MALLA REDDY ENGINEERING COLLEGE

(AN UGC Autonomous Institution Approved by AICTE New Delhi & Affiliated to JNTU, Hyderabad
Accredited by NAAC with 'A++' Grade (cycle III) NBA Tier -I Accredited
IIC-Four star Rating, NIRF Ranking 210-250, RIIA Brnd Performer
Maisammaguda(H),Medchal -Malkajgiri District, Secunderabad, Telangana State-500100,www.mrec.ac.in

Department of Master of Business Administration

E-Content File

I MBA I Semester

Subject

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

Code: C1E05

Academic Year 2023-24

Regulations: MR22

MBA-I Semester **Paper Code: AB105**

Research Methodology and Statistical Analysis

UNIT – I

Meaning Of Research:

Research in simple terms refers to search for knowledge. It is a scientific and systematic search for information on a particular topic orissue. 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, concepts or symbols for the purpose of generalizing to extend, correct or verify knowledge, whether that knowledge aids in the construction of theory or in the practice of an art".

According to Redman and Mory (1923), 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 hypotheses 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 formulated hypotheses".

Thus, research is an original addition to the available knowledge, which contributes to its further advancement. It is an attempt to pursuetruth 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.

Objectives Of Research:

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

1. To gain familiarity with 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 the hypothesis of a causal relationship between two variables (i.e., hypothesis-testing research studies).

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 that the researchers adopt for conducting the research studies.

on the other hand, research methodology is the way in which research problems are solved systematically. 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.

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 subjectiveassessment 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 are either in non-quantitativeform or in the form which cannot be put to rigorous quantitative analysis. Usually, this approach uses techniques like indepth interviews, focus group interviews, and projective techniques.

Types Of Research:

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

1. Descriptive Versus Analytical:

Descriptive research consists of 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 postfacto 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 ofthe 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 Versus Fundamental:

Research can also be applied or fundamental in nature. An attempt to find a solution to an immediate problem encountered by a firm, anindustry, 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 thealready existing organized body of scientific knowledge.

3. Quantitative Versus Qualitative:

Quantitative research relates to aspects that can be quantified or can be expressed in terms of quantity. It involves the measurement ofquantity or amount. Various available statistical and econometric methods are adopted for analysis in such research. Which includes correlation, regressions and time series analysis etc,.

On the other hand, Qualitative research is concerned with qualitative phenomena, or more specifically, the aspects related to or involving quality or kind. For example, an important type of qualitative research is 'Motivation Research', which investigates into the reasons for certain human behaviour. The main aim of this type of research isdiscovering the underlying motives and desires of human beings by 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 helps 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 Versus Empirical:

The 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 the observation or experience with hardly any regard for theory and system. Such researchis data based, which often comes up with conclusions that can be verified through experiments or observation. Empirical research is also known as experimental type of research, in which it is important to first collect the facts and their sources, and actively take steps to stimulate the production of desired information. In this type of research, the researcher first formulates a working hypothesis, and then gathers sufficient facts to prove or disprove the stated hypothesis. He/she formulates the experimental design, which according to him/her would manipulate the variables, so as to obtain the desired information. This type of research is thus characterized by the researcher's control over the variables under study. In simple term, empirical research is most appropriate when an attempt is made to provethat certain variables influence the other variables in some way. Therefore, the results obtained by using the experimental or empirical studies are considered to be the most powerful evidences for a given hypothesis.

5. Other Types Of Research:

The remaining types of research are variations of one or more ofthe afore-mentioned type of research. They vary in terms of the purposeof research, or the time required to complete it, or may be based on some

other similar factor. On the basis of time, research may either be in the nature of one-time or longitudinal time series 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 can 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 method may be employed to analyse the basic causal relations. These studies usually undertake a detailed in-depth analysis of the causesof certain events of interest, and use very small samples and sharp data collection methods. The research may also be explanatory in nature. Formalized research studies consist of substantial structure and specifichypotheses to be verified. As regards to historical research, sources like historical documents, remains, etc. Are utilized to study past events orideas. It also includes philosophy of persons and groups of the past or any remote point of time.

Research has also been classified into decision-oriented and conclusion-oriented categories. The decision-oriented research is always carried out as per the need of a decision maker and hence, the researcher hasno freedom to conduct the research according to his/her own desires. On the other hand, in the case of 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. Operations research is a kind of decision-oriented research, where in scientific method is used in providing the departments, a quantitative basis for decision-making with respect to the activities under their purview.

Importance Of Knowing How To Conduct Research:

The importance of knowing how to conduct research are listed below:

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 inculcates the ability to evaluate and utilize the research findings with confidence;

iii. The knowledge of research methodology equips the researcher with the tools that help him/her to make the observations objectively;and

iv. The knowledge of methodology helps the research consumers to evaluate research and make rational decisions.

Qualities Of A Researcher:

It is important for a researcher to possess certain qualities to conduct research. First and foremost, he 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 cited by (Wilkinson and Bhandarkar, 1979) identified a few distinct qualities of a scientist. According to him, a true research scientist should possess the following qualities:

(1) First of all, the nature of a researcher must be of the temperamentthat 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. The difficulty may arise dueto untrained eye, which fails to see anything beyond what it has the powerof seeing and sometimes even less than that. This may also be due to thelack of discipline in the method of science. An unscientific individualoften remains satisfied with the expressions like approximately, almost,or nearly, which is never what nature is. A real research cannot see two things which differ, however minutely, as the same.

(2) A researcher must possess an alert mind. 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. An individual who is ignorant or not alert and receptive during his research will notmake a good researcher. He will fail as a good researcher if he has no keen eyes or mind to observe the unusual changes behind the routine. Research

demands a systematic immersion into the subject matter by the researcher grasp even the slightest hint that may culminate into significant research problems. In this context, Cohen and Negal cited by (Selltiz et al, 1965; 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".

(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 process of conducting research is not an easytask. There are occasions when a research scientist might feel defeated or completely lost. This is the stage when a researcher would need immense courage and the sense of conviction. The researcher must learn the art of enduring intellectual hardships. In the words of Darwin, "It's dogged that does it".

In order to cultivate the afore-mentioned three qualities of a researcher, a fourth one may be added. This is the quality of makingstatements 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.

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". It brings out the significance of research, increased amount of which makes the 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 policy-makers, althoughin itself 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 are engaged by the governmentthese days to undertake this work. Thus, research as a tool of government economic policy formulation involves three distinct stages of operation:
(i) investigation of economic structure through continual compilationof facts; (ii) diagