i + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

where y_{ij} is the jth observation of the ith treatment,

μ is the general mean effect which is fixed,

a_i is the fixed effect due to ith treatment and

e_{ij} is the random error effect which is distributed normally with zero mean and constant variance.

Let $\sum_{ij} y_{ij} = G$ be the Grand total of all the observations.

In $\sum y_{ij}$, fix *i* and vary *j*. Then the sum gives the ith treatment total, denoted by T_i. i.e., $\sum_j y_{ij} = T_i$ *(i=1,2,...,k).*

Apply the ANOVA for one-way classified data and compute the total sum of squares (TSS) and treatment sum of squares (T_rSS) as follows:

$$TSS = \sum_{ij} y_{ij}^2 - \frac{G^2}{N} = Q$$

$$TrSS = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N} = Q_T$$

G^2/N is called the correction factor or the correction term.

The error sum of squares (ESS) can be obtained by subtraction. All these values are represented in the form of an ANOVA table provided below.

ANOVA Table for CRD

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F
Treatments	k– 1	$Q_T = \sum_i \frac{T_i}{N_i} - \frac{G^2}{N}$	$M_T = \frac{Q_T}{k-1}$	$F_T = \frac{M_T}{M_E} \sim F_{k-1,N-k}$
Error	N– k	Q_E : By subtraction	$M_E = \frac{Q_E}{N-k}$	-

Total	N– 1	$Q = \sum_{ij} y_{ij}^2 - \frac{G^2}{N}$	-	-

Application of ANOVA:

Objective of ANOVA: We apply ANOVA to find out whether there is any significant difference in the performance of the treatments. We formulate the following **null hypothesis:**

H_0: There is no significant difference in the performance of the treatments.

The null hypothesis has to be tested against the following **alternative hypothesis**:

H_1: There is a significant difference in the performance of the treatments.

We have to decide whether the null hypothesis has to be accepted or rejected at a desired level of significance (α).

Inference

If the observed value of F is less than the expected value of F, i.e., $F_o < F_e$, then the null-hypothesis H_o is accepted for a given level of significance (α) and we conclude that the effects due to various treatments do not differ significantly.

If the observed value of F is greater than the expected value of F, i.e., $F_o > F$, then the null-hypothesis H_o is rejected for a given level of significance (α) and we conclude that the effects due to various treatments differ significantly.

Working rule for an example:

We have to consider three quantities G, N and the Correction Factor (denoted by CF) defined as follows:

G = Sum of the values for all the treatments,

N = The sum of the number of times each treatment is applied

The correction factor $CF = G^2 / N$.

Let us consider an example of CRD. Suppose there are 3 treatments A, B, C. Suppose the number of times the treatment is applied is n_1 in the case of A, n_2 for B and n_3 for C. The sum of the values for the 3 treatments are denoted by T_1, T_2 and T_3. With these notations, we have

$N = n_1 + n_2 + n_3$,
$G = T_1 + T_2 + T_3$,
$CF = G^2/N = (T_1 + T_2 + T_3)^2 / (n_1 + n_2 + n_3)$.

Define the following quantities:

TSS = Sum of the squares of the observed values – Correction Factor
T_r SS = $(T_1^2 / n_1 + T_2^2 / n_2 + T_3^2 / n_3)$ – Correction Factor
ESS = TSS – T_r SS
Calculation of the Degrees of Freedom (df):
The df for treatments = No. of treatments – 1.
The df for the total = Total no. of times all the treatments have been applied – 1 = $N - 1 = n_1 + n_2 + n_3 - 1$.
The df for the Error = (Total no. of times all the treatments have been applied - No. of treatments) – 2.
We have the following ANOVA table for this example.

ANOVA Table for CRD

Source of variation	Degrees of freedom	SS	MSS	Variance ratio F
Treatment	3– 1 = 2	T_r SS	T_r SS / df = T_r SS / 2	
Error	8– 2 = 6	ESS	ESS / df = ESS / 6	
Total	9– 1 = 8	TSS		

After these steps, carry out the Analysis of Variance and draw the inference.

Problem 1

Examine the CRD with the following Data Allocation Table and determine whether or not the treatments differ significantly.

Treatment Data Allocation

A	28	36	32	34

B	40	38	36

C	32	34

Solution:
The treatments in the design are A, B and C.
We have
n_1 = The number of times A is applied = 4,
n_2 = The number of times B is applied = 3,
n_3 = The number of times C is applied = 2.
$N = n_1 + n_2 + n_3 = 4 + 3 + 2 = 9$.
The sum of the values for the 3 treatments are denoted by T_1, T_2 and T_3, respectively.
For the given data on experimental values, we obtain
$T_1 = 28 + 36 + 32 + 34 = 130$,
$T_2 = 40 + 38 + 36 = 114$ and
$T_3 = 32 + 34 = 66$.
$G = T_1 + T_2 + T_3 = 130 + 114 + 66 = 310$.
The correction factor $= G^2/N = 310^2/9 = 10677.8$
$\sum y^2_{ij} = 28^2 + 36^2 + 32^2 + 34^2 + 40^2 + 38^2 + 36^2 + 32^2 + 34^2$
$= 784 + 1296 + 1024 + 1156 + 1600 + 1444 + 1296 + 1024 + 1156$
$= 10780$
$\sum (T^2_i / n_i) = 130^2 / 4 + 114^2 / 3 + 66^2 / 2$
$= 16900 / 4 + 12996 / 3 + 4356 / 2 = 4225 + 4332 + 2178 = 10735$
The total sum of squares (TSS) and treatment sum of squares (T_rSS) are calculated as follows:
$TSS = \sum y^2_{ij} - CF = 10780 - 10677.8 = 102.2$
$T_rSS = \sum T^2_i / n_i - CF = 10735 - 10677.8 = 57.2$
$ESS = TSS - T_rSS$

We apply ANOVA to find out whether there is any significant difference in the performance of the treatments. We formulate the following **null hypothesis**:

H_0: There is no significant difference in the performance of the treatments.

The null hypothesis has to be tested against the following **alternative hypothesis**:

H_1: There is a significant difference in the performance of the treatments.

We have to decide whether the null hypothesis has to be accepted or rejected at a desired level of significance (α).

ANOVA Table for CRD

Source of variation	Degrees of freedom	SS	MSS = SS/DF	Variance ratio F
Treatment	3– 1 = 2	57.2	57.2 / 2 = 28.6	28.6 / 7.5 = 3.81
Error	8– 2 = 6	45.0	45 / 6 = 7.5	
Total	9– 1 = 8	102.2		

In the table, first enter the values of SS for 'Total' and 'Treatment'. From Total, subtract Treatment to obtain SS for 'Error'.
i.e., ESS = TSS – T_rSS = 102.2 – 57.2 = 45.0

Calculation of F value: F = Greater variance / Smaller variance = 28.6 / 7.5
= 3.81

Degrees of freedom for greater variance (df_1) = 2
Degrees of freedom for smaller variance (df_2) = 6
Table value of F at 5% level of significance = 5.14

Inference:
Since the calculated value of F is less than the table value of F, the null hypothesis is accepted and it is concluded that there is no significant difference in the treatments A, B and C, at 5% level of significance.

III. Randomized Block Design (RBD)

In CRD, note that the site is not split into blocks. An improvement of CRD can be obtained by providing the blocking (local control) measure in the experimental design. One such design is Randomized Block Design (RBD). In a block design, the site is split into different blocks such that each block is

homogeneous in itself, with respect to the particular attribute under experiment. The result from a RBD will be better than that from a CRD. While we use one-way ANOVA in CRD, we use two-way ANOVA in RBD.

Example of lay out of RBD:
Experimental area

Treatment	Block 1	Block 2	Block 3
A	19	16	17
B	16	17	20
C	23	24	22

This is an example of a RBD with 3 treatments and 3 blocks.

Statistical Analysis of RBD

Suppose there are k treatments each replicated r times. Then the total number of experimental units is rk. These units are rearranged into r groups (Blocks) of size k. The local control measure is adopted in this design in order to make the units of each group to be homogeneous. The group units in these bocks are known as plots or cells. The k treatments are allocated at random in the k plots of each of the blocks selected randomly one by one. This type of homogeneous grouping of experimental units and random allocation of treatments to randomly selected blocks are two main features of RBD.

The technique of ANOVA for two-way classified data is applicable to an experiment with RBD lay out. The data collected from the experiment is classified according to the levels of two factors namely treatments and blocks. The linear model for RBD is defined by the relation

$$y_{ij} = \mu + a_i + b_j + e_{ij} \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,r \end{pmatrix}$$

where y_{ij} is the observation corresponding to i^{th} treatment and j^{th} block,

μ is the general mean effect which is fixed,

a_i is the fixed effect due to i^{th} treatment,

b_j is the fixed effect due to j[th] block and

e_{ij} is the random error effect which is distributed normally with zero mean and constant variance.

Applying the method of ANOVA for two-way classified data, the sum of squares due to treatments, blocks and error can be obtained.

Let $\sum_{ij} y_{ij} = G$ be the Grand total of all the rk observations.

In $\sum y_{ij}$, fix *i* and vary *j*. Then the sum gives the i[th] treatment total, denoted by T_i. i.e., $\sum_j y_{ij} = T_i$ *(i=1,2,...,k).*

In $\sum y_{ij}$, fix *j* and vary *i*. Then the sum gives the j[th] block total, denoted by Bj. i.e., $\sum_j y_{ij} = Bj$ *(j=1,2,...,r).*

We take $\frac{G^2}{rk}$ as the correction factor. The number of treatments is k and the number of blocks is r. Various sums of squares are computed as follows.

$$TSS = \sum_{ij} y_{ij}^2 - \frac{G^2}{rk} = Q$$

$$TrSS = \sum_i \frac{T_i^2}{r} - \frac{G^2}{rk} = Q_T,$$

$$BSS = \sum_j \frac{B_j^2}{k} - \frac{G^2}{rk} = Q_B,$$

$$ESS = Q - Q_T - Q_B = Q_E$$

All these values are represented in the form of an ANOVA table provided below.

ANOVA Table for RBD

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F

Treatments	k – 1	$Q_T = \sum_i \frac{T_i^2}{r} - \frac{G^2}{rk}$	$M_T = \frac{Q_T}{k-1}$	$F_T = \frac{M_T}{M_E} \sim F_{k-1,(k-1)(r-1)}$
Blocks	r – 1	$Q_B = \sum_j \frac{B_j^2}{k} - \frac{G^2}{rk}$	$M_B = \frac{Q_B}{r-1}$	$F_B = \frac{M_B}{M_E} \sim F_{r-1,(k-1)(r-1)}$
Error	(k – 1)(r – 1)	Q_E : By subtraction	$M_E = \frac{Q_E}{(k-1)(r-1)}$	
Total	(rk – 1)	$Q = \sum_{ij} y_{ij}^2 - \frac{G^2}{rk}$		

We have to find out whether there is any significant difference in the performance of the treatments. Also we can determine whether there is any significant difference in the performance of different blocks. We formulate the following two null hypotheses:

Null hypothesis-1

H_{01}: There is no significant difference in the performance of the treatments.

Null hypothesis-2

H_{02}: There is no significant difference in the performance of the blocks.

Each null hypotheses has to be tested against the alternative hypothesis. Even though there are two null hypotheses, the important one is the null hypothesis on the treatments. We have to decide whether to accept or reject the null hypothesis on the treatments at a desired level of significance (α).

Inference

If the observed value of F is less than the expected value of F, i.e., $F_o < F_e$, then the null-hypothesis H_o is accepted for a given level of significance (α) and we conclude that the effects due to various treatments do not differ significantly.

If the observed value of F is greater than the expected value of F, i.e., $F_o > F$ then the null-hypothesis H_o is rejected for a given level of significance (α) and we conclude that the effects due to various treatments differ significantly.

Similarly, the blocks' effects may also be tested, if necessary.

Working rule for an example:

Consider the following example:

Treatment	Block 1	Block 2	Block 3	Block 4
A	72	68	70	56
B	55	60	62	55
C	65	70	70	60

In this case, we have
$T_1 = 72 + 68 + 70 + 56 = 266$,
$T_2 = 55 + 60 + 62 + 55 = 232$,
$T_3 = 65 + 70 + 70 + 60 = 265$,
$T_1 + T_2 + T_3 = 266 + 232 + 265 = 763$.
$B_1 = 72 + 55 + 65 = 192$,
$B_2 = 68 + 60 + 70 = 198$,
$B_3 = 70 + 62 + 70 = 202$,
$B_4 = 56 + 55 + 60 = 171$,
$B_1 + B_2 + B_3 + B_4 = 192 + 198 + 202 + 171 = 763$.
For easy reference, let us take the number of treatments as t and the number of blocks as b. Then we have t = 3 and b = 4.
Calculate T_r SS and BSS as follows:
T_r SS = $(T_1^2 / b + T_2^2 / b + T_3^2 / b + T_4^2 / b)$ – Correction Factor
BSS = $(B_1^2 / t + B_2^2 / t + B_3^2 / + B_3^2 / t)$ – Correction Factor
After these steps, carry out the Analysis of Variance and draw the inference.

Problem 2

Analyse the following RBD and determine whether or not the treatments differ significantly.

Experimental area

Treatment	Block 1	Block 2	Block 3
A	9	5	7

B	6	8	5
C	4	5	8

Solution:

The treatments in the design are A, B and C. There are 3 blocks namely, Block 1, Block 2 and Block 3.
We have
n_1 = The number of times A is applied = 3,
n_2 = The number of times B is applied = 3,
n_3 = The number of times C is applied = 3.
$N = n_1 + n_2 + n_3 = 3 + 3 + 3 = 9$.

The sum of the values for the 3 treatments are denoted by T_1, T_2 and T_3, respectively.

For the given data on experimental values, we obtain
$T_1 = 9 + 5 + 7 = 21$,
$T_2 = 6 + 8 + 5 = 19$,
$T_3 = 4 + 5 + 8 = 17$,
$T_1 + T_2 + T_3 = 21 + 19 + 17 = 57$.
$B_1 = 9 + 6 + 4 = 19$,
$B_2 = 5 + 8 + 5 = 18$,
$B_3 = 7 + 5 + 8 = 20$,
$B_1 + B_2 + B_3 = 19 + 18 + 20 = 57$.
$G = T_1 + T_2 + T_3 = 57$.
The correction factor $= G^2/N = 57^2 / 9 = 3249 / 9 = 361$
$\sum y^2_{ij} = 9^2 + 5^2 + 7^2 + 6^2 + 8^2 + 5^2 + 4^2 + 5^2 + 8^2$
$= 81 + 25 + 49 + 36 + 64 + 25 + 16 + 25 + 64 = 385$
No. of blocks = b = 3
No. of treatments = t = 3
$\sum (T^2_i / b) = 21^2 / 3 + 19^2 / 3 + 17^2 / 3$
$= 441 / 3 + 361 / 3 + 289 / 3 = 147 + 120.3 + 96.3 = 363.6$
$\sum (B^2_j / t) = 19^2 / 3 + 18^2 / 3 + 20^2 / 3$
$= 361 / 3 + 324 / 3 + 400 / 3 = 120.3 + 108 + 13.3 = 361.6$
The total sum of squares (TSS), treatment sum of squares (T_rSS) and block sum of squares (BSS) are calculated as follows:
$TSS = \sum y^2_{ij} - CF = 385 - 361 = 24$
$T_rSS = \sum (T^2_i / b) - CF = 363.6 - 361 = 2.6$

$BSS = \sum (B^2_j/t) - CF = 361.6 - 361 = 0.6$

$ESS = TSS - T_rSS - BSS = 24 - 2.6 - 0.6 = 24 - 3.2 = 20.8$

We apply ANOVA to find out whether there is any significant difference in the performance of the treatments. We formulate the following **null hypothesis**:

H_0: There is no significant difference in the performance of the treatments.

The null hypothesis has to be tested against the following **alternative hypothesis**:

H_1: There is a significant difference in the performance of the treatments.

We have to decide whether the null hypothesis has to be accepted or rejected at a desired level of significance (α).

ANOVA Table for RBD

Source of variation	Degrees of freedom	SS	MSS = SS/DF	Variance ratio F
Treatment	3– 1 = 2	2.6	2.6 / 2 = 1.3	5.2 / 1.3 = 4.0
Block	3– 1 = 2	0.6	0.6 / 2 = 0.3	5.2 / 0.3 = 17.3
Error	8– 4 = 4	20.8	20.8 / 4 = 5.2	
Total	9– 1 = 8	24.0		

In the table, first enter the values of SS for 'Total' , 'Treatment' and 'Block'. From Total, subtract (Treatment + Block) to obtain SS for 'Error'.

i.e., ESS = 24.0 - 3.2 = 20.8
Calculation of F value: We consider '**Treatment**'.
F = Greater variance / Smaller variance = 5.2 / 1.3 = 4
Degrees of freedom for greater variance (df_1) = 4
Degrees of freedom for smaller variance (df_2) = 2
Table value of F at 5% level of significance = 19.25
Inference:

Since the calculated value of F for the treatments is less than the table value of F, the null hypothesis is accepted and it is concluded that there is no significant difference in the treatments A, B and C at 5% level of significance.

Note: If required, by using the same table, we can also test whether there is any significant difference in the blocks, at 5% level of significance.

IV Latin Square Design (LSD)

It was pointed out earlier that RBD is an improvement of CRD, since RBD provides an error control measure for the elimination of block variation.

In RBD, the source of variation is eliminated in only one direction, namely block wise. This idea can be further generalized to improve RBD by eliminating more sources of variation. One such design with a provision for elimination of two sources of variation is 'Latin Square Design'. The result from an LSD will be better than that from a RBD.

Suppose there are n treatments each replicated n times. Then the total number of experimental units is $n \times n = n^2$. Let $p \times q$ denote the factors whose variations are to be eliminated from the experimental error. Then both the factors P and Q should be related to the variable under study. In that case, these two factors are control factors of variation.

Therefore the total number of level combinations of the two factors is $n \times n = n^2$. Now the experimental units are so chosen that each unit contains different level combinations of these two factors. Further the n^2 experimental units are arranged in the form of an $n \times n$ array so that there are n rows and n columns of the n^2 units. Then each unit belongs to different row-column combination. i.e., The two factors P and Q become the rows and columns of the design.

Though it is not necessary that the two factors P and Q should always be called as rows and columns, it has become a convention to define an LSD by means of two factors, namely rows and columns.

After the experimental units are obtained, the n treatments are allocated to the n^2 units such that each treatment occurs once and only once in each row and each column. This ensures that each treatment is replicated n times. If a two-way table is formed with the levels of the factor P (rows) and the levels of the factor Q (columns), then the n treatments should be allocated to the n^2 units such that each treatment occurs once and only once in each level of the factor P and each level of the factor Q. Such an arrangement is called a Latin Square Design of order $n \times n$.

Example of lay out of LSD

Example 1:

Experimental area

A	B	C
B	C	A
C	A	B

In this design, the first row consists of the experiments A, B, C, in this order. The second row is got by a cyclic permutation of the first row elements. The third row is got by a cyclic permutation of the second row elements.

Example 2:

Experimental area

A	B	C
C	A	B
B	C	A

In this design, the first row consists of the experiments A, B, C in this order. The third row is got by a cyclic permutation of the first row elements. The second row is got by a cyclic permutation of the third row elements.

Example 3:

Experimental area

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

In this design, the first row consists of the experiments A, B, C, D in this order. The second row is got by a cyclic permutation of the first row elements. The third row is got by a cyclic permutation of the second row elements. The fourth row is got by a cyclic permutation of the third row elements.

Example 4:

Suppose there are 5 treatments denoted by A, B, C, D, E. Then the following arrangement of the treatments is a Latin Square Design of order 5×5.

Column / **Row**		Factor Q (Column)				
		Q_1	Q_2	Q_3	Q_4	Q_5
Factor P	P_1	A	B	C	D	E
	P_2	B	C	D	E	A
	P_3	C	D	E	A	B
	P_4	D	E	A	B	C
	P_5	E	A	B	C	D

Note that every treatment appears in each row and column exactly once.

In the lay out of an LSD, apart from indicating the treatment, the experimental value also has to be mentioned in each cell.

Statistical Analysis of LSD

In an LSD, we have to consider there factors namely rows, columns and treatments. Therefore, the data collected from this design must be analyzed as a three-way classified data. For this purpose actually there must be n^3 observations, since there are three factors each with n-levels. However, because of the particular allocation of the treatment to each cell, there is only one observation per cell, instead of n-observations per cell, according to a three-way classified data. Consequently, there is no interaction between any of the factors namely rows, columns and treatments. Hence the appropriate linear model for LSD is defined by the relation

$$y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk} \quad (i, j, k = 1, 2, ..., n)$$

where y_{ijk} is the general observation corresponding to ith row, jth column and kth treatment,

μ is the general mean effect which is fixed,

r_i is the fixed effect due to ith row,

c_j is the fixed effect due to jth column,

t_k is the fixed effect due to kth treatment and

e_{ijk} is the random error effect which is distributed normally with zero mean and constant variance.

Application of ANOVA:

The analysis here is similar to the analysis of two-way classified data.

First of all, the data is arranged in a row-column table. Let y_{ij} denote the observation corresponding to ith row and jth column in the table.

In $\sum y_{ij}$, fix *i* and vary *j*. Then the sum gives the i^{th} row total, denoted by Ri .

i.e., $\sum_j y_{ij} = R_i$ *(i=1,2,...,n).*

In $\sum y_{ij}$, fix j and vary *i.* Then the sum gives the j^{th} column total, denoted by

Cj . i.e., $\sum_j y_{ij} = Cj$ *(j=1,2,...n).*

Let

T_k = k^{th} treatment total (k=1,2,…,n).

We have $\sum_i R_i = \sum_j C_j = \sum_k T_k = G$ which is the Grand total of all the n^2 observations. The correction factor CF is defined by $CF = \frac{G^2}{N}$ where $N = n^2$ is the total number of observations. We have

$$\sum_{ij} y_{ij} = G .$$

Various sums of squares are computed through the CF as follows:

$$TSS = \sum_{ij} y_{ij}^2 - \frac{G^2}{n^2} \quad which\ has\ (n^2 - 1)\ dF$$

$$RSS = \sum_i \frac{R_i^2}{n} - \frac{G^2}{n^2} \quad which\ has\ (n-1)\ dF$$

$$CSS = \sum_j \frac{C_j^2}{n} - \frac{G^2}{n^2} \quad which\ has\ (n-1)\ dF$$

$$TSS = \sum_k \frac{T_k^2}{n} - \frac{G^2}{n^2} \quad which\ has\ (n-1)\ dF$$

$$ESS = TSS - RSS - CSS - TrSS$$

which has *(n-1)(n-2) dF.*

All these values are represented in the form of an ANOVA table below.

ANOVA Table for $n \times n$ Latin Square Design

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F
Rows	(n-1)	$Q_R = \sum_i \frac{R_i^2}{n} - \frac{G^2}{n^2}$	$M_R = \frac{Q_R}{n-1}$	$F_R = \frac{M_R}{M_E} \sim F_{n-1,(n-1)(n-2)}$
Columns	(n-1)	$Q_C = \sum_j \frac{C_j^2}{n} - \frac{G^2}{n^2}$	$M_c = \frac{Q_c}{n-1}$	$F_C = \frac{M_C}{M_E} \sim F_{n-1,(n-1)(n-2)}$
Treatments	(n-1)	$Q_T = \sum_k \frac{T_k^2}{n} - \frac{G^2}{n^2}$	$M_T = \frac{Q_T}{n-1}$	$F_T = \frac{M_T}{M_E} \sim F_{n-1,(n-1)(n-2)}$
Error	(n-1) (n-2)	Q_E : By subtraction	$M_E = \frac{Q_E}{(n-1)(n-2)}$	
Total	(n^2-1)	$Q = \sum_{ij} y^2{}_{ij} - \frac{G^2}{n^2}$		

The following hypotheses are formed:

Null hypothesis-1

H_{01}: There is no significant difference in the performance of the treatments.

Null hypothesis-2

H_{02}: There is no significant difference in the performance of the rows.

Null hypothesis-3

H_{03}: There is no significant difference in the performance of the columns.

Each null hypothesis has to be tested against the alternative hypothesis.

Even though there are three null hypotheses, the important one is the null hypothesis

on the treatments. We have to decide whether to accept or reject the null hypothesis on the treatments at a desired level of significance (α).

Inference

If the observed value of *F* is less than the expected value of F, i.e., $F_o < F_e$, for a given level of significance α, then the null hypothesis of equal treatment effect is accepted. Otherwise, it is rejected.

Problem 3

Examine the following experimental values on the output due to four different training methods A, B, C and D for sales persons and find out whether there is any significant difference in the training methods.

A 28	B 20	C 32	D 28
B 36	C 30	D 28	A 20
C 25	D 30	A 22	B 35
D 30	A 26	B 36	C 28

Solution :

In this design, there are 4 treatments A, B, C and D. In the lay out of the design, each treatment appears exactly once in each row as well as each column. Therefore this design is an LSD. The name of the treatment and the observed value under that treatment are specified together in each cell.

R_1 = $\sum$ first row elements = 28 + 20 + 32 + 28 = 108
R_2 = $\sum$ second row elements = 36 + 30 + 28 + 20 = 114
R_3 = $\sum$ third row elements = 25 + 30 + 22 + 35 = 112
R_4 = $\sum$ fourth row elements = 30 + 26 + 36 + 28 = 120
C_1 = $\sum$ first column elements = 28 + 36 + 25 + 30 = 119
C_2 = $\sum$ second column elements = 20 + 30 + 30 + 26 = 106
C_3 = $\sum$ third column elements = 32 + 28 + 22 + 36 = 118
C_4 = $\sum$ fourth column elements = 28 + 20 + 35 + 28 = 111

From the given table, rewrite the experimental values for each treatment separately as follows:

Treatment			
A	B	C	D
28	20	32	28
20	36	30	28
22	35	25	30
26	36	28	30

$T_1 = \sum A = 28 + 20 + 22 + 26 = 96$
$T_2 = \sum B = 20 + 36 + 35 + 36 = 127$
$T_3 = \sum C = 32 + 30 + 25 + 28 = 115$
$T_4 = \sum D = 28 + 28 + 30 + 30 = 116$
$G = T_1 + T_2 + T_3 + T_3 = 96 + 127 + 115 + 116 = 454$
n = No. of treatments = 4
$N = n^2 = 16$
Correction Factor = $G^2/N = 454^2 / 16 = 206116 / 16 = 12882.25$
The total sum of squares (TSS), Row sum of squares (RSS), Column sum of squares (CSS) and Treatment sum of squares (T_rSS) are calculated as follows:

$TSS = \sum y^2_{ij} -$ Correction Factor
$RSS = \sum (R_i^2 / n) -$ Correction Factor
$CSS = \sum (C_j^2 / n_j) -$ Correction Factor
$T_rSS = \sum (T^2_k /n) -$ Correction Factor
$\sum y^2_{ij} = 28^2 + 20^2 + 32^2 + 28^2 + 36^2 + 30^2 + 28^2 + 20^2 + 25^2 + 30^2 + 22^2 + 35^2 + 30^2 + 26^2 + 36^2 + 28^2$

$= 784 + 400 + 1024 + 784 + 1296 + 900 + 784 + 400 + 625 + 900 + 484 + 1225 + 900 + 676 + 1296 + 784$
$= 13262$
$TSS = \sum y^2_{ij} - CF = 13262 - 12882.25 = 379.75$

$RSS = R_1^2 / 4 + R_2^2 / 4 + R_3^2 / 4 + R_4^2 / 4 - CF$
$= 108^2 / 4 + 114^2 / 4 + 112^2 / 4 + 120^2 / 4 - 12882.25$
$= 11664 / 4 + 12996 / 4 + 12544 / 4 + 14400 / 4 - 12882.25$

$= 2916 + 3249 + 3136 + 3600 – 12882.25 = 12901 – 12882.25 = 18.75$

$CSS = C_1^2 / 4 + C_2^2 / 4 + C_3^2 / 4 + C_4^2 / 4 – CF$
$= 119^2 / 4 + 106^2 / 4 + 118^2 / 4 + 111^2 / 4 – 12882.25$
$= 14161 / 4 + 11236 / 4 + 13924 / 4 + 12321/ 4 – 12882.25$
$= 3540.25 + 2809 + 3481 + 3080.25 – 12882.25 = 12910.5 –12882.25$
$= 28.25$

$T_rSS = T_1^2 / 4 + T_2^2 / 4 + T_3^2 / 4 + T_4^2 / 4 – CF$
$= 96^2 / 4 + 127^2 / 4 + 115^2 / 4 + 116^2 / 4 – 12882.25$
$= 9216 / 4 + 16129 / 4 + 13225 / 4 +13456/ 4 – 12882.25$
$= 2304 + 4032.25+ 3306.25 + 3364 – 12882.25 = 13006.5 – 12882.25$
$= 124.25$
ESS = Error sum of squares = TSS – RSS – CSS – T_rSS
$= 379.75 – (18.75 + 28.25 + 124.25) = 379.75 –171.25 = 208.50$

We apply ANOVA to find out whether there is any significant difference in the performance of the treatments. We formulate the following **null hypothesis**:

H_0: There is no significant difference in the training methods.

The null hypothesis has to be tested against the following **alternative hypothesis**:

H_1: There is a significant difference in the training methods.

We have to decide whether the null hypothesis has to be accepted or rejected at a desired level of significance (α).

We have the following ANOVA table.

ANOVA Table for LSD

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F
Row	4 – 1 = 3	18.75	18.75 / 3 = 6.25	34.75 / 6.25 = 5.56
Column	4 – 1 = 3	28.25	28.25 / 3 = 9.42	34.75 / 9.42 = 3.69
Treatment	4 – 1 = 3	124.25	124.25 / 3 = 41.42	41.42 / 34.75 = 1.19
Error	3 x 2 = 6	208.50	208.50 / 6 = 34.75	

Total	16 – 1 = 15	379.75		

Calculation of F value: We consider '**Treatment**'.
F = Greater variance / Smaller variance = 41.42 / 34.75 = 1.19
Degrees of freedom for greater variance (df_1) = 3
Degrees of freedom for smaller variance (df_2) = 6
Table value of F at 5% level of significance = 4.76

Inference:
Since the calculated value of F for the treatments is less than the table value of F, the null hypothesis is accepted and it is concluded that there is no significant difference in the training methods A, B, C and D, at 5% level of significance.

Problem 4
Examine the following production values got from four different machines A, B, C and D and determine whether there is any significant difference in the machines.

A 131	D 129	C 126	B 126
C 125	B 125	A 127	D 124
D 125	C 120	B 123	A 126
B 123	A 126	D 127	C 121

Solution :
In this design, there are 4 treatments A, B, C and D. In the lay out of the design, each treatment appears exactly once in each row as well as each column. Therefore this design is an LSD.

Since the entries in the design are large, we will follow the **coding method**. Subtract 120 from each entry. We get the following LSD.

A 11	D 9	C 6	B 6
C 5	B 5	A 7	D 4

D 5	C 0	B 3	A 6
B 3	A 6	D 7	C 1

R_1 = ∑ first row elements = 11 + 9 + 6 + 6 = 32
R_2 = ∑ second row elements = 5 + 5 + 7 + 4 = 21
R_3 = ∑ third row elements = 5 + 0 + 3 + 6 = 14
R_4 = ∑ fourth row elements = 3 + 6 + 7 + 1 = 17
C_1 = ∑ first column elements = 11 + 5 + 5 + 3 = 24
C_2 = ∑ second column elements = 9+ 5 + 0 + 6 = 20
C_3 = ∑ third column elements = 6 + 7 + 3 + 7 = 23
C_4 = ∑ fourth column elements = 6 + 4 + 6 + 1 = 17

From the given table, rewrite the experimental values for each treatment separately as follows: Treatment

A	B	C	D
11	6	6	9
7	5	5	4
6	3	0	5
6	3	1	7

T_1 = ∑ A = 11 +7 + 6 + 6 = 30
T_2 = ∑ B = 6 +5 + 3 + 3 = 17
T_3 = ∑ C = 6 + 5 + 0 + 1 = 12
T_4 = ∑ D = 9 + 4 + 5 + 7 = 25
$G = T_1 + T_2 + T_3 + T_3 = 30 + 17 + 12 + 25 = 84$
n = No. of treatments = 4
$N = n^2 = 16$
Correction Factor = $G^2/N = 84^2 / 16 = 7056 / 16 = 441$
$\sum y^2_{ij} = 11^2 + 9^2 + 6^2 + 6^2 + 5^2 + 5^2 + 7^2 + 4^2 + 5^2 + 0^2 + 3^2 + 6^2 + 3^2 + 6^2 + 7^2 + 1^2$
$= 121+81+36+36+25+25+49+16+25+0+9+36+9+36+49+1 = 554$

The total sum of squares (TSS), Row sum of squares (RSS), Column sum of squares (CSS) and Treatment sum of squares (T_rSS) are calculated as follows:

$TSS = \sum y^2_{ij} - CF = 554 - 441 = 113$

$RSS = R_1^2 / 4 + R_2^2 / 4 + R_3^2 / 4 + R_4^2 / 4 - CF$
$= 32^2 / 4 + 21^2 / 4 + 14^2 / 4 + 17^2 / 4 - 441$

$= 1024 / 4 + 441 / 4 + 196 / 4 + 289 / 4 – 441$

$= 256 + 110.25 + 49 + 72.25 – 441 = 487.5 – 441 = 46.5$

$CSS = C_1^2 / 4 + C_2^2 / 4 + C_3^2 / 4 + C_4^2 / 4 – CF$

$= 24^2 / 4 + 20^2 / 4 + 23^2 / 4 + 17^2 / 4 – 441$

$= 576/ 4 + 400 / 4 + 529 / 4 + 289 / 4 – 441$

$= 144 + 100 + 132.25 + 72.25 – 441 = 448.5 – 441 = 7.5$

$T_rSS = T_1^2 / 4 + T_2^2 / 4 + T_3^2 / 4 + T_4^2 / 4 – CF$

$= 30^2 / 4 + 17^2 / 4 + 12^2 / 4 + 25^2 / 4 – 441$

$= 900 / 4 + 289 / 4 + 144 / 4 + 625 / 4 – 441$

$= 225 + 72.25 + 36 + 156.25 – 441 = 489.5 – 441 = 48.5$

$ESS = TSS – RSS – CSS – T_rSS$

$= 113 – (46.5 + 7.5 + 48.5) = 113 – 102.5 = 10.5$

We formulate the following **null hypothesis**:

H_0: There is no significant difference in the performance of the machines.

The null hypothesis has to be tested against the following **alternative hypothesis**:

H_1: There is a significant difference in the performance of the machines.

We have to decide whether the null hypothesis has to be accepted or rejected at a desired level of significance (α).

We have the following ANOVA table.

ANOVA Table for LSD

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F
Row	4 – 1 = 3	46.5	46.5 / 3 = 15.50	15.50 / 1.75 = 8.857
Column	4 – 1 = 3	7.5	7.5 / 3 = 2.50	2.50 / 1.75 = 1.429
Treatment	4 – 1 = 3	48.5	48.5 / 3 = 16.17	16.17 / 1.75 = 9.240
Error	3 x 2 = 6	10.5	10.5 / 6 = 1.75	
Total	16 – 1 = 15	113.0		

Calculation of F value: We consider '**Treatment**'.

F = Greater variance / Smaller variance = 16.17 / 1.75 = 9.240

Degrees of freedom for greater variance (df_1) = 3
Degrees of freedom for smaller variance (df_2) = 6
Table value of F at 5% level of significance = 4.76

Inference:

Since the calculated value of F for the treatments is greater than the table value of F, the null hypothesis is rejected and the alternative hypothesis is accepted. It is concluded that there is a significant difference in the performance of the machines A, B, C and D at 5% level of significance.

Problem 5

The financial manager of a company obtained the following details on the LSD concerning the resources mobilized through 4 different schemes.

Source of Variation	Degrees of Freedom	SS
Row	3	270
Column	3	150
Treatment	3	1380
Error	6	156
Total	15	1956

Examine the data and find out whether there is any significant difference in the schemes.

Solution :

ANOVA Table for LSD

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance ratio F
Row	3	270	270 / 3 = 90	90 / 26 = 3.462
Column	3	150	150 / 3 = 50	50 / 26 = 1.923
Treatment	3	1380	1380 / 3 = 460	460 / 26 = 17.692
Error	6	156	156 / 6 = 26	
Total	15	1956		

Null hypothesis:

$\mathbf{H_0}$: There is no significant difference in the performance of the schemes.

Alternative hypothesis:

$\mathbf{H_1}$: There is a significant difference in the performance of the schemes.

Calculation of F value: We consider '**Treatment**'.
F = Greater variance / Smaller variance = 460 / 26 = 17.692
Degrees of freedom for greater variance (df_1) = 3
Degrees of freedom for smaller variance (df_2) = 6
Table value of F at 5% level of significance = 4.76

Inference:

Since the calculated value of F for the treatments is greater than the table value of F, the null hypothesis is rejected and the alternative hypothesis is accepted. It is concluded that there is a significant difference in the financial schemes A, B, C and D, at 5% level of significance.

QUESTIONS

1. What is an experimental design? Explain.
2. Explain the key concepts in experimental design.
3. Explain the steps in experimental design.
4. Explain the terms Replication, Randomization and Local Control.
5. What is meant by the lay out of an experimental design? Explain with an example.
6. What is a data allocation table? Give an example.
7. Describe a Completely Randomized Design.
8. Describe a Randomized Block Design.
9. Describe a Latin Square Design.
10. Explain the construction of a lay out of a Latin Square Design.
11. Explain the managerial application of an experimental design.

LESSON 4 PARTIAL AND MULTIPLE CORRELATION

LESSON OUTLINE

- The concept of partial correlation
- The concept of multiple correlation

LEARNING OBJECTIVES

After reading this lesson you should be able to

- determine partial correlation coefficient
- determine multiple correlation coefficient

Lesson 4

PARTIAL AND MULTIPLE CORRELATION

I. PARTIAL CORRELATION

Recall that simple correlation is a measure of the relationship between a dependent variable and another independent variable. For example, if the performance of a sales person depends only on the training that he has received, then the relationship between the training and the sales performance is measured by the simple correlation coefficient r. However, a dependent variable may depend on several variables. For example, the yarn produced in a factory may depend on the efficiency of the machine, the quality of cotton, the efficiency of workers, etc. It becomes necessary to have a measure of relationship in such complex situations. Partial correlation is used for this purpose. The technique of partial correlation proves useful when one has to develop a model with 3 to 5 variables.

Suppose Y is a dependent variable, depending on n other variables X_1, X_2, …, X_n. Partial correlation is a measure of the relationship between Y and any one of the variables X_1, X_2,…,X_n, as if the other variables have been eliminated from the situation.

The partial correlation coefficient is defined in terms of simple correlation coefficients as follows.

Let $r_{12.3}$ denote the correlation of X_1 and X_2 by eliminating the effect of X_3.
Let r_{12} be the simple correlation coefficient between X_1 and X_2.
Let r_{13} be the simple correlation coefficient between X_1 and X_3.
Let r_{23} be the simple correlation coefficient between X_2 and X_3.

Then we have

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{(1-r^2_{13})\ (1-r^2_{23})}}$$

Similarly, $$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1-r^2_{12})\ (1-r^2_{32})}}$$

and $$r_{32.1} = \frac{r_{23} - r_{21}\ r_{13}}{\sqrt{(1-r^2_{21})\ (1-r^2_{13})}}$$

Problem 1

Given that $r_{12} = 0.6$, $r_{13} = 0.58$, $r_{23} = 0.70$ determine the partial correlation coefficient $r_{12.3}$

Solution:

We have

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{(1-r^2_{13})\ (1-r^2_{23})}}$$

$$= \frac{0.6 - 0.58x0.70}{\sqrt{(1-(0.58)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.6 - 0.406}{\sqrt{(1-0.3364)\ (1-0.49)}}$$

$$= \frac{0.194}{\sqrt{0.6636x\ 0.51}}$$

$$= \frac{0.194}{0.8146x0.7141} = \frac{0.194}{0.5817} = 0.3335$$

Problem 2

If $r_{12} = 0.75$, $r_{13} = 0.80$, $r_{23} = 0.70$, find the partial correlation coefficient $r_{13.2}$

Solution:

We have

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1-r^2_{12})\ (1-r^2_{32})}}$$

$$= \frac{0.8 - 0.75X0.70}{\sqrt{(1-(0.75)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.8 - 0.525}{\sqrt{(1-0.5625)\ (1-0.49)}}$$

$$= \frac{0.275}{\sqrt{(0.4375)\ (0.51)}}$$

$$= \frac{0.275}{0.6614X0.7141} \qquad = \frac{0.275}{0.4723} \quad = 0.5823$$

II. MULTIPLE CORRELATION

When the value of a variable is influenced by another variable, the relationship between them is a simple correlation. In a real life situation, a variable may be influenced by many other variables. For example, the sales achieved for a product may depend on the income of the consumers, the price, the quality of the product, sales promotion techniques, the channels of distribution, etc. In this case, we have to consider the joint influence of several independent variables on the dependent variable. Multiple correlation arises in this context.

Suppose Y is a dependent variable, which is influenced by n other variables $X_1, X_2, \ldots, X_n$. The multiple correlation is a measure of the relationship between Y and $X_1, X_2, \ldots, X_n$ considered together.

The multiple correlation coefficient is denoted by the letter R. The dependent variable is denoted by X_1. The independent variables are denoted by X_2, X_3, X_4,…, etc.

Meaning of Notations:

$R_{1.23}$ denotes the multiple correlation of the dependent variable X_1 with two independent variables X_2 and X_3 . It is a measure of the relationship that X_1 has with X_2 and X_3 .

$R_{2.13}$ is the multiple correlation of the dependent variable X_2 with two independent variables X_1 and X_3.

$R_{3.12}$ is the multiple correlation of the dependent variable X_3 with two independent variables X_1 and X_2.

$R_{1.234}$ is the multiple correlation of the dependent variable X_1 with three independent variables X_2 , X_3 and X_4.

Coefficient of Multiple Linear Correlation

The coefficient of multiple linear correlation is given in terms of the partial correlation coefficients as follows:

$$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$$

$$R_{3.12} = \frac{\sqrt{r^2_{31} + r^2_{32} - 2\, r_{31}\, r_{32}\, r_{12}}}{\sqrt{1 - r^2_{12}}}$$

Properties of the coefficient of multiple linear correlation:

1. The coefficient of multiple linear correlation R is a non-negative quantity. It varies between 0 and 1.

2. $R_{1.23} = R_{1.32}$
   $R_{2.13} = R_{2.31}$
   $R_{3.12} = R_{3.21,}$ etc.

3. $R_{1.23} \geq |r_{12}|$,
   $R_{1.32} \geq |r_{13}|$, etc.

Problem 3

If the simple correlation coefficients have the values $r_{12} = 0.6$, $r_{13} = 0.65$, $r_{23} = 0.8$, find the multiple correlation coefficient $R_{1.23}$

Solution:

We have $$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$= \frac{\sqrt{(0.6)^2 + (0.65)^2 - 2x0.6x0.65x0.8}}{\sqrt{1 - (0.8)^2}}$$

$$= \frac{\sqrt{0.36 + 0.4225 - 0.624}}{\sqrt{1 - 0.64}}$$

$$= \frac{\sqrt{0.7825 - 0.624}}{\sqrt{0.36}}$$

$$= \frac{\sqrt{0.1585}}{\sqrt{0.36}}$$

$$= \sqrt{0.4403} \qquad = 0.6636$$

Problem 4

Given that $r_{21} = 0.7$, $r_{23} = 0.85$ and $r_{13} = 0.75$, determine $R_{2.13}$

Solution:

We have $$R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$$

$$= \frac{\sqrt{(0.7)^2 + (0.85)^2 - 2 \times 0.7 \times 0.85 \times 0.75}}{\sqrt{1 - (0.75)^2}}$$

$$= \frac{\sqrt{0.49 + 0.7225 - 0.8925}}{\sqrt{1 - 0.5625}}$$

$$= \frac{\sqrt{1.2125 - 0.8925}}{\sqrt{0.4375}}$$

$$= \frac{\sqrt{0.32}}{\sqrt{0.4375}} \qquad = \sqrt{0.7314} \qquad = 0.8552$$

QUESTIONS

1. Explain partial correlation.
2. Explain multiple correlation.
3. State the properties of the coefficient of multiple linear correlation.

UNIT IV

LESSON 5 DISCRIMINATE ANALYSIS

LESSON OUTLINE

- An overview of Matrix Theory
- The objective of Discriminate Analysis
- The concept of Discriminant Function
- Determination of Discriminant Function
- Pooled covariance matrix

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the basic concepts in Matrix Theory
- understand the objective of Discriminate Analysis
- understand Discriminant Function
- calculate the Discriminant Function

Lesson 5

DISCRIMINATE ANALYSIS

PART – I: AN OVERVIEW OF MATRIX THEORY

First, we have an overview of matrix theory required for discriminate analysis.

A matrix is a rectangular or square array of numbers. The matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ & & \\ a_{m1} & a_{m2} & a_{mn} \end{bmatrix}$$

is a rectangular matrix with m rows and n columns. We say that it is a matrix of type $m \times n$. A matrix with n rows and n columns is called a square matrix. We say that it is a matrix of type $n \times n$.

A matrix with just one row is called a row matrix or a row vector.

Eg: $\begin{pmatrix} a_1 & a_2 & a_n \end{pmatrix}$

A matrix with just one column is called a column matrix or a column vector.

Eg: $\begin{bmatrix} b_1 \\ b_2 \\ \\ b_m \end{bmatrix}$

A matrix in which all the entries are zero is called a zero matrix.

Addition of two matrices is accomplished by the addition of the numbers in the corresponding places in the two matrices. Thus we have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{bmatrix}$$

Multiplication of a matrix by a scalar is accomplished by multiplying each element in the matrix by that scalar. Thus we have

$$k\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} ka_{11} & ka_{12} \\ ka_{21} & ka_{22} \end{bmatrix}$$

$$k\begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} = \begin{pmatrix} ka_1 & ka_2 & \cdots & ka_n \end{pmatrix}$$

$$k\begin{bmatrix} b_1 \\ b_2 \\ \\ b_m \end{bmatrix} = \begin{bmatrix} kb_1 \\ kb_2 \\ \\ kb_m \end{bmatrix}$$

When a matrix A of type $m \times n$ and a matrix B of type $n \times p$ are multiplied, we obtain a matrix C of type $m \times p$. To get the element in the ith row, jth column of C, consider the elements of the ith row in A and the elements in the jth column of B, multiply the corresponding elements and take the sum. Thus we have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

The matrix I = $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is called the identity matrix of order 2. Similarly we can consider identity matrices of higher order. The identity matrix has the following property: If the matrices A and I are of type $n \times n$, then A I = I A = A.

Consider a square matrix of order 2. Denote it by A = $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. The determinant of A = det A = $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ = ad – bc. If it is zero, we say that A is a singular matrix. If it is not zero, we say that A is a non-singular matrix. When $ad - bc \neq 0$, A has a multiplicative inverse, denoted by A^{-1} with the property that $AA^{-1} = A^{-1}A = I$.

We have

$$A^{-1} = \frac{1}{\det A}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Note that

$$\frac{1}{ad-bc}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A symmetric matrix is the one in which the first row and first column are identical; the second row and second column are identical; and so on.

Eg:

$$\begin{bmatrix} a & b \\ b & d \end{bmatrix} \text{ and } \begin{bmatrix} a & h & g \\ h & b & f \\ g & f & c \end{bmatrix}$$

are similar matrices.

PART – II: **DISCRIMINATE ANALYSIS**

The objective of discriminate analysis

The objective of discriminate analysis (also known as discriminant analysis) is to separate a population (or samples from the population) into two distinct groups or two distinct conditionalities. After such a separation is made, we should be able to discriminate one group against the other. In other words, if some sample data is given, it should be possible for us to say with certainty whether that sample data has come from the first group or the second group. For this purpose, a function called '**Discriminant function'** is constructed. It is a linear function and it is used to describe the differences between two groups.

It is to be noted that the concept of discriminant function is applicable when there are more than 2 distinct groups also. However, we restrict ourselves to a situation of two distinct groups only. The discriminant function is the linear combination of the observations from the two groups which minimizes the distance between the mean vectors of the two groups after some transformation of the vectors.

Suppose we consider 2 variables both taking values under two different conditions denoted by condition I and condition II. Suppose there are m samples for each variable under condition I and n samples for each variable under condition II.

Let the values of the samples be as follows:

Condition I		Condition II	
Variable 1	Variable 2	Variable 1	Variable 2
p_1	q_1	α_1	β_1
p_2	q_2	α_2	β_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_m	q_m	α_n	β_n

Determine the means of the samples for the two variables under the two conditions.

Let $\overline{p}$ be the mean of the values of variable 1 under condition I.
Let $\overline{q}$ be the mean of the values of variable 2 under condition I.
Let $\overline{\alpha}$ be the mean of the values of variable 1 under condition II.
Let $\overline{\beta}$ be the mean of the values of variable 2 under condition II.
Let $\overline{y}_1, \overline{y}_2$ denote the column vectors whose entries are the mean values under conditions I, II respectively.

$$\text{i.e.,} \qquad \overline{y}_1 = \begin{bmatrix} \overline{p} \\ \overline{q} \end{bmatrix}, \quad \overline{y}_2 = \begin{bmatrix} \overline{\alpha} \\ \overline{\beta} \end{bmatrix}$$

Calculate the column vector $\overline{y}_1 - \overline{y}_2 = \begin{bmatrix} (\overline{p} - \overline{\alpha}) \\ (\overline{q} - \overline{\beta}) \end{bmatrix}$. The pooled covariance matrix S is obtained as follows:

$$S=\frac{1}{m+n-2}\begin{bmatrix} \sum_{i=1}^{m}(p_i-\bar{p})^2+\sum_{j=1}^{n}(\alpha_j-\bar{\alpha})^2 & \sum_{i=1}^{m}(p_i-\bar{p})(q_i-\bar{q})+\sum_{j=1}^{n}(\alpha_j-\bar{\alpha})(\beta_j-\bar{\beta}) \\ \sum_{i=1}^{m}(p_i-\bar{p})(q_i-\bar{q})+\sum_{j=1}^{n}(\alpha_j-\bar{\alpha})(\beta_j-\bar{\beta}) & \sum_{i=1}^{m}(q_i-\bar{q})^2+\sum_{j=1}^{n}(\beta_j-\bar{\beta})^2 \end{bmatrix}$$

Note that the inverse of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, provided $ad-bc\neq 0$.

Calculate the inverse of the matrix S. Denote it by S^{-1}. Find the matrix product $S^{-1}(\bar{y}_1-\bar{y}_2)$. The result is a column vector order 2. Denote it by δ and the entries by λ and μ. Then $\delta=\begin{bmatrix} \lambda \\ \mu \end{bmatrix}$

Fisher's discriminant function Z is obtained as

$$Z=\lambda y_1+\mu y_2 .$$

Application:

Given an observation of the attributes, we can use the discriminant function to decide whether it arose from condition I or condition II.

Problem

A tourism manager adopts two different strategies. Under each strategy, the number of tourists and the profits earned (in thousands of rupees) are as recorded below.

Strategy I	
No. of tourists	Profit earned
30	60
32	64
30	65
38	61
40	65

Strategy II	
No. of tourists	Profit earned

38	55
40	61
37	57
36	55
46	58
41	61
42	59

Construct Fisher's discriminant function and examine whether the strategies provide an effective tool of discrimination of the tourist operations.

Solution:

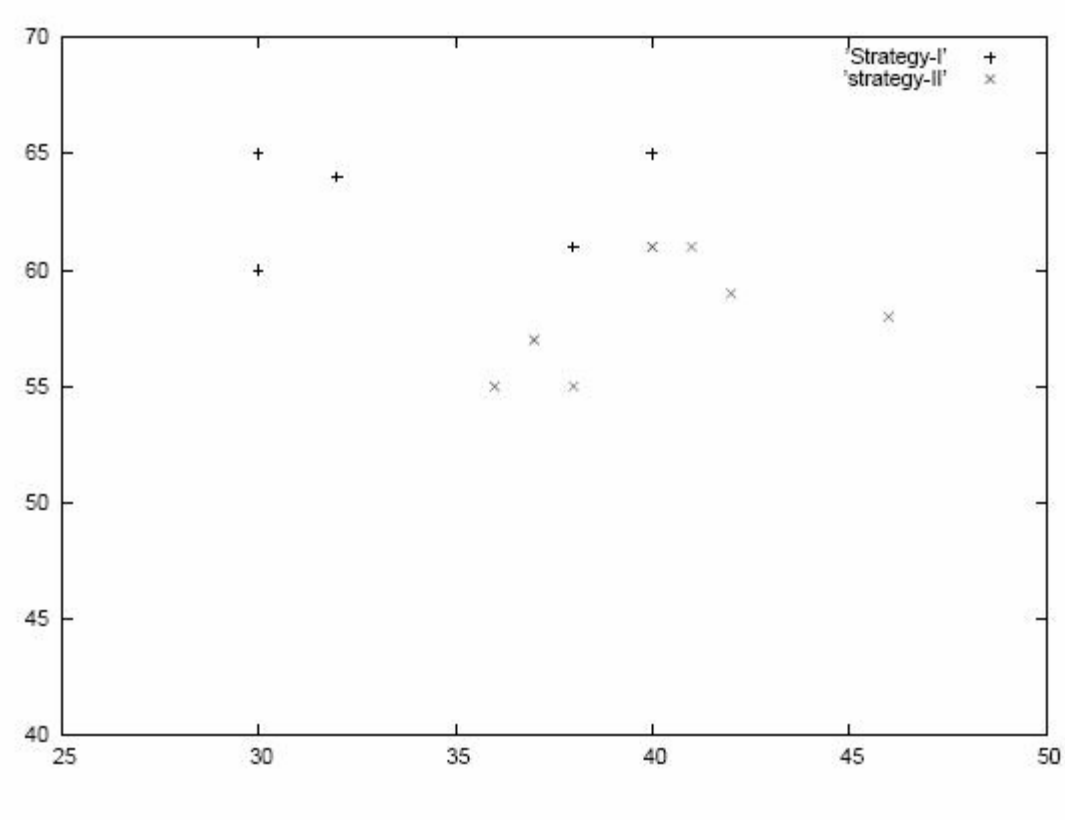
The given values are plotted in a graph. One point belonging to Strategy I seems to be an outlier as it is closer to the points of Strategy II. The other points seem to fall in two clusters. We shall examine this phenomenon by means of Fisher's discriminant function.

We have

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} 30 \\ 32 \\ 30 \\ 38 \\ 40 \end{bmatrix}, \quad \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix} = \begin{bmatrix} 60 \\ 64 \\ 65 \\ 61 \\ 65 \end{bmatrix}, \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \end{bmatrix} = \begin{bmatrix} 38 \\ 40 \\ 37 \\ 36 \\ 46 \\ 41 \\ 42 \end{bmatrix}, \quad \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix} = \begin{bmatrix} 55 \\ 61 \\ 57 \\ 55 \\ 58 \\ 61 \\ 59 \end{bmatrix}$$

The means of the above 4 columns are obtained as

$$\bar{p} = \frac{170}{5} = 34, \quad \bar{q} = \frac{315}{5} = 63, \quad \bar{\alpha} = \frac{280}{7} = 40, \quad \bar{\beta} = \frac{406}{7} = 58$$



$\bar{y}_1$ = column vector containing the mean values under strategy I

$$= \begin{bmatrix} \bar{p} \\ \bar{q} \end{bmatrix} = \begin{bmatrix} 34 \\ 63 \end{bmatrix}$$

$\bar{y}_2$ = column vector containing the mean values under strategy II

$$= \begin{bmatrix} \bar{\alpha} \\ \bar{\beta} \end{bmatrix} = \begin{bmatrix} 40 \\ 58 \end{bmatrix}$$

Therefore we get

$$\bar{y}_1 - \bar{y}_2 = \begin{bmatrix} 34 \\ 63 \end{bmatrix} - \begin{bmatrix} 40 \\ 58 \end{bmatrix} = \begin{bmatrix} -6 \\ 5 \end{bmatrix}$$

Calculation of $p_i - \overline{p}$, $q_i - \overline{q}$ etc.,

P	q	$p_i - \overline{p}$ $= p - 34$	$q_i - \overline{q}$ $= q - 63$	$(p_i - \overline{p})^2$	$(p_i - \overline{p})(q_i - \overline{q})$	$(q_i - \overline{q})^2$
30	60	- 4	- 3	16	12	9
32	64	- 2	1	4	- 2	1
30	65	- 4	2	16	- 8	4
38	61	4	- 2	16	- 8	4
40	65	6	2	36	12	4
				88	6	22

Calculation of $\alpha_j - \overline{\alpha}$, $\beta_j - \overline{\beta}$, etc.,

α	β	$\alpha_j - \overline{\alpha}$ $= \alpha - 40$	$\beta_j - \overline{\beta}$ $= \beta - 58$	$(\alpha_j - \overline{\alpha})^2$	$(\alpha_j - \overline{\alpha})(\beta_j - \overline{\beta})$	$(\beta_j - \overline{\beta})^2$
38	55	- 2	- 3	4	6	9
40	61	0	3	0	0	9
37	57	- 3	- 1	9	3	1
36	55	- 4	- 3	16	12	9
46	58	6	0	36	0	0
41	61	1	3	1	3	9
42	59	2	1	4	2	1
				70	26	38

$$\sum_{i=1}^{5}(p_i - \overline{p})^2 + \sum_{j=1}^{7}(\alpha_j - \overline{\alpha})^2 = 88 + 70 = 158$$

$$\sum_{i=1}^{5}(p_i - \overline{p})(q_i - \overline{q}) + \sum_{j=1}^{7}(\alpha_j - \overline{\alpha})(\beta_j - \overline{\beta}) = 6 + 26 = 32$$

$$\sum_{i=1}^{5}(q_i - \overline{q})^2 + \sum_{j=1}^{7}(\beta_j - \overline{\beta})^2 = 22 + 38 = 60$$

m + n – 2 = 5 + 7 – 2 = 10.

The pooled covariance matrix

$$S = \frac{1}{10}\begin{bmatrix} 158 & 32 \\ 32 & 60 \end{bmatrix} = \begin{bmatrix} 15.8 & 3.2 \\ 3.2 & 6 \end{bmatrix}$$

$$\det S = 94.8 - 10.24 = 84.56$$

$$S^{-1} = \frac{1}{84.56}\begin{bmatrix} 6 & -3.2 \\ -3.2 & 15.8 \end{bmatrix} = \begin{bmatrix} 0.071 & -0.038 \\ -0.038 & 0.187 \end{bmatrix}$$

$$\delta = \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = S^{-1}\left(\bar{y}_1 - \bar{y}_2\right)$$

$$= \begin{bmatrix} 0.071 & -0.038 \\ -0.038 & 0.187 \end{bmatrix}\begin{bmatrix} -6 \\ 5 \end{bmatrix} = \begin{bmatrix} -0.616 \\ 1.163 \end{bmatrix}$$

Fisher's discriminant function is obtained as

$$Z = \lambda y_1 + \mu y_2$$

$$= -0.616 y_1 + 1.161 y_2$$

where y_1 denotes the number of tourists and y_2 is the profit earned

Inference

We evaluate the discriminant function for the data given in the problem.

Strategy I		
No. of tourists (y_1)	Profit earned (y_2)	Z
30	60	51.3
32	64	54.72
30	65	57.12
38	61	47.54
40	65	50.96
Strategy II		
No. of tourists (y1)	Profit earned (y_2)	Z
38	55	40.56
40	61	46.30
37	57	43.50
36	55	41.79
46	58	39.12
41	61	45.69
42	59	42.75

By referring to the projected values of the discriminant function, it is seen that the discrimination function is able to separate the two strategies.

QUESTIONS

1. Explain the objective of discriminate analysis.
2. Briefly describe how discriminate analysis is carried out.

LESSON 6 CLUSTER ANALYSIS

LESSON OUTLINE

- The objective of cluster analysis
- Cluster analysis for qualitative data
- Resemblance matrix
- Simple matching coefficient
- Pessimistic, moderate, optimistic estimates of similarity
- Object-attribute incidence matrix
- Matching coefficient matrix
- Cluster analysis for quantitative data
- Hierarchical cluster analysis
- Euclidean distance matrix
- Dendogram

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the objective of cluster analysis
- perform cluster analysis for qualitative data
- perform cluster analysis for quantitative data
- understand resemblance matrix
- determine simple matching coefficient
- understand the properties of simple matching coefficient
- determine pessimistic, moderate, optimistic estimates of similarity
- understand object-attribute incidence matrix
- understand matching coefficient matrix
- find out Euclidean distance matrix
- construct Dendogram

Lesson 6

CLUSTER ANALYSIS

THE OBJECTIVE OF CLUSTER ANALYSIS

A cluster means a group of objects which remain together as far as a certain characteristic is concerned. When several objects are examined systematically, the cluster analysis seeks to put similar objects in the same cluster and dissimilar objects in different clusters so that each object will be allotted to one and only one cluster. Thus, it is a method for estimation of similarities among multivariate data. Similarity or dissimilarity is concerned with a certain attribute like magnitude, direction, shape, distance, colour, smell, taste, performance, etc.

Thus, it is to be seen that objects with similar description are pooled together to form a single cluster and objects with dissimilar properties will contribute to distinct clusters. For this purpose, given a set of objects, one has to determine which objects in that set are similar and which objects are dissimilar.

Method of cluster analysis

Cluster analysis is a complex task. However, we can have a broad outline of this analysis. One has to carry out the following steps:

1. Identify the objects that are required to be put in different clusters.
2. Prepare a list of attributes possessed by the objects under consideration. If they are too many, identify the important ones with the help of experts.
3. Identify the common attributes possessed by two or more objects.
4. Find out the attributes which are present in one object and absent in other objects.
5. Evolve a measure of similarity or dissimilarity. In other words, evolve a measure of "togetherness" or "standing apart".
6. Apply a standard algorithm to separate the objects into different clusters.

Application of cluster analysis

The concept of cluster analysis has applications in a variety of areas. A few examples are listed below:

1. A marketing manager can use it to find out which brands of products are perceived to be similar by the consumers.
2. A doctor can apply this method to find out which diseases follow the same pattern of occurrence.
3. An agriculturist may use it to determine which parts of his land are similar as regards the cultivating crop.
4. Once a set of objects have been put in different clusters, the top level management can take a policy decision as to which cluster has to be paid more attention and which cluster needs less attention, etc. Thus it will help the management in the decision on market segmentation.
   In short, cluster analysis finds applications in so many contexts.

I. Method of Cluster Analysis for Qualitative Data

We consider a case of binary attributes. They have two states, namely present or absent. Suppose we have to evolve a measure of resemblance between two objects P and Q. Suppose we take into consideration certain pre-determined attributes. If a certain attribute is present in an object, we will indicate it by 1 and if that attribute is absent we indicate it by 0. Count the number of attributes which are present in both the objects, which are absent in both the objects and which are present in one object but not in the other. We use the following notations.

a	=	Number of attributes present in both P and Q,
b	=	Number of attributes present in P but not in Q,
c	=	Number of attributes present in Q but not in P,
d	=	Number of attributes absent in both P and Q.

Among these quantities, a and d are counts for matched pairs of attributes while b and c are counts for unmatched pairs of attributes.

Resemblance matrix of two objects

The resemblance matrix of two objects P and Q consists of the values a, b, c, d as its entries. It is shown below.

		Q	
		1	0
P	1	a	b
	0	c	d

Simple matching coefficient

We consider a similarity coefficient called simple matching coefficient C(P,Q), defined as the ratio of the matched pairs of attributes to the total number of attributes. i.e., $C(P,Q) = \frac{a+d}{a+b+c+d}$

Properties of simple matching coefficient

1. The denominator in C(P,Q) shows that he simple matching coefficient gives equal
   weight for the unmatched pairs of attributes as well as the matched pairs.
2. The minimum value of C(P,Q) is 0.
3. The maximum value of C(P,Q) is 1.
4. A value of C(P,Q) = 1 indicates perfect similarity between the objects P and Q. This occurs when there are no unmatched pairs of attributes. i.e., b = c = 0.
5. A value of C(P,Q) = 0 indicates maximum dissimilarity between the objects P and Q. This occurs when there are no matched pairs of attributes. i.e., a = d = 0.
6. C(P,Q) = C(Q,P).
7. Using C(P,Q), we can estimate the percentage of similarity between P and Q.
8. C(P,P) = 1 since b = c = 0.

Illustrative Problem 1

A tourist is interested to evaluate two tourist spots P, Q with regard to their similarity and dissimilarity. He considers 10 attributes of the tourist spots and collects the following data matrix.

Attribute	Tourist Spot 1	Tourist Spot 2
1	1	1

2	0	0
3	1	1
4	0	0
5	0	1
6	1	1
7	1	1
8	1	1
9	1	0
10	1	1

Determine whether the two tourist spots are similar or not.

Solution:

We obtain the following resemblance matrix.

		Q	
		1	0
P	1	a = 6	b = 1
	0	c = 1	d = 2

We obtain the **similarity coefficient** as

$$C(P,Q) = \frac{a+d}{a+b+c+d}$$

$$= \frac{6+2}{6+1+1+2}$$

$$= \frac{8}{10} = 0.8$$

Inference

It is estimated that there is 80% similarity between the two tourist spots P and Q.

Matching coefficient with correction term

The correction term in the matching coefficient can be defined in several ways. We consider two specific approaches.

(a) Rogers and Tanimoto coefficient of matching

By giving double weight for unmatched pairs of attributes, the matching coefficient with correction term is defined as

$$C(P,Q) = \frac{a+d}{a+d+2(b+c)}.$$

Perfect similarity between P and Q occurs when b = c = 0. In this case, C(P,Q) = 1.

Maximum dissimilarity between P and Q occurs when a = d = 0. In this case, C(P,Q) = 0.

(b) Sokal and Sneath coefficient of matching

By giving double weight for matched pairs of attributes, the matching coefficient with correction term is defined as

$$C(P,Q) = \frac{2(a+d)}{2(a+d)+b+c}.$$

Perfect similarity between P and Q occurs when b = c = 0. In this case, C(P,Q) = 1.

Maximum dissimilarity between P and Q occurs when a = d = 0. In this case, C(P,Q) = 0.

Example

If we adopt Rogers and Tanimoto principle in the above problem, we get

$$C(P,Q) = \frac{6+2}{6+2+2(1+1)} = \frac{8}{12} = 0.67.$$

So the estimate of similarity between P and Q is is 67%

If we adopt Sokal and Sneath principle in the above example, we get

$$C(P,Q) = \frac{2(6+2)}{2(6+2)+1+1} = \frac{16}{18} = 0.89.$$

Thus the similarity between P and Q is estimated as 89%

Comparison of the three coefficients of similarity:

One can verify the following relation:

$$\frac{a+d}{a+d+2(b+c)} \le \frac{a+d}{a+b+c+d} \le \frac{2(a+d)}{2(a+d)+b+c}.$$

i.e., Rogers-Tanimoto Coefficient ≤ Simple matching Coefficient ≤ Sokal-Sneath Coefficient.

It is observed that Rogers and Tanimoto principle provides a **pessimistic estimate** of similarity. On the other hand, Sokal and Sneath principle gives an

optimistic estimate of similarity. The simple matching coefficient (without any correction term) gives a **moderate estimate** of similarity.

Clustering through object-attribute incidence matrix

Consider a set of objects. Enumerate the attributes of the objects. Not all the attributes will be present in all the objects. The object-attribute incidence matrix consists of the entries 0 and 1. If a certain attribute is present in an object, the corresponding place in the matrix is marked by 1; otherwise it is marked by 0. This matrix is useful in separating the objects into clusters.

Illustrative Problem 2

An expert of fashion designs identifies six fashions and five important attributes of fashions. He obtains the following object-attribute incidence matrix.

		Object					
		1	2	3	4	5	6
Attribute	1	1	0	0	0	0	1
	2	0	0	0	1	1	0
	3	0	1	0	0	1	0
	4	0	1	0	1	0	0
	5	1	0	1	0	0	1

Separate the objects into two clusters.

Solution:

Method I: By examination of the entries in the object-attribute incidence matrix

Denote the 6 objects by $O_1, O_2, O_3, O_4, O_5, O_6$ and the 5 attributes by A_1, A_2, A_3, A_4, A_5.

Consider the object O_1. Attributes A_1 and A_5 are present in object O_1 and the other 3 attributes are absent in it. Compare other objects with object O_1 and find which object possesses similar attributes. For this, consider the columns of the matrix. It is noticed that columns 1 and 6 in the matrix are identical. i.e., Attributes A_1 and A_5 are present in both the objects O_1 and O_6. All the other attributes are absent in both the objects. So the objects O_1 and O_6 can be put in a cluster. Denote this cluster by $\{O_1, O_6\}$.

The remaining objects are O_2, O_3, O_4, O_5. Consider the columns 2,3,4,5 in the matrix. No other column is identical to column 2. The object O_2 possesses the attributes A_3 and A_4. Identify other objects which possess at least one of these attributes. Objects O_4 possess attribute A_4. So put the objects O_2 and O_4 in a cluster. Denote this cluster by $\{O_2, O_4\}$.

The remaining objects are O_3 and O_5. The object O_3 possesses only the attribute A_5 and the same is possessed by objects O_1 and O_6. So the object O_3 is closer to the cluster $\{O_1, O_6\}$ rather than the cluster $\{O_2, O_4\}$. So enlarge the cluster $\{O_1, O_6\}$ by including the object O_3. Thus we get the cluster $\{O_1, O_6, O_3\}$.

The remaining object is O_5. It possesses attributes A_2 and A_3. These attributes are absent in the objects O_1, O_6, O_3. Attribute A_3 in present in object

O_2 and attribute A_2 is present in object O_4. So enlarge the cluster $\{O_2, O_4\}$ by including the object O_5. In this way we get the cluster $\{O_2, O_4, O_5\}$.

Result: Thus we obtain the following two clusters.

Cluster I: $\{O_1, O_3, O_6\}$ and

Cluster II: $\{O_2, O_4, O_5\}$.

The attributes present in cluster I are absent in cluster II and vice verse.

Method II: Application of simple matching coefficient

Calculate the matching coefficients of pairs of distinct objects. Since there are 6 objects, we have (6 x 5) / 2 = 15 such pairs. Tabulate the results as follows:

Counts of matched and unmatched pairs of attributes

Ordered pairs of objects	a	b	c	D	Simple matching coefficient = (a+b)/(a+b+c+d)
O_1 , O_2	0	2	2	1	0.2
O_1 , O_3	1	1	0	3	0.8
O_1 , O_4	0	2	2	1	0.2
O_1 , O_5	0	2	2	1	0.2
O_1 , O_6	2	0	0	3	1.0
O_2 , O_3	0	2	1	2	0.4
O_2 , O_4	1	1	1	2	0.6
O_2 , O_5	1	1	1	2	0.6
O_2 , O_6	0	1	2	2	0.4
O_3 , O_4	0	1	2	2	0.4
O_3 , O_5	0	1	2	2	0.4
O_3 , O_6	1	0	1	3	0.8

O_4, O_5	1	1	1	2	0.6
O_4, O_6	0	2	2	1	0.2
O_5, O_6	0	2	2	1	0.2

We form the **matching coefficient matrix** for the objects under consideration by entering the simple matching coefficients against the pairs of objects. It is a symmetric matrix since C(P,Q) = C(Q,P). In the present problem, we get the following matrix.

		Object					
		1	2	3	4	5	6
Object	1	1	0.2	0.8	0.2	0.2	1
	2	0.2	1	0.4	0.6	0.6	0.4
	3	0.8	0.4	1	0.4	0.4	0.8
	4	0.2	0.6	0.4	1	0.6	0.2
	5	0.2	0.6	0.4	0.6	1	0.2
	6	1	0.4	0.8	0.2	0.2	1

Consider the matching coefficients of pairs of distinct objects. Here there are 15 such pairs. The maximum among them is 1 = C(O_1, O_6). Thus O_1 and O_6 have the maximum similarity. Therefore, they can be put in a cluster. The next maximum matching coefficient is 0.8 possessed by the pairs (O_1,O_3) and (O_3,O_6). Therefore the objects O_1, O_3,O_6 can be clubbed together. The next maximum matching coefficient is 0.6 possessed by the pairs (O_2, O_4), (O_2,O_5) and (O_4, O_5). So the objects O_2, O_4, O_5 can be considered together. Since we have exhausted all the objects, the process is now complete.

Result: Thus we have arrived at Cluster I: $\{O_1, O_3, O_6\}$ and Cluster II: $\{O_2, O_4, O_5\}$.

II. Method of Cluster analysis for Quantitative Data Hierarchical cluster analysis

The aim of the hierarchical cluster analysis is to put the given objects into various clusters and to arrange the clusters in a hierarchical order. A cluster will consist of similar objects. Dissimilar objects will be put into different clusters. The clusters so formed will be arranged such that two clusters which contain somewhat similar objects will be grouped together. Two clusters which contain extremely dissimilar objects will stand apart in the hierarchical order.

Steps in hierarchical cluster analysis

The hierarchical cluster analysis comprises of the following steps.

1. Collect the necessary data in a matrix form. The columns in the matrix denote the objects taken for examination and the rows denote the attributes that describe the objects. This matrix is called the **data matrix**.
2. Standardize the data matrix.
3. Use the data matrix or the standardized data matrix to determine the values of "resemblance coefficient". It is measure of similarities among pairs of objects.
4. By means of the values of the resemblance coefficient, construct a diagram called a **dendogram.** It is a **tree-like structure**. A tree will exhibit the different clusters into which the given set of objects is decomposed. The tree will indicate the hierarchy of similarities among different pairs of objects. This is the reason for calling the method as hierarchical cluster analysis.

Illustrative problem 3

A marketing manager wishes to examine the sales performance of 4 sales persons P,Q,R,S in his division by means of cluster analysis. Records indicating their performance in the past 6 months are collected in the following table.

Unit: Rs. In lakhs

Month	Sales Performance			
	P	Q	R	S
January	20	22	25	23
February	22	23	27	24
March	24	24	28	25
April	19	21	22	20
May	20	22	24	21
June	21	23	25	24

Help the manager in arranging the sales persons in a hierarchical order according to their sales performance.

Solution:

First we construct a **Euclidean distance matrix**. This matrix is formed by entering the Euclidean distances against the pairs of objects. In our context, Euclidean distance does not refer to any geographical distance. It is a relative measure of the performance of two sales persons over the given period of time. It will indicate which two sales persons are similar in their performance and which two sales persons are extremely different in their performance.

Assume that there are n data values for each sales person. Denote the sales data of two persons by vectors P and Q as follows:

$$P = (X_1, X_2, ..., X_n)$$
$$Q = (Y_1, Y_2, ..., Y_n)$$

Then the **Euclidean distance** between them is denoted by d(P,Q) and is defined by the following relation:

$$\text{d(P,Q)} = \sqrt{\left[(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_n - Y_n)^2\right]}$$

(1)

Note that d(P,P) = 0 and d(Q,P) = d(P,Q). In the problem under consideration, n = 6. For the 4 sales persons P,Q,R,S, we have to calculate the 6 quantities d(P,Q), d(P,R), d(P,S), d(Q,R), d(Q,S), d(R,S). We have

$$P = (20, 22, 24, 19, 20, 21)$$
$$Q = (22, 23, 24, 21, 22, 23)$$
$$R = (25, 27, 28, 22, 24, 25)$$
$$S = (23, 24, 25, 20, 21, 24)$$

Using formula (1), calculate the Euclidean distances. We obtain

$$\begin{aligned} d(P,Q) &= \sqrt{(20-22)^2 + (22-23)^2 + (24-24)^2 + (19-21)^2 + (20-22)^2 + (21-23)^2} \\ &= \sqrt{(-2)^2 + (-1)^2 + (0)^2 + (-2)^2 + (-2)^2 + (-2)^2} \\ &= \sqrt{4+1+0+4+4+4} \\ &= \sqrt{17} \\ &= 4.1 \end{aligned}$$

correct to 1 place of decimals. Next we get

$$\begin{aligned} d(P,R) &= \sqrt{(20-25)^2 + (22-27)^2 + (24-28)^2 + (19-22)^2 + (20-24)^2 + (21-25)^2} \\ &= \sqrt{(-5)^2 + (-5)^2 + (-4)^2 + (-3)^2 + (-4)^2 + (-4)^2} \\ &= \sqrt{25+25+16+9+16+16} \\ &= \sqrt{107} \\ &= 10.3 \end{aligned}$$

$$\begin{aligned} d(P,S) &= \sqrt{(20-23)^2 + (22-24)^2 + (24-25)^2 + (19-20)^2 + (20-21)^2 + (21-24)^2} \\ &= \sqrt{9+4+1+1+1+9} \\ &= \sqrt{25} \\ &= 5 \end{aligned}$$

$$d(Q,R) = \sqrt{(22-25)^2 + (23-27)^2 + (24-28)^2 + (21-22)^2 + (22-24)^2 + (23-25)^2}$$
$$= \sqrt{9+16+16+1+4+4}$$
$$= \sqrt{50}$$
$$= 7.1$$

$$d(Q,S) = \sqrt{(22-23)^2 + (23-24)^2 + (24-25)^2 + (21-20)^2 + (22-21)^2 + (23-24)^2}$$
$$= \sqrt{1+1+1+1+1+1}$$
$$= \sqrt{6}$$
$$= 2.4$$

$$d(R,S) = \sqrt{(25-23)^2 + (27-24)^2 + (28-25)^2 + (22-20)^2 + (24-21)^2 + (25-24)^2}$$
$$= \sqrt{4+9+9+4+9+1}$$
$$= \sqrt{36}$$
$$= 6$$

The following Euclidean distance matrix is got for the sales persons P,Q,R and S.

$$\begin{array}{c} \\ P \\ Q \\ R \\ S \end{array} \begin{array}{c} \begin{array}{cccc} P & Q & R & S \end{array} \\ \begin{bmatrix} - & 4.1 & 10.3 & 5 \\ 4.1 & - & 7.1 & 2.4 \\ 10.3 & 7.1 & - & 6 \\ 5 & 2.4 & 6 & - \end{bmatrix} \end{array}$$

Determination of Dendogram:

We adopt a procedure called single linkage clustering method **(SLINK)**. This is based on the concept of nearest neighbours.

Consider the distance between different persons. They are d(P,Q), d(P,R), d(P,S), d(Q,R), d(Q,S), d(R,S). i.e., 4.1, 10.3, 5, 7.1, 2.4, 6

The minimum among them is 2.4 = d(Q,S). Thus Q and S are the nearest neighbors. Therefore, Q and S are selected to form a cluster at the first level, denoted by {Q,S}. Next, we have to add another object to the list {Q,S}. The remaining elements are P and R. We have to decide whether P should be added to the list {Q,S} or R should be added. So we have to determine which among P, R is nearer to the set {Q,S}. We consider the quantities

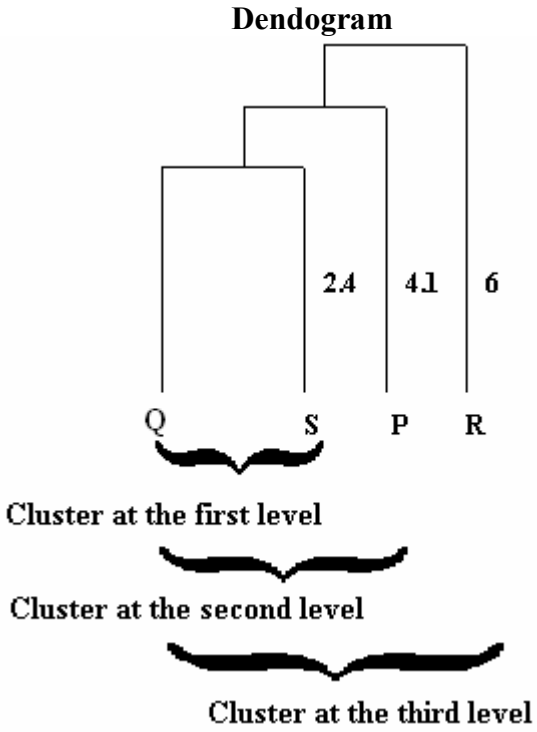
$$\begin{aligned} d\left((Q,S),P\right) &= Minimum\left[d\left(Q,P\right),d\left(S,P\right)\right] \\ &= Minimum\left[4.1,5\right] = 4.1 \\ d\left((Q,S),R\right) &= Minimum\left[d\left(Q,R\right),d\left(S,R\right)\right] \\ &= Minimum\left[7.1,6\right] = 6 \end{aligned}$$

Among these two quantities, we find Minimum [d((Q,S),P), d((Q,S),R)] = Minimum [4.1,6] = 4.1 = d((Q,S),P).

Therefore, P is nearer to the cluster {Q,S} rather than R. Consequently P is attached with the set {Q,S} and so we obtain the cluster {{Q,S}, P}. This is the cluster at the second level. If there are other objects remaining, we have to repeat the above procedure. In the present case, there is only one object remaining i.e., R. We add R to the cluster ((Q,S),P) to the form the cluster at the third level. We note that

$$\begin{aligned} d\left[\left((Q,S),P\right),R\right] &= Minimum\left[d\left(Q,R\right),\ d\left(S,R\right),\ d\left(P,R\right)\right] \\ &= Minimum\left[7.1,6,10.3\right] = 6 \end{aligned}$$

Using these values, we obtain the following diagram:



Inference

It is seen that sales persons Q, S are similar in their performance over the given of time. The next sales person somewhat similar to them is P. The sales person R stands apart.

QUESTIONS

1. Explain the objective of cluster analysis.
2. Briefly describe how cluster analysis is carried out.
3. State the properties of simple matching coefficient.
4. Describe the methods of obtaining pessimistic, moderate and optimistic estimates of the similarity between two objects.

5. Explain object-attribute incidence matrix.
6. Explain matching coefficient matrix.
7. What are the steps in hierarchical cluster analysis?
8. What is Euclidean distance matrix? Explain.
9. What is a dendogram? Explain.

LESSON 7

FACTOR ANALYSIS AND CONJOINT ANALYSIS

LESSON OUTLINE

- Factor Analysis
- Conjoint Analysis
- Steps in Development of Conjoint Analysis
- Applications of Conjoint Analysis
- Advantages and disadvantages of Conjoint Analysis
- Illustrative problems
- Multi-factor evaluation approach in Conjoint Analysis
- Two-factor evaluation approach in Conjoint Analysis

LEARNING OBJECTIVES

After reading this lesson you should be able to

- understand the concept of Factor Analysis
- understand the managerial applications of Factor Analysis
- understand the concept of Conjoint Analysis
- apply rating scale technique in Conjoint Analysis
- apply ranking method in Conjoint Analysis
- apply mini-max scaling method in Conjoint Analysis
- understand Multi-factor evaluation approach
- understand Two-factor evaluation approach
- understand the managerial applications of Conjoint Analysis

Lesson 7

FACTOR ANALYSIS AND CONJOINT ANALYSIS

PART I - FACTOR ANALYSIS

In a real life situation, several variables are operating. Some variables may be highly correlated among themselves.

Suppose, for example, a manager of a restaurant has to analyse six attributes of a new product. He undertakes a sample survey and finds out the responses of potential consumers. Suppose he obtains the following attribute correlation matrix.

		Attribute					
		1	2	3	4	5	6
Attribute	1	1.00	0.05	0.10	0.95	0.20	0.02
	2	0.05	1.00	0.15	0.10	0.60	0.85
	3	0.10	0.15	1.00	0.50	0.55	0.10
	4	0.95	0.10	0.50	1.00	0.12	0.08
	5	0.20	0.60	0.55	0.12	1.00	0.80
	6	0.02	0.85	0.10	0.08	0.80	1.00

Attribute Correlation Matrix

We try to group the attributes by their correlations. The high correlation values are observed for the following attributes.

Attributes 1, 4 with a very high correlation coefficient of 0.95.

Attributes 2, 4 with a high correlation coefficient of 0.85.

Attributes 3, 4 with a high correlation coefficient of 0.85.

As a result, it is seen that not all the attributes are independent. The attributes 1 and 4 have mutual influence on each other while the attributes 2,5 and 6 have mutual influence among themselves.

As far as attribute 3 is concerned, it has little correlation with the attributes 1, 2 and 6. Even with the other attributes 4 and 5, its correlation is not high. However, we can say that attribute 3 is somewhat closer to the variables 4 and 5 rather than the attributes 1, 2 and 6. Thus, from the given list of 6 attributes, it is possible to find out 2 or 3 common factors as follows:

I. 1) The common features of the attributes 1,3,4 will give a factor
2) The common features of the attributes 2,5,6 will give a factor

or

II. 1) The common features of the attributes 1,4 will give a factor
2) The common features of the attributes 2,5,6 will give a factor
3) The attribute 3 can be considered to be an independent factor

The factor analysis is a multivariate method. It is a statistical technique to identify the underlying factors among a large number of interdependent variables. It seeks to extract common factor variances from a given set of observations. It splits a number of attributes or variables into a smaller group of uncorrelated factors. It determines which variables belong together. This method is suitable for the cases with a number of variables having a high degree of correlation.

In the above example, we would like to filter down the attributes 1, 4 into a single attribute. Also we would like to do the same for the attributes 2, 5, 6. If a set of attributes (variables) A_1, A_2, …, A_k filter down to an attribute A_i ($1 \leq i \leq k$), we say that these attributes are loaded on the factor A_i or saturated with the factor A_i. Sometimes, more than one factor also may be identified.

Basic concepts in factor analysis

The following are the key concepts on which factor analysis is based.

Factor: A factor plays a fundamental role among a set of attributes or variables. These variables can be filtered down to the factor. A factor represents the combined effect of a set of attributes. Either there may be one such factor or several such factors in a real life problem based on the complexity of the situation and the number of variables operating.

Factor loading: A factor loading is a value that explains how closely the variables are related to the factor. It is the correlation between the factor and the variable. While interpreting a factor, the absolute value of the factor is taken into account.

Communality: It is a measure of how much each variable is accounted for by the underlying factors together. It is the sum of the squares of the loadings of the variable on the common factors. If A,B,C,… are the factors, then the communality of a variable is computed using the relation

h^2 = (The factor loading of the variable with respect to factor A)^2 +
(The factor loading of the variable with respect to factor B)^2 +
(The factor loading of the variable with respect to factor C)^2 + …..

Eigen value: The sum of the squared values of factor loadings pertaining to a factor is called an eigen value. It is a measure of the relative importance of each factor under consideration.

Total Sum of Squares (TSS)

It is the sum of the eigen values of all the factors.

Application of Factor Analysis:

1. Model building for new product development:

As pointed out earlier**,** a real life situation is highly complex and it consists of several variables**.** A model for the real life situation can be built by incorporating

as many features of the situation as possible. But then, with a multitude of features, it is very difficult to build such a highly idealistic model. A practical way is to identify the important variables and incorporate them in the model. Factor analysis seeks to identify those variables which are highly correlated among themselves and find a common factor which can be taken as a representative of those variables. Based on the factor loading, some of variables can be merged together to give a common factor and then a model can be built by incorporating such factors. Identification of the most common features of a product preferred by the consumers will be helpful in the development of new products.

2. Model building for consumers:

Another application of factor analysis is to carry out a similar exercise for the respondents instead of the variables themselves. Using the factor loading, the respondents in a research survey can be sorted out into various groups in such a way that the respondents in a group have more or less homogeneous opinions on the topics of the survey. Thus a model can be constructed on the groups of consumers. The results emanating from such an exercise will guide the management in evolving appropriate strategies towards market segmentation.

PART II - CONJOINT ANALYSIS

Introduction

Everything in the world is undergoing a change. There is a proverb saying that "the old order changes, yielding place to new". Due to rapid advancement in science and technology, there is fast communication across the world. Consequently, the whole world has shrunk into something like a village and thus now-a-days one speaks of the "global village". Under the present set-up, one

can purchase any product of his choice from whatever part of the world it may be available. Because of this reason, what was a seller's market a few years back has transformed into a buyer's market now.

In a seller's market of yesterday, the manufacturer or the seller could pass on a product according to his own perceptions and prescriptions. In the buyer's market of today, a buyer decides what he should purchase, what should be the quality of the product, how much to purchase, where to purchase, when to purchase, at what cost to purchase, from whom to purchase, etc. A manager is perplexed at the way a consumer takes a decision on the purchase of a product. In this background, conjoint analysis is an effective tool to understand a buyer's preferences for a good or service.

Meaning of Conjoint Analysis

A product or service has several attributes. By an attribute, we mean a characteristic, a property, a feature, a quality, a specification or an aspect. A buyer's decision to purchase a good or service is based on not just one attribute but a combination of several attributes. i.e., He is concerned with a join of attributes.

Therefore, finding out the consumer's preferences for individual attributes of a product or service may not yield accurate results for a marketing research problem. In view of this fact, conjoint analysis seeks to find out the consumer's preferences for a 'join of attributes'. i.e., a combination of several attributes.

Let us consider an example. Suppose a consumer desires to purchase a wrist watch. He would take into consideration several attributes of a wrist watch, namely the configuration details such as mechanism, size, dial, appearance and colour and other particulars such as strap, price, durability, warranty, after-sales service, etc. If a consumer is asked what is the important

aspect among the above list, he would reply that all attributes are important for him and so a manger cannot arrive at a decision on the design of a wrist watch. Conjoint analysis assumes that the buyer will base his decision not on just the individual attributes of the product but rather he would consider various combinations of the attributes, such as

'mechanism, colour, price, after-sales service',
or 'dial, colour, durability, warranty',
or 'dial, appearance, price, durability', etc.

This analysis would enable a manager in his decision making process in the identification of some of the preferred combinations of the features of the product.

The rank correlation method seeks to assess the consumer's preferences for individual attributes. In contrast, the conjoint analysis seeks to assess the consumer's preferences for combinations (or groups) of attributes of a product or a service. This method is also called an '**unfolding technique**' because preferences on groups of attributes unfold from the rankings expressed by the consumers. Another name for this method is '**multi-attribute compositional model**' because it deals with combinations of attributes.

Steps in the Development of Conjoint Analysis

The development of conjoint analysis comprises of the following steps.

1. Collect a list of the attributes (features) of a product or a service.
2. For each attribute, fix a certain number of points or marks. The more the number of points for an attribute, the more serious the consumers' concern on that attribute.
3. Select a list of combinations of various attributes.
4. Decide a mode of presentation of the attributes to the respondents of the study. i.e., whether it should be in written form, or oral form, or a pictorial representation, etc.
5. Inform the combinations of the attributes to the prospective customers.
6. Request the respondents to rank the combinations, or to rate them on a suitable scale, or to choose between two different combinations at a time.

7. Decide a procedure to aggregate the responses from the consumers. Any one of the following procedures may be adopted:
   (i). Go by the individual responses of the consumers.
   (ii). Put all the responses together and construct a single utility function.
   (iii). Split the responses into a certain number of segments such that within each segment, the preferences would be similar.
8. Choose the appropriate technique to analyze the data collected from the respondents.
9. Identify the most preferred combination of attributes.
10. Incorporate the result in designing a new product, construction of an advertisement copy, etc.

Applications of Conjoint Analysis

1. An idea of consumer's preferences for combinations of attributes will be useful in designing new products or modification of an existing product.
2. A forecast of the profits to be earned by a product or a service.
3. A forecast of the market share for the company's product.
4. A forecast of the shift in brand loyalty of the consumers.
5. A forecast of differences in responses of various segments of the product.
6. Formulation of marketing strategies for the promotion of the product.
7. Evaluation of the impact of alternative advertising strategies.
8. A forecast of the consumers' reaction to pricing policies.
9. A forecast of the consumes' reaction on the channels of distribution.
10. Evolving an appropriate marketing mix.
11. Even though the technique of conjoint analysis was developed for the formulation of corporate strategy, this method can be used to have a comprehensive knowledge of a wide range of areas such as family decision making process, pharmaceuticals, tourism development, public transport system, etc.

Advantages of Conjoint Analysis

1. The analysis can be carried out on physical variables.
2. Preferences by different individuals can be measured and pooled together to arrive at a decision.

Disadvantages of Conjoint Analysis

1. When more and more attributes of a product are included in the study, the number of combinations of attributes also increases, rendering the study highly difficult. Consequently, only a few selected attributes can be included in the study.
2. Gathering of information from the respondents will be a tough job.
3. Whenever novel combinations of attributes are included, the respondents will have difficulty in capturing such combinations.
4. The psychological measurements of the respondents may not be accurate.

In spite of the above stated disadvantages, conjoint analysis offers more scope to the researchers in identifying the consumers' preferences for groups of attributes.

Illustrative Problem 1 : Application of Rating Scale Technique

A wrist watch manufacturer desires to find out the combinations of attributes that a consumer would be interested in. After considering several attributes, the manufacturer identifies the following combinations of attributes for carrying out marketing research.

Combination – I	Mechanism, colour, price, after-scales service
Combination – II	Dial, colour, durability, warranty
Combination – III	Dial, appearance, price, durability
Combination – IV	Mechanism, dial, price, warranty

12 respondents are asked to rate the 4 combinations on the following 3 – point rating scale.

Scale – 1	:	Less important
Scale – 2	:	Somewhat important
Scale – 3	:	Very important

Their responses are given in the following table.

Rating of Combination

Respondent No	Combination I	Combination II	Combination III	Combination IV
1	Less important	Somewhat	Very important	Somewhat
2	Somewhat	Very important	Less important	Somewhat
3	Somewhat	Less important	Somewhat	Very important
4	Less important	Less important	Very important	Somewhat
5	Somewhat	Very important	Very important	Less important
6	Somewhat	Very important	Somewhat	Less important
7	Somewhat	Less important	Very important	Less important
8	Very important	Somewhat	Less important	Somewhat
9	Very important	Less important	Somewhat	Somewhat
10	Somewhat	Very important	Less important	Somewhat
11	Very important	Somewhat	Very important	Somewhat
12	Very important	Less important	Very important	Somewhat

Determine the most important and the least important combinations of the attributes.

Solution:

Let us assign scores to the scales as follows:

Sl. No.	Scale	Score
1	Less important	1
2	Somewhat important	3
3	Very important	5

The scores for the four combinations are calculated as follows:

Combination	Response	Score for Response	No. of Respondents	Total Score
I	Less important Somewhat important Very important	1 3 5	2 6 4	$1 X 2 = 2$ $3 X 6 = 18$ $5 X 4 = 20$
			12	40
II	Less important Somewhat important Very important	1 3 5	5 3 4	$1 X 5 = 5$ $3 X 3 = 9$ $5 X 4 = 20$
			12	34
III	Less important Somewhat important Very important	1 3 5	3 3 6	$1 X 3 = 3$ $3 X 3 = 9$ $5 X 6 = 30$
			12	42
IV	Less important Somewhat important Very important	1 3 5	3 8 1	$1 X 3 = 3$ $3 X 8 = 24$ $5 X 1 = 5$
			12	32

Let us tabulate the scores earned by the four combinations as follows:

Combination	Total scores
I	40
II	34
III	42
IV	32

Inference:

It is concluded that the consumers consider combination III as the most important and combination IV as the least important.

Note: For illustrating the concepts involved, we have taken up 12 respondents in the above problem. In actual research work, we should take a large number of

respondents, say 200 or 100. In any case, the number of respondents shall not be less than 30.

Illustrative Problem 2: Application of Ranking Method

A marketing manager selects four combinations of features of a product for study. The following are the ranks awarded by 10 respondents. Rank one means the most important and rank 4 means the least important.

Respondent No.	Rank Awarded			
	Combination I	Combination II	Combination III	Combination IV
1	2	1	3	4
2	1	4	2	3
3	1	2	3	4
4	3	2	4	1
5	4	1	2	3
6	1	2	3	4
7	4	3	2	1
8	3	1	2	4
9	3	1	4	2
10	4	1	2	3

Determine the most important and the least important combinations of the features of the product.

Solution:

Let us assign scores to the ranks as follows:

Rank	Score
1	10
2	8
3	6
4	4

The scores for the 4 combinations are calculated as follows:

Combination	Rank	Score for rank	No. of Respondents	Total Score
I	1	10	3	10 *X* 3 = 30
	2	8	1	8 *X* 1= 8
	3	6	3	6 *X* 3 = 18
	4	4	3	4 *X* 3 = 12
			10	68
II	1	10	5	10 *X* 5 = 50
	2	8	3	8 *X* 3 = 24
	3	6	1	6 *X* 1 = 6
	4	4	1	4 *X* 1 = 4
			10	84
III	1	10	Nil	--
	2	8	5	8 *X* 5 = 40
	3	6	3	6 *X* 3 = 18
	4	4	2	4 *X* 2 = 8
			10	66
IV	1	10	2	10 *X* 2 = 20
	2	8	1	8 *X* 1 = 8
	3	6	3	6 *X* 3 = 18
	4	4	4	4 *X* 4 = 16
			10	62

The final scores for the 4 combinations are as follows:

Combination	Score
I	68
II	84
III	66
IV	62

Inference:

It is seen that combination II is the most preferred one by the consumers and combination IV is the least preferred one.

Illustrative Problem 3: Application of Mini-Max Scaling Method

An insurance manager chooses 5 combinations of attributes of a social security plan for analysis. He requests 10 respondents to indicate their perceptions on the importance of the combinations by awarding the minimum score and the maximum score for each combination in the range of 0 to 100. The details of the responses are given below. Help the manager in the identification of the most important and the least important combinations of the attributes of the social security plan.

Respondent Number	Combination I		Combination II		Combination III		Combination IV		Combination V	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1	30	60	45	85	50	70	40	75	50	80
2	35	65	50	80	50	80	35	75	40	75
3	40	70	35	80	60	80	40	70	50	80
4	40	80	40	80	60	85	50	75	60	80
5	30	75	50	80	60	75	60	75	60	85
6	35	70	35	85	50	80	40	80	40	80
7	40	80	40	75	45	75	50	70	40	80
8	30	80	40	75	50	80	50	70	60	80
9	45	75	45	75	50	80	50	80	50	80
10	55	75	40	85	35	75	45	80	40	80

Solution:

For each combination, consider the minimum score and the maximum score separately and calculate the average in each case.

	Combination I		Combination II		Combination III		Combination IV		Combination V	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Total	380	730	420	800	510	780	460	750	490	800
Average	38	73	42	80	51	78	46	75	49	80

Consider the mean values obtained for the minimum and maximum of each combination and calculate the range for each combination as

Range = Maximum value – Minimum value

The measure of importance for each combination is calculated as follows:

$$\text{Measure of Importance for a combination of attributes} = \frac{\text{Range for that combination}}{\text{Sum of the ranges for all the combinations}} \times 100$$

Tabulate the results as follows:

Combination	Max. Value	Min. Value	Range	Measure of Importance
I	73	38	35	21.875
II	80	42	38	23.750
III	78	51	27	16.875
IV	75	46	29	18.125
V	80	49	31	19.375
Sum of the ranges			160	100.000

Inference:

It is concluded that combination II is the most important one and combination III is the least important one.

APPROACHES FOR CONJOINT ANALYSIS

The following two approaches are available for conjoint analysis.

i. Multi-factor evaluation approach
ii. Two-factor evaluation approach

MULTI-FACTOR EVALUATION APPROACH IN CONJOINT ANALYSIS

Suppose a researcher has to analyze n factors. It is possible that each factor can assume a value in different levels.

Product Profile

A product profile is a description of all the factors under consideration, with any one level for each factor.

Suppose, for example, there are 3 factors with the levels given below.

Factor 1	:	3 levels
Factor 2	:	2 levels
Factor 3	:	4 levels

Then we have $3 \times 2 \times 4 = 24$ product profiles. For each respondent in the research survey, we have to provide 24 data sheets such that each data sheet contains a distinct profile. In each profile, the respondent is requested to indicate his preference for that profile in a rating scale of 0 to 10. A rating of 10 indicates that the respondent's preference for that profile is the highest and a rating of 0 means that he is not all interested in the product with that profile.

Example. Consider the product 'Refrigerator' with the following factors and levels:

Factor 1	:	capacity of 180 liters; 200 liters; 230 liters
Factor 2	:	number of doors: either 1 or 2
Factor 3	:	Price : Rs. 9000; Rs. 10,000; Rs. 12,000

Sample profile of the product

Profile Number	:	
Capacity	:	200 liters
Number of Doors	:	1
Price	:	Rs. 10,000
Rating of Respondent:		
(in the scale of 0 to 10)		

Steps in multi-factor evaluation approach:

1. Identify the factors or features of a product to be analyzed. If they are too many, select the important ones by discussion with experts.
2. Find out the levels for each factor selected in Step 1.

3. Design all possible product profiles. If there are n factors with levels L_1, L_2,…,L_n respectively, then the total number of profiles = $L_1L_2\ldots L_n$.
4. Select the scaling technique to be adopted for multi-factor evaluation approach (rating scale or ranking method)
5. Select the list of respondents using the standard sampling technique.

6. Request each respondent to give his rating scale for all the profiles of the product. Another way of collecting the responses is to request each respondent to award ranks to all the profiles: i.e., rank 1 for the best profile, rank 2 for the next best profile, etc.
7. For each factor profile, collect all the responses from all the participating respondents in the survey work.
8. With the rating scale awarded by the respondents, find out the score secured by each profile.
9. Tabulate the results in Step 8. Select the profile with the highest score. This is the most preferred profile.
10. Implement the most preferred profile in the design of a new product.

TWO-FACTOR EVALUATION APPROACH IN CONJOINT ANALYSIS

When several factors with different levels for each factor have to be analyzed, the respondents will have difficulty in evaluating all the profiles in the multi-factor evaluation approach. Because of this reason, two-factor evaluation approach is widely used in conjoint analysis.

Suppose there are several factors to be analyzed, with different levels of values for each factor.

Then we consider any two factors at a time with their levels of values. For each such case, we have a data sheet called a **two-factor table**. If there are n factors, then the number of such data sheets is $\binom{n}{2} = \frac{n(n-1)}{2}$.

Let us consider the example of 'Refrigerator' described in the multi-factor approach. For the two factors (i) capacity and (ii) price, we have the following **data sheet**.

Data Sheet (Two Factor Table) No:

Factor: Price of refrigerator

Factor: Capacity of Refrigerator	Price		
	Rs. 9,000	Rs. 10,000	Rs. 12,000
180 liters			
200 liters			
230 liters			

In this case, the data sheet is a matrix of 3 rows and 3 columns. Therefore, there are $3 \times 3 = 9$ places in the matrix. The respondent has to award ranks from 1 to 9 in the cells of the matrix. A rank of 1 means the respondent has the maximum preference for that entry and a rank of 9 means he has the least preference for that entry. Compared to multi-factor evaluation approach, the respondents will find it easy to respond to two-factor evaluation approach since only two factors are considered at a time.

Steps in two-factor evaluation approach:

1. Identify the factors or features of a product to be analyzed.
2. Find out the levels for each factor selected in Step 1.
3. Consider all possible pairs of factors. If there are n factors, then the number of pairs is $\binom{n}{2} = \frac{n(n-1)}{2}$. For each pair of factors, prepare a two-factor table, indicating all the levels for the two factors. If L_1 and L_2 are the respective levels for the two factors, then the number of cells in the corresponding table is L_1L_2.
4. Select the list of respondents using the standard sampling technique.
5. Request each respondent to award ranks for the cells in each two-factor table. i.e., rank 1 for the best cell, rank 2 for the next best cell, etc.
6. For each two-factor table, collect all the responses from all the participating respondents in the survey work.

7. With the ranks awarded by the respondents, find out the score secured by each cell in each two-factor table.
8. Tabulate the results in Step 7. Select the cell with the highest score. Identify the two factors and their corresponding levels.
9. Implement the most preferred combination of the factors and their levels in the design of a new product.

Application:

The two factor approach is useful when a manager goes for market segmentation to promote his product. The approach will enable the top level management to evolve a policy decision as to which segment of the market has to be concentrated more in order to maximize the profit from the product under consideration.

QUESTIONS

1. Explain the purpose of 'Factor Analysis'.
2. What is the objective of 'Conjoint Analysis'? Explain.
3. State the steps in the development of conjoint analysis.
4. State the applications of conjoint analysis.
5. Enumerate the advantages and disadvantages of conjoint analysis.
6. What is a 'product profile'? Explain.
7. What are the steps in multi-factor evaluation approach in conjoint analysis?
8. What is a 'two-factor table'? Explain.
9. Explain two-factor evaluation approach in conjoint analysis.

REFERENCES

Green, P.E. and Srinivasan, V., Conjoint Analysis in Consumer Research: Issues and Outlook, Journal of Consumer Research, 5, 1978, 103 – 123.

Green, P.E., Carrol, J. and Goldberg, A general approach to product design optimization via conjoint analysis, Journal of Marketing, 43, 1981, 17 – 35.

Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Pearson Education, Delhi, 2005.

Kanji, G.K., 100 Statistical Tests, Sage Publications, New Delhi, 1994.

Kothari, C.R., Quantitative Techniques, Vikas Publishing House Private Ltd., New Delhi, 1997.

Marrison, D.F., Multivariate Statistical Methods, McGraw Hill, New York, 1986.

Panneerselvam, R., Research Methodology, Prentice Hall of India, New Delhi, 2004.

Rencher, A.V., Methods of Multivariate Analysis, Wiley Inter-science, Second Edition, New Jersey, 2002.

Romesburg, H.C., Cluster Analysis for Researchers, Lifetime Learning Publications, Belmont, California, 1984.

Statistical Table-1: F-values at 1% level of significance

df_1: degrees of freedom for greater variance

df_2: degrees of freedom for smaller variance

df2/df1	1	2	3	4	5	6	7	8	9	10
1	4052.1	4999.5	5403.3	5624.5	5763.6	5858.9	5928.3	5981.0	6022.4	6055.8
2	98.5	99.0	99.1	99.2	99.2	99.3	99.3	99.3	99.3	99.3
3	34.1	30.8	29.4	28.7	28.2	27.9	27.6	27.4	27.3	27.2
4	21.1	18.0	16.6	15.9	15.5	15.2	14.9	14.7	14.6	14.5
5	16.2	13.2	12.0	11.3	10.9	10.6	10.4	10.2	10.1	10.0
6	13.7	10.9	9.7	9.1	8.7	8.4	8.2	8.1	7.9	7.8
7	12.2	9.5	8.4	7.8	7.4	7.1	6.9	6.8	6.7	6.6
8	11.2	8.6	7.5	7.0	6.6	6.3	6.1	6.0	5.9	5.8
9	10.5	8.0	6.9	6.4	6.0	5.8	5.6	5.4	5.3	5.2
10	10.0	7.5	6.5	5.9	5.6	5.3	5.2	5.0	4.9	4.8
11	9.6	7.2	6.2	5.6	5.3	5.0	4.8	4.7	4.6	4.5
12	9.3	6.9	5.9	5.4	5.0	4.8	4.6	4.4	4.3	4.2
13	9.0	6.7	5.7	5.2	4.8	4.6	4.4	4.3	4.1	4.1
14	8.8	6.5	5.5	5.0	4.6	4.4	4.2	4.1	4.0	3.9
15	8.6	6.3	5.4	4.8	4.5	4.3	4.1	4.0	3.8	3.8
16	8.5	6.2	5.2	4.7	4.4	4.2	4.0	3.8	3.7	3.6
17	8.4	6.1	5.1	4.6	4.3	4.1	3.9	3.7	3.6	3.5
18	8.2	6.0	5.0	4.5	4.2	4.0	3.8	3.7	3.5	3.5
19	8.1	5.9	5.0	4.5	4.1	3.9	3.7	3.6	3.5	3.4
20	8.0	5.8	4.9	4.4	4.1	3.8	3.6	3.5	3.4	3.3
21	8.0	5.7	4.8	4.3	4.0	3.8	3.6	3.5	3.3	3.3
22	7.9	5.7	4.8	4.3	3.9	3.7	3.5	3.4	3.3	3.2
23	7.8	5.6	4.7	4.2	3.9	3.7	3.5	3.4	3.2	3.2
24	7.8	5.6	4.7	4.2	3.8	3.6	3.4	3.3	3.2	3.1
25	7.7	5.5	4.6	4.1	3.8	3.6	3.4	3.3	3.2	3.1
26	7.7	5.5	4.6	4.1	3.8	3.5	3.4	3.2	3.1	3.0
27	7.6	5.4	4.6	4.1	3.7	3.5	3.3	3.2	3.1	3.0
28	7.6	5.4	4.5	4.0	3.7	3.5	3.3	3.2	3.1	3.0
29	7.5	5.4	4.5	4.0	3.7	3.4	3.3	3.1	3.0	3.0
30	7.5	5.3	4.5	4.0	3.6	3.4	3.3	3.1	3.0	2.9

Statistical Table-2: F-values at 2.5% level of significance

df_1: degrees of freedom for greater variance

df_2: degrees of freedom for smaller variance

df2/df1	1	2	3	4	5	6	7	8	9	10
1	647.7	799.5	864.1	899.5	921.8	937.1	948.2	956.6	963.2	968.6
2	38.5	39.0	39.1	39.2	39.2	39.3	39.3	39.3	39.3	39.3
3	17.4	16.0	15.4	15.1	14.8	14.7	14.6	14.5	14.4	14.4
4	12.2	10.6	9.9	9.6	9.3	9.1	9.0	8.9	8.9	8.8
5	10.0	8.4	7.7	7.3	7.1	6.9	6.8	6.7	6.6	6.6
6	8.8	7.2	6.5	6.2	5.9	5.8	5.6	5.5	5.5	5.4
7	8.0	6.5	5.8	5.5	5.2	5.1	4.9	4.8	4.8	4.7
8	7.5	6.0	5.4	5.0	4.8	4.6	4.5	4.4	4.3	4.2
9	7.2	5.7	5.0	4.7	4.4	4.3	4.1	4.1	4.0	3.9
10	6.9	5.4	4.8	4.4	4.2	4.0	3.9	3.8	3.7	3.7
11	6.7	5.2	4.6	4.2	4.0	3.8	3.7	3.6	3.5	3.5
12	6.5	5.0	4.4	4.1	3.8	3.7	3.6	3.5	3.4	3.3
13	6.4	4.9	4.3	3.9	3.7	3.6	3.4	3.3	3.3	3.2
14	6.2	4.8	4.2	3.8	3.6	3.5	3.3	3.2	3.2	3.1
15	6.1	4.7	4.1	3.8	3.5	3.4	3.2	3.1	3.1	3.0
16	6.1	4.6	4.0	3.7	3.5	3.3	3.2	3.1	3.0	2.9
17	6.0	4.6	4.0	3.6	3.4	3.2	3.1	3.0	2.9	2.9
18	5.9	4.5	3.9	3.6	3.3	3.2	3.0	3.0	2.9	2.8
19	5.9	4.5	3.9	3.5	3.3	3.1	3.0	2.9	2.8	2.8
20	5.8	4.4	3.8	3.5	3.2	3.1	3.0	2.9	2.8	2.7
21	5.8	4.4	3.8	3.4	3.2	3.0	2.9	2.8	2.7	2.7
22	5.7	4.3	3.7	3.4	3.2	3.0	2.9	2.8	2.7	2.7
23	5.7	4.3	3.7	3.4	3.1	3.0	2.9	2.8	2.7	2.6
24	5.7	4.3	3.7	3.3	3.1	2.9	2.8	2.7	2.7	2.6
25	5.6	4.2	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.6
26	5.6	4.2	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.5
27	5.6	4.2	3.6	3.3	3.0	2.9	2.8	2.7	2.6	

										2.5
28	5.6	4.2	3.6	3.2	3.0	2.9	2.7	2.6	2.6	2.5
29	5.5	4.2	3.6	3.2	3.0	2.8	2.7	2.6	2.5	2.5
30	5.5	4.1	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.5

Statistical Table-3: F-values at 5% level of significance

df_1: degrees of freedom for greater variance

df_2: degrees of freedom for smaller variance

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.5	230.1	233.9	236.7	238.8	240.5	241.8
2	18.5	19.0	19.1	19.2	19.2	19.3	19.3	19.3	19.3	19.3
3	10.1	9.5	9.2	9.1	9.0	8.9	8.8	8.8	8.8	8.7
4	7.7	6.9	6.5	6.3	6.2	6.1	6.0	6.0	5.9	5.9
5	6.6	5.7	5.4	5.1	5.0	4.9	4.8	4.8	4.7	4.7
6	5.9	5.1	4.7	4.5	4.3	4.2	4.2	4.1	4.0	4.0
7	5.5	4.7	4.3	4.1	3.9	3.8	3.7	3.7	3.6	3.6
8	5.3	4.4	4.0	3.8	3.6	3.5	3.5	3.4	3.3	3.3
9	5.1	4.2	3.8	3.6	3.4	3.3	3.2	3.2	3.1	3.1
10	4.9	4.1	3.7	3.4	3.3	3.2	3.1	3.0	3.0	2.9
11	4.8	3.9	3.5	3.3	3.2	3.0	3.0	2.9	2.8	2.8
12	4.7	3.8	3.4	3.2	3.1	2.9	2.9	2.8	2.7	2.7
13	4.6	3.8	3.4	3.1	3.0	2.9	2.8	2.7	2.7	2.6
14	4.6	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.6
15	4.5	3.6	3.2	3.0	2.9	2.7	2.7	2.6	2.5	2.5
16	4.4	3.6	3.2	3.0	2.8	2.7	2.6	2.5	2.5	2.4
17	4.4	3.5	3.1	2.9	2.8	2.6	2.6	2.5	2.4	2.4
18	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.5	2.4	2.4
19	4.3	3.5	3.1	2.8	2.7	2.6	2.5	2.4	2.4	2.3
20	4.3	3.4	3.0	2.8	2.7	2.5	2.5	2.4	2.3	2.3
21	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3
22	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3
23	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2
24	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2
25	4.2	3.3	2.9	2.7	2.6	2.4	2.4	2.3	2.2	2.2
26	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2
27	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2
28	4.1	3.3	2.9	2.7	2.5	2.4	2.3	2.2	2.2	2.1
29	4.1	3.3	2.9	2.7	2.5	2.4	2.3	2.2	2.2	2.1
30	4.1	3.3	2.9	2.6	2.5	2.4	2.3	2.2	2.2	2.1

Statistical Table-4: F-values at 10% level of significance

df_1: degrees of freedom for greater variance

df_2: degrees of freedom for smaller variance

df2/df1	1	2	3	4	5	6	7	8	9	10
1	39.8	49.5	53.5	55.8	57.2	58.2	58.9	59.4	59.8	60.1
2	8.5	9.0	9.1	9.2	9.2	9.3	9.3	9.3	9.3	9.3
3	5.5	5.4	5.3	5.3	5.3	5.2	5.2	5.2	5.2	5.2
4	4.5	4.3	4.1	4.1	4.0	4.0	3.9	3.9	3.9	3.9
5	4.0	3.7	3.6	3.5	3.4	3.4	3.3	3.3	3.3	3.2
6	3.7	3.4	3.2	3.1	3.1	3.0	3.0	2.9	2.9	2.9
7	3.5	3.2	3.0	2.9	2.8	2.8	2.7	2.7	2.7	2.7
8	3.4	3.1	2.9	2.8	2.7	2.6	2.6	2.5	2.5	2.5
9	3.3	3.0	2.8	2.6	2.6	2.5	2.5	2.4	2.4	2.4
10	3.2	2.9	2.7	2.6	2.5	2.4	2.4	2.3	2.3	2.3
11	3.2	2.8	2.6	2.5	2.4	2.3	2.3	2.3	2.2	2.2
12	3.1	2.8	2.6	2.4	2.3	2.3	2.2	2.2	2.2	2.1
13	3.1	2.7	2.5	2.4	2.3	2.2	2.2	2.1	2.1	2.1
14	3.1	2.7	2.5	2.3	2.3	2.2	2.1	2.1	2.1	2.0
15	3.0	2.6	2.4	2.3	2.2	2.2	2.1	2.1	2.0	2.0
16	3.0	2.6	2.4	2.3	2.2	2.1	2.1	2.0	2.0	2.0
17	3.0	2.6	2.4	2.3	2.2	2.1	2.1	2.0	2.0	2.0
18	3.0	2.6	2.4	2.2	2.1	2.1	2.0	2.0	2.0	1.9
19	2.9	2.6	2.3	2.2	2.1	2.1	2.0	2.0	1.9	1.9
20	2.9	2.5	2.3	2.2	2.1	2.0	2.0	1.9	1.9	1.9
21	2.9	2.5	2.3	2.2	2.1	2.0	2.0	1.9	1.9	1.9
22	2.9	2.5	2.3	2.2	2.1	2.0	2.0	1.9	1.9	1.9
23	2.9	2.5	2.3	2.2	2.1	2.0	1.9	1.9	1.9	1.8
24	2.9	2.5	2.3	2.1	2.1	2.0	1.9	1.9	1.9	1.8
25	2.9	2.5	2.3	2.1	2.0	2.0	1.9	1.9	1.8	1.8
26	2.9	2.5	2.3	2.1	2.0	2.0	1.9	1.9	1.8	1.8
27	2.9	2.5	2.2	2.1	2.0	2.0	1.9	1.9	1.8	1.8

28	2.8	2.5	2.2	2.1	2.0	1.9	1.9	1.9	1.8	1.8
29	2.8	2.4	2.2	2.1	2.0	1.9	1.9	1.8	1.8	1.8
30	2.8	2.4	2.2	2.1	2.0	1.9	1.9	1.8	1.8	1.8

UNIT V

LESSON: 1

STRUCTURE AND COMPONENTS OF RESEARCH REPORTS

Lesson Objectives:

- What is a Report?
- Characteristics of a good report
- Framework of a Report
- Practical Reports Vs Academic Reports
- Parts of a Research Report
- A note on Literature Review

Learning Objectives:

After reading this lesson, you should be able to :

- Understand the meaning of a research report
- Analyze the components of a good report
- Structure of a report
- Characteristic differences in Research Reporting

WHAT IS A REPORT?

A report is a written document on a particular topic, which conveys information and ideas and may also make recommendations. Reports often from the basis of crucial decision making. Inaccurate, incomplete and poorly written reports will fail to achieve their purpose and reflect on the decision, which will ultimately be made. This will also be the case if the report is excessively long, jargonistic and/ or structure less.

Good reports can be written by following these rules:

1. All points in the report should be clear to the intended reader.
2. The report should be concise with information kept to a necessary minimum and arranged logically under headings.
3. All information should be correct and supported by evidence.
4. All relevant material should be included in a complete report.

Purpose of Research Report:

1. Why am I writing this report? Do I want to inform/ explain/ persuade, or indeed all of these.
2. Who is going to read this report? Managers/ academicians/ researchers what do they know already? What do they need to know? Do any of them have certain attitudes or prejudices?
3. What resources do we have? Do I have access to a computer? Do I have enough time? Can any of my colleagues help?
4. Think about the content of your report – what am I going to put in it? What are my main themes? How much should be text, and how much should be illustrations?

Framework of a Report:

Various frameworks can be used depending on the content of the report, but generally the same rules apply. Introduction, method, results and discussion

with references and bibliography at the end and, an abstract at the beginning could from the framework.

STRUCTURE OF A REPORT:

Structure your writing around the IMRaD framework and you will ensure a beginning, middle and end to your report.

I	Introduction	Why did I do this research?	(beginning)
M	Method	What did I do and how did I go about doing it?	(middle)
R	results	What did I find?	(middle)

AND

D	Discussion	What does it all mean?	(end)

What do I put in the beginning part?

TITLE PAGE	Title of project Sub –title (where appropriate) Date Author Organization Logo
BACKGROUND	History(if any) behind project
ACKNOWLEDGEMENT	Author thanks people and organization who helped during the project
SUMMARY(sometimes called abstract of synopsis)	A condensed version of a report – outlines salient points, emphasis main conclusions and (where appropriate) the main recommendations. N.B this is often difficult to write and it is suggested that you write it last.
LIST OF CONTENTS	An at- a – glance list that tells the reader what is in the report and what page number(s) to find it on.
LIST OF TABLES	As above, specifically for tables.
LIST OF APPENDICES	As above, specifically for appendices.
INTRODUCTION	Author sets the scene and states his/ her intentions.
AIMA NAD OBJECTIVES	AIMS – general aims of the audit/ project, broad statement of intent. OBJECTIVES – specific things except to do/ deliver(e.g.

	expected outcomes)

What do I put in the middle part?

METHOD	Work steps; what was done – how, by whom, when?
RESULT/FINDINGS	Honest presentation of the findings, whether these were as expected or not. give the facts, including any inconsistencies or difficulties encountered

What do I put in the end part?

DISCUSSION	Explanation of the results.(you might like to keep the SWOT analysis in mind and think about your project's strengths, weakness, opportunities and threats, as you write)
CONCLUSIONS	The author links the results/ findings with the points made in the introduction and strives to reach clear, simply stated and unbiased conclusions. Make sure they are fully supported by evidence and arguments of the main body of your audit/project.
RECOMMENDATIONS	The author states what specifies actions should be taken, by whom and why. They must always like to the future and should always be realistic. Don't make them unless asked to.
REFERENCES	A section of a report, which provides full details of publications mentioned in the text, or from which extracts have been quoted.
APPENDIX	The purpose of an appendix is to supplement the information contained in the main body of the report.

PRACTICAL REPORTS VS ACADEMIC REPORT

Practical Reports:

In a practical world of business or government, a report convey a information and (sometimes) recommendations from a researcher who has investigated

a topic in detail. A report like this will usually be requested by people who need the information for a specific purpose and their request may be written in terms of reference or the brief. . whatever the report, it is important to look at the instruction for what is wanted.

A report like this differs from an essay in that it is designed to provide information which will be acted on, rather than to be read by people interested in the ideas for their own sake. Because of this, it has a different structure and layout.

Academic Reports:

A report written for an academic course can be thought of as a simulation. We can imagine that someone wants the report for a practical purpose, although we are really writing the report as an academic exercise for assessment. Theoretical ideas will be more to the front in an academic report than in a practical one.

Sometimes a report seems to serve academic and practical purposes. Students on placement with organizations often have to produce a report for the organization and for assessment on the course. Although the background work for both will be related, in practice, the report the student produces for academic assessment will be different from the report produced for the organization, because the needs of each are different.

RESEARCH REPORT: PRELIMINARIES

It is not sensible to leave all your writing until the end. There is always the possibility that it will take much longer than you anticipate and you will not have enough time. There could also be pressure upon available word processors as other students try to complete their own reports. It is wise to begin writing up some aspects of your research as you go along. Remember that you do not have

to write your report in the order than it will be read. Often it is easiest to start with the method section. Leave the introduction and the abstract to last. The use of a word processor makes it very straightforward to modigy and rearrange what you have written as your research progresses and your ideas change. The very process of writing will help your ideas to develop. Last but by no means least, ask someone to proofread your work.

STRUCTURE OF A RESEARCH REPORT

A research report has a different structure and layout to a project research. / A research report is for reference and is often quite a long document. It has to be clearly structured for the readers to quickly find the information wanted to come need to plan carefully to make sure that the information which has been gets put under the correct headings.

PARTS OF RESEARCH REPORT:

Cover sheet: this should contain some or all the following: full title of the report name of the research; the name of the unit of which the project is a part ; the name of the institution ; the date.

Title page: full title of the report. Your name

Acknowledgement: a thank you to the people who helped you.

Contents of table of contents

Headings and subheadings used in the report with their page numbers. Remember that each new chapter should begin on a new page. Use a consistent system in dividing the report into parts. The simplest may be to use chapters for each major part and subdivide these into sections and subsections. 1, 2, 3, etc, can be used as the numbers for each chapter. The sections for chapter 3 (for example) would be 3.1, 3.2, 3.3, and so on. For a further subdivision of a subsection you can use 3.2.1, 3.2.2, and so on.

Abstract or **Summary** or **Executive Summary** or **Introduction**

This is the overview of the whole report. It should let the reader see, in advance, what is in it. This includes what you set out to do, how reviewing literature focused and narrowed your research, the relation of the methodology you chose to your aims, a summary of your findings and of your analysis of the findings

BODY

Aims and Purpose or **Aims and Objectives**

Why did you do the work? What was the problem you were investigating? If you are not including a literature review, mention here the other research which is relevant to your work.

Literature Review: This should help to put your research into a background context and to explain its importance. Include only the books and articles which relate directly to your topic. Remember that you need to be analytical and critical and not just describe the works that you have read.

Methodology

Methodology deals with the methods and principles used in an activity, in this case research. In the methodology chapter you explain the method/s you used for the research and why you thought they were the appropriate ones. You may, for example, be doing mostly documentary research or you may have collected you own data. You should explain the methods of data collection, materials used, subjects interviewed, or places you visited. Give a detailed account of how and when you carried out your research and explain why you used the particular methods which you did use, rather than other methods. Included in this discussion should be an examination of ethical issues.

Results or Findings

What did you find out? Give a clear presentation of your results. Show the essential data and calculations here. You may want to use tables, graphs and figures.

Analysis and Discussion

Interpret your results. What do you make of them? How do they compare with those of others who have done research in this area? The accuracy of your measurements/results should be discussed and any deficiencies in the research design should be mentioned.

Conclusions

What do you conclude? You should summarize briefly the main conclusions which you discussed under "Results." Were you able to answer some or all of the questions which you raised in your aims? Do not be tempted to draw conclusions which are not backed up by your evidence. Note any deviation from expected results and any failure to achieve all that you had hoped.

Recommendations

Make your recommendations, **if required**. Positive or negative suggestions for either action or further research.

Appendix

You may not need an appendix, or you may need several. If you have used questionnaires, it is usual to include a blank copy in the appendix. You could include data or calculations, not used in the body, that are necessary, or useful,

to get the full benefit from your report. There may be maps, drawings, photographs or plans that you want to include. If you have used special equipment, you may want to include information about it.

The plural of an **appendix** is two or more **appendices** or **appendixes**. If an appendix or appendices are needed, design them thoughtfully in a way that your readers will find convenient to use.

Bibliography

List all the sources to which you refer in the body of the report. These will be referenced in the body of the text using the Harvard method. You may also list all the relevant sources you consulted even if you did not quote them.

A more confusing method is sometimes asked for in which you provide two lists of sources, one labelled "References" and the other "Bibliography". If you can avoid doing this, do so.

LITERATURE REVIEW

All investigations require for small projects this may not be in the form of a critical review of the literature, but this is often asked for and is a standard part of larger projects. Sometimes students are asked to produce a Literature Review on a topic as a piece of work in its own right. In its simplest form, a literature review is a list of relevant books and other sources, each followed by a description and comment on its relevance.

- A literature review should demonstrate that you have read and analysed literature **relevant** to your topic. From your reading you may get ideas about

methods of data collection and analysis If the review is part of a project, you will relate your reading to the issues in the project. As well a describing the reading, you should apply it to your topic.
A review should include only relevant items. The review should provide the reader with a picture of the state of knowledge in the subject. Your literature search should establish what previous research has been carried out in the subject area. Broadly speaking, there are three kinds of sources you will want to consult:

1. **introductory materials,**
2. journal articles
3. books.

To get a background idea of your topic you may wish to consult one or more textbooks at the appropriate level(s). As with must academic writing, it is a good idea to do your review in cumulative stages - That is, do not think you can do it all at once. **But keep a careful record of what you have searched**, how you have gone about it, and the exact citations and page numbers of your reading. Write notes as you go along. Record suitable notes on everything that you read, note methods of investigation. Make sure that you keep a full reference, complete with page numbers. You will have to find your own balance between taking notes that are too long and detailed and ones too brief to be of any use. It is best to write your notes in complete sentences and paragraphs, because research has shown that you are more likely to understand your notes later if they are written in a way that other people would understand. Keep your notes from different sources and/or about different points on separate index cards or on separate sheets of paper. You will do mainly basic reading while you are trying to decide on your topic. You may scan and make notes on the abstracts or summaries of work in the area. Then do a more thorough job of reading later on,

when you are more sure of what you are doing. If your project spans several months, it would be sensible towards the end to check whether there are any brand new useful references.

REFERENCES

There are many different methods of referencing your work; the most common perhaps is the Numbered Style, and the Harvard Method, with many other variations. We do not ask for any particular method, just use the one you are most familiar and most comfortable with. Also we do ask that you do reference your work.

THE PRESENTATION OF REPORT

Well-produced, appropriate illustrations enhance a report. With today's computer packages, almost anything is possible. However, histograms, bar charts and pie charts are still the three 'staples'. Readers like illustrated information because it is easier to absorb - and it's more memorable! Illustrations are useful ONLY when they are easier to understand than words or figures and they MUST BE relevant to the text. Use the *algorithm* included to help you decide whether or not to use an illustration. They should never be included for their own sake, and don't overdo it; too many illustrations will overwhelm your readers.

LESSON 2

TYPES OF REPORTS: CHARACTERISTICS OF GOOD RESEARCH REPORT

Lesson Outline:

- Different types of Reports
- Technical Reports
- General Reports
- Reporting Styles
- Characteristics of a Good Report

Learning Objectives:

After reading this lesson, you should be able to:

- Understand different types of reports
- Technical Reports and contents of them
- General Reports
- Different types of Writing styles
- Essential characteristics of a Good Report

Reports vary in length and type. Students study reports are often called as term papers, project reports, theses, dissertations depending on the nature of the report. Reports of Researchers are in the form of monographs, research papers, research thesis, etc. In business organizations a wide variety of reports are under use. Project reports, annual reports of financial statements, report of consulting groups, Project proposals, etc. News items in daily papers are also one form of report writing. In this lesson, let us identify different forms of reports and their major components

Types of Reports:

Reports may be categorized broadly as Technical Reports and General Reports based on the nature of methods, terms of reference and extent in-depth enquiry made, etc. On the basis of usage pattern, the reports may also be classified as Information oriented reports, decision oriented reports and research based reports. Further, kind of reports may also differ based on the communication situation. For example, the reports may be in the form of Memo, which is appropriate for informal situations or for short periods. On the other hand, the projects that extend over a period of time, often calls for project report. Thus, there is no standard format of reports. The most important thing that helps in classifying the reports the outline of its purpose and answers for the following questions:

- What did you do?
- Why did you choose the particular research methods you used?
- What did you learn and what are the implications of what you learned?
- If you are writing a recommendation report, what action are you recommending in response to what you learned?

Two types of report formats are described below:

A Technical Report:

A Technical report mainly focuses on methods employed, assumptions made while conducting a study, detailed presentation of findings and drawing inferences and comparisons with earlier findings based on the type of data drawn from the empirical work.

An outline of a Technical Report mostly consists of the following

Title and Nature of Study	: Brief title on the nature of work some times followed by subtitle to indicate more appropriately either the method or tools used. Description of objectives of the study, research design, operational terms, working hypothesis, type of analysis and data required.
Abstract of Findings	: A brief review of the main findings just either in a paragraph or in one/two pages.
Review of current status	: A quick review of past observations and contradictions reported, applications observed and reported be reviewed based on the in-house resources or based on published observations
Sampling and Methods employed	: Specific methods used in the study and their limitations. In case of experimental methods, the nature of subjects, control conditions are to be specified. In case of sample studies, details about the sample design, i.e., sample size, sample selection, etc
Data sources and experiment conducted	: Sources of data, their characteristics and limitations to be specified. In case of primary survey, the manner in which data has been collected to be described.

Analysis of data and tools used.	: The analysis of data and presentation of findings of the study with supporting data in the form of tables and charts be narrated. This constitutes the major component of the research report
Summary of findings	: A detailed summary of findings of the study and the major observations be stated. Decision inputs if any, policy implications from the observations be specified
Bibliography	: A brief list of studies conducted in similar lines, either preceding the present study or conducted under different experimental conditions be listed
Technical appendices	: These appendices include the design of experiments or questionnaires used in conducting the study, mathematical derivations, elaboration on particular techniques of analysis, etc.

General Reports :

General reports often relates a popular policy issues mostly related to social issues. These reports are generally simple, less technical, good use of tables and charts. Most often they reflect the journalistic style. Examples for this type of report is the " Best B-Schools survey in Business Magazines. The outline of these reports is as follows:

1. Major Finding and its implication
2. Recommendations for Action
3. Objectives of the Study
4. Method employed in collecting data
5. Results

Writing Styles:

There are at least 3 distinct report writing styles that can be applied by students of Business Studies. They are called:

i. Conservative*
ii. Key points*
iii. Holistic

i. Conservative Style

Essentially, the conservative approach takes the best structural elements from essay writing and integrates these with appropriate report writing tools. Thus headings would be used to deliberate different sections of the answer. In addition, space would be well utilised by ensuring that each paragraph is distinct (perhaps separated from other paragraphs by leaving two blank lines in between).

ii. Key Point Style

This style utilises all of the report writing tools and is thus more overtly 'report-looking'. Use of headings, underlining, margins, diagrams and tables are common. Occasionally reporting might even use indentation and dot points.

The important thing to remember is that the tools should be applied in a way that adds to the report. The question must be addressed and the tools applied should assist in doing that.An advantage of this style is the enormous amount of information that can be delivered relatively quickly.

iii. Holistic Style

The most complex and unusual of the styles, holistic report writing aims to answer the question from a thematic and integrative perspective. This style of report writing requires that

researcher to have a strong understanding of the course and are able to see which outcomes are being targeted by the question.

Essentials of a good report :

Good research report should satisfy some of the following basic characteristics

STYLE

Reports should be easy to read and understand. The style of the writer should ensure that sentences are succinct and the language used simple, to the point and avoiding excessive jargon.

LAYOUT

A good layout enables the reader to follow the report's intentions and aids the communication process. Sections and paragraphs should be given headings and subheadings. You may also consider a system of numbering or lettering to identify the relative importance of paragraphs and subparagraphs. Bullet points are an option for highlighting important points in your report.

ACCURACY

Make sure everything you write is factually accurate. If you mislead, misinform or unfairly persuade your readers, you will be doing a disservice not only to yourself but also to your practice/ health centre etc and your credibility will be destroyed. Remember to reference any information you have used to support your work.

CLARITY

Take a break from writing. When you come back to it you'll have that degree of objectivity that you need. Remember tell them what you're going to say, say it, and then tell them you said it.

READABILITY

Experts agree that the factors, which most affect readability, are:

> Attractive appearance
> Non-technical subject matter
> Clear and direct style
> Short sentences
> Short and familiar words

REVISION

When the first draft of the report is completed, it should be put to one side or at least 24 hours. The report should then be read as if with eyes of the intended reader. It should be checked for spelling and grammatical errors. Remember the spell and grammar check on your computer. Use it!

REINFORCEMENT

Usually gets the message across. This old adage is well known and is used to good effect in all sorts of circumstances, e.g. presentations - not just report writing.

> TELL THEM WHAT YOU ARE GOING TO SAY: in the introduction and summary you set the scene for what follows in your report.
> THEN SAY IT : you spell things out in results/findings
> THEN TELL THEM WHAT YOU SAID: you remind your readers through the discussion what it was all about.

REFERENCES

There are many different methods of referencing your work; the most common perhaps is the Numbered Style, and the Harvard Method, with many other variations. We do not ask for any particular method, just use the one you are most familiar and most comfortable with. Also we do ask that you do reference your work.

FEEDBACK MEETING

It is useful to circulate copies of your report prior to the feedback meeting. Meaningful discussion can then take place during the feedback meeting with recommendations for change more likely to be agreed upon which can then be included in your conclusion.

The following questions should be asked at this stage to check whether the Report served the purpose:

> Does the report have impact?
> Does the summary /abstract do justice to the report?
> Does the introduction encourage the reader to read more?
> Is the content consistent with the purpose of the report?
> Have the objectives been met?
> Is the structure logical and clear?
> Have the conclusions been clearly stated?
> Are the recommendations based on the conclusions and expressed clearly and logically?

USING ILLUSTRATIONS TO IMPROVE THE PRESENTATION OF YOUR REPORT

Well-produced, appropriate illustrations enhance a report. With today's computer packages, almost anything is possible. However, histograms, bar charts and pie charts are still the three 'staples'.

Readers like illustrated information because it is easier to absorb - and it's more memorable! Illustrations are useful ONLY when they are easier to understand than words or figures and they MUST BE relevant to the text. Use the *algorithm* included to help you decide whether or not to use an illustration. They should never be included for their own sake, and don't overdo it; too many illustrations will overwhelm your readers.

UNIT V

LESSON 3

FORMAT AND PRESENTATION OF A REPORT

Lesson Outline:

- *Importance of Presentation of a Report*
- *Common Elements of a Format*
- *Title Page*
- *Introductory Pages*
- *Body of the Text*
- *References*
- *Appendix*
- *Dos and Do n'ts*
- *Presentation of Reports*

Learning Objectives:
After reading this Lesson, you should be able to :

- *Understand the importance of Format of a Report*
- *Contents of a Title Page*
- *What should be in Introductory pages*
- *Contents of a Body Text*
- *How to report other studies*
- *Contents of an Appendix*
- *Dos and Don'ts a Report*

Any report serves its purpose, if it is finally presented before the stake holders of the work. It is an MBA student Project Work in a Industrial enterprise, the findings of the study would be more relevant, if they were presented before the internal managers of the company. In case of reports prepared out of consultancy projects, a presentation would help the users to

interact with the research team and get greater clarification on any issue of their interest. Business Reports, Feasibility Reports do need a summary presentation, if they have to serve the intended purpose. Finally, the Research Reports of the scholars would help in achieving the intended academic purpose, if they are made public in academic symposiums, seminars or in Public Viva Voce examinations. Thus, the presentation of a report goes along with preparation of good report. Further, the use of Graphs, Charts and citations, pictures would definitely draw the attention of audience of any time. In this lesson, it is intended to provide a general outline relating the presentation of any type of report. See Exhibit I

Exhibit I

Common Elements of a Report

A report may contain some or all of the following, please refer to your departmental guidelines.

MEMORANDUM OR COVERING LETTER

A brief note stating the purpose of or giving an explanation for something. Used when the report is sent to someone within the same organization.

TITLE PAGE

Addressed to the receiver of a report giving an explanation for it. Used when the report is for someone who does not belong to the same organization as the writer.

Contains a descriptive heading or name, may also contain author's name, position, company name and so on.

EXECUTIVE SUMMARY

Summarizes the main contents. Usually 300-350words.

TABLE OF CONTENTS

A list of the main sections, indicating the page on which each section begins.

INTRODUCTION

Informs the reader of what the report is about—aim and purpose, significant issues, any relevant background information.

DISCUSSION

Describes reasoning and research in detail.

CONCLUSION/S

Summarizes the main points made in the written work. It often includes an overall answer to the problem addressed; or an overall statement synthesizing the strands of information dealt with.

RECOMMENDATION/S

Gives suggestions relating to the issue(s) or problem(s) dealt with.

REFERENCES

An alphabetical list of all sources referred to in the report.

APPENDICES

Extra information of further details placed after the main body of the text.

FORMATS OF REPORTS:

Before attempting to look into Presentation dimensions of a Report, a quick look into standard format associated with a Research Report is examined

hereunder. The format generally includes the steps one should follow while writing and finalizing their research report.

Different Parts of a Report

Generally different parts of a report include:

1. Cover Page / Title Page
2. Introductory Pages (Foreward, Preface, Acaknolwedgement, Table of contents, List of Tables, List of Illustrations or Figures, Key words / Abrevations used etc)
3. Contents of the Report (Which generally includes a Macro setting, Research Problem, Methodology used, Objectives of the study, Review of studies, Data tools used, Empirical results in one/two sections, Summary of Observations, etc)
4. References (including Bibliography, Appendices, Glossary of terms used, Source data, Derivations of Formulas for Models used in the analysis, etc)

Title Page:

The Cover page or Title Page of a Research Report should contain the following information:

1. **Title of the Project / Subject**
2. **Who has conducted the study**
3. **For What purpose**
4. **Organization**
5. **Period of submission**

A Model:

An example of a Summer Project Report conducted by an MBA student generally follows the following Title Page

---

A STUDY ON THE USE OF COMPUTER TECHNOLOGY IN BANKING OPERATIONS IN XXX BANK LTD., PONDICHERRY

A SUMMER PROJECT REPORT

PREPARED BY
Ms MADAVI LATHA

Submitted at

SCHOOL OF MANAGEMENT
PONDICHERRY UNIVERSITY
PONDICHERRY – 605 014
2006

Introductory Pages:

Introductory pages generally does not constitute the Write up of the Research work done. These introductory pages basically form the Index of the work done. These pages are usually numbered in Roman numerical (eg, I, ii, iii, etc). The introductory pages include the following components

- Foreword
- Preface
- Acknowledgements
- Table of contents
- List of tables
- List of Figures / Charts

Foreword is usually one page write up or a citation about the work by any eminent / popular personality or a specialist in the given field of study. Generally the write up include a brief background on the contemporary issues and the suitability of the present subject and its timelyness, major highlights of the present work, brief background of the author, etc. The writer of the Foreward generally gives this Foreword on his letter head

Preface is again one/two pages write up by the author of the book / report stating circumstances under which the present work is taken up, importance of the work, major dimensions examined and intended audience for the given work. The author gives his signature and address at the bottom of the page along with date and year of the work

Acknowledgements is a short section, mostly a paragraph. It mostly consists of sentences giving thanks for all those associated and encouraged to carry out the present work. Generally authors takes time to acknowledge the liberal funding by any funding agencies to carry out the work, agencies given permission to use their resources, etc. At the end, the authors thanks every body and gives his signature

Table of Contents refers to the index of all pages of the said Research Report. These contents provide the information about the chapters, sub sections, annexure for each chapter, if any, etc. Further, the page numbers of each content of the report greatly helps any one to refer to those pages for necessary details. Most authors use different forms while listing the sub contents. These include alphabet classification and decimal classification. An example for both of them are given below

Example of content sheet (alphabet classification)

CONTENTS

Foreword	i
Preface	ii
Acknowledgement	iv
Chapter I (Title of the Chapter) INTRODUCTION	
A. Macro Economic Background	1
B. Performance of a specific industry sector	6
C. Different studies conducted so far	9
D. Nature and Scope	17
1. Objectives of the study	18
2. Methodology adopted	19
a. Sampling Procedure adopted	20
b. Year of the study	20
Chapter II (Title of the Chapter) : Empirical Results I	22
A. Test results of H1	22
B. Test Results of H2	27
C Test Results of H3	32
1. Sub Hypothesis of H3	33
2. Sub Hypothesis of H2	37
Chapter III	45
Chapter IV	85
Chapter V (Summary & Conclusions)	120
Appendices	132
Bibliography	135
Glossary	140

An example of Content Sheet with decimal classification

CONTENTS

Foreword	i
Preface	iii
Acknowledgement	v

Chapter I (Title of the Chapter) INTRODUCTION

1. Macro Economic Background	1
2. Performance of a specific industry sector	6
3. Different studies conducted so far	9
4 Nature and Scope	17
4.1. Objectives of the study	18
4.2. Methodology adopted	19
4.2. a. Sampling Procedure adopted	20
4.2.b. Year of the study	20
Chapter II (Title of the Chapter) : Empirical Results I	22
1. Test results of H1	22
2. Test Results of H2	27
3 Test Results of H3	32
3.1. Sub Hypothesis of H3	33
3.2. Sub Hypothesis of H2	37
Chapter III	45
Chapter IV	85
Chapter V (Summary & Conclusions)	120
Appendices	132
Bibliography	135
Glossary	140

List of Tables and Charts - Details of Charts and Tables given in the research Report are numbered and presented in separate pages and the list of such tables and Charts are given in a separate page. Tables are generally numbered either in Arabic numerals or in decimal form. In case of decimal form, it is possible to indicate the chapter to which the said table belongs to. For example, Table 2.1 refers to Table 1 in Chapter 2.

Executive Summary : Most Business Reports or Project works conducted on a specific issue, carries one or two pages of Executive Summary. This summary

precedes the Chapters of the Regular Research Report. This summary generally contains a brief description of problem under enquiry, methods used and the findings. A line about the possible alternatives for decision making would be the last line of the Executive Summary.

BODY OF THE REPORT:

The body of the Report is the most important part of the report. This body of report may be segmented into a handful of Units / Chapters arranged in a sequential order. Research Report often present the Methodology, Objectives of the study, Data tools, etc in the first/ second chapters along with a brief background of the study, review of relevant studies.

The major findings of the study are incorporated into two or three chapters based on the major or minor hypothesis tested or based on the sequence of objectives of the study. Further, the chapter plan may also likely to base on different dimensions of the problem under enquiry.

Each Chapter may be divided into sections. While the first section may narrate the descriptive characteristics of the problem under enquiry, the second and subsequent sections may focus on empirical results based deeper insights of the problem of study. Each chapter based on Research Studies mostly contain Major Headings, Sub headings, quotations drawn from observations made by earlier writers, footnotes and exhibits

Use of References:

There are two types of reference formatting. The first is the 'in-text' reference format, where previous researchers and authors are cited during the building of arguments in the Introduction and Discussion sections. The second type of format is that adopted for the Reference section for writing food notes or Bibliography.

Citations in the text

The names and dates of researchers go in the text as they are mentioned, e.g. "This idea has been explored in the work of Smith (1992)." It is generally unacceptable to refer to authors and previous researchers, etc

Examples of Citing References Single author

Duranti (1995) has argued **or** It has been argued that (Duranti, 1995)

In case of More authors,

Moore, Maguire, and Smyth (1992) proposed **or** It has been proposed that (Moore, Macquire, & Smyth, 1992)

For subsequent citations in the same report: Moore et al.(1992) also proposed... **or** It has also been proposed that.... (Moore et al., 1992)

The reference section

The end of report reference section comes immediately after the Discussion and is begun on a new page. It is headed 'References' in upper and lower case letters centered across the page. Psychology reports should only include reference sections, not bibliographies.

Published journal articles

Beckerian, D.A. (1993). In search of the typical eyewitness. *American Psychologist, 48,* 574-576.

Gubbay, S.S., Ellis, W., Walton, J.N., & Court, S.D.M. (1965). Clumsy children: A study of apraxic and agnosic defects in 21 children. *Brain, 88,* 295-312.

Authored Books

Cone, J.D., & Foster, S.L. (1993). *Dissertations and theses from start to finish: Psychology and related fields.* Washington, DC: American Psychological Association.

Cone, J.D., & Foster, S.L. (1993). *Dissertations and theses from start to finish: Psychology and related fields* (2nd ed.). Washington, DC: American Psychological Association.

APPENDICES

Your report should be sufficiently detailed that the reader should never have to refer to the appendices to know what happened in your study, what questions were asked of your participants and/or what you found. Rather the purpose of the appendices is to supplement the main body of your text and provide additional information that may be of interest to the reader.

There is no major heading for the Appendices. You simply need to include each one, starting on a new page, numbered using capital letters, and headed with a centered brief descriptive title, for example:

Appendix A: List of stimulus words presented to participants

Dos and Don'ts of Report Writing

1. Choose a font size that is not too small or too large; 11 or 12 is a good font size to use.

2. Acknowledgment need not be a separate page, except in the final report. In fact, you could just drop it altogether for the first- and second-stage reports. Your guide already knows how much you appreciate his/her support. Express your gratitude by working harder instead of writing a flowery acknowledgment!

3. Make sure your paragraphs have some indentation and that it is not too large. Refer to some text books or journal papers if you are not sure.

4. If figures, equations, or trends are taken from some reference, the reference must be cited right there, even if you have cited it earlier.

5. The correct way of referring to a figure is Fig. 4 or Fig. 1.2 (note that there is a space after Fig.). The same applies to Section, Equation, etc. (e.g., Sec. 2, Eq. 3.1).

6. Cite a reference as, for example, "The threshold voltage is a strong function of the implant dose [1]." Note that there must be a space before the bracket.

7. Follow some standard format while writing references. For example, you could look up any IEEE transactions issue and check out the format for journal papers, books, conference papers, etc.

8. Do not type references (for that matter, any titles or captions) entirely in capital letters. About the only capital letters required are (i) the first letter of a name, (ii) acronyms, (iii) the first letter of the title of an article (iv) the first letter of a sentence.

9. The order of references is very important. In the list of your references, the first reference must be the one which is cited before any other reference, and so on. Also, every reference in the list must be cited at least once (this also applies to figures). In handling references and figure numbers, Latex turns out to be far better than Word.

12. Many commercial packages allow "screen dump" of figures. While this is useful in preparing reports, it is often very wasteful (in terms of toner or ink) since the background is black. Please see if you can invert the image or use a plotting program with the raw data such that the background is white.

    The following tips may be useful: (a) For Windows, open the file in Paint and select Image/Invert Colors. (b) For Linux, open the file in Image Magick (this can be done by typing display&) and then selecting Enhance/Negate.

14. As far as possible, place each figure close to the part of the text where it is referred to.

15. A list of figures is not required except for the final project report. It generally does not do more than wasting paper.

16. The figures, when viewed together with the caption, must be, as far as possible, self-explanatory. There are times when one must say, "see text for details". However, this is an exception and not a rule.

17. The purpose of a figure caption is simply to state what is being presented in the figure. It is not the right place for making comments or comparisons; that should appear only in the text.

18. If you are showing comparison of two (or more) quantities, use the same notation through out the report. For example, suppose you want to compare measured data with analytical model in four different figures. In each figure, make sure that the measured data is rep resented by the same line type or symbol. The same should be followed for the analytical model. This makes it easier for the reader to focus on the important aspects of the report rather than getting lost in lines and symbols.

19. If you must resize a plot or a figure, make sure that you do it simultaneously in both x and y directions. Otherwise, circles in the original figure will appear as ellipses, letters will appear too fat or too narrow, and other similar calamities will occur.

20. In the beginning of any chapter, you need to add a brief introduction and then start sections. The same is true about sections and subsections. If you have sections that are too small, it only means that there is not enough material to make a separate section. In that case, do not make a separate section! Include the same material in the main section or elsewhere.

    Remember, a short report is perfectly acceptable if you have put in the effort and covered all important aspects of your work. Adding unnecessary sections and subsections will create the impression that you are only covering up the lack of effort.

22. Do not make one-line paragraphs.

23. Always add a space after a full stop, comma, colon, etc. Also, leave a space before opening a bracket. If the sentence ends with a closing bracket, add the full stop (or comma or semicolon, etc) after the

bracket.

24. Do not add a space before a full stop, comma, colon, etc.

25. Using a hyphen can be tricky. If two (or more) words form a single adjective, a hyphen is required; otherwise, it should not be used. For example, (a) A short-channel device shows a finite output conductance. (b) This is a good example of mixed-signal simulation. (c)Several devices with short channels were studied.

26. If you are using Latex, do not use the quotation marks to open. If you do that, you get "this". Use the single opening quotes (twice) to get "this".

27. Do not use very informal language. Instead of "This theory should be taken with a pinch of salt," you might say, "This theory is not convincing," or "It needs more work to show that this theory applies in all cases."

28. Do not use "&"; write "and" instead. Do not write "There're" for "There are" etc.

29. If you are describing several items of the same type (e.g., short-channel effects in a MOS transistor), use the "list" option; it enhances the clarity of your report.

30. Do not use "bullets" in your report. They are acceptable in a presentation, but not in a formal report. You may use numerals or letters instead.

31. Whenever in doubt, look up a text book or a journal paper to verify whether your grammar and punctuation are correct.

32. Do a spell check before you print out your document. It always helps.

33. Always write the report so that the reader can easily make out what your contribution is. Do not leave the reader guessing in this respect.

34. Above all, be clear. Your report must have a flow, i.e., the reader must be able to appreciate continuity in the report. After the first reading, the reader should be able to understand (a) the overall theme and (b) what is new (if it is a project report).

35. Plagiarism is a very serious offense. You simply cannot copy material

from an existing report or paper and put it verbatim in your report. The idea of writing a report is to convey in your words what you have understood from the literature.

The above list may seem a little intimidating. However, if you make a sincere effort, most of the points are easy to remember and practice. A supplementary exercise that will help you immensely is that of looking for all major and minor details when you read an article from a newspaper or a magazine, such as grammar, punctuation, organization of the material, etc

PRESENTATION OF A REPORT

In this section, we will look into the issues associated with presentation of a Research Report by the Researcher or principal investigator. While preparing for the presentation of a report, the researchers have to focus on the following issues:

1. What is the purpose of the report and issues on which the Presentation has to focus?
2. Who are the stake holders and their areas of interest
3. The mode and media of presentation
4. Extent of Coverage and depth to address at
5. Time, Place and cost associated with presentation
6. Audio – Visual aids intended to be used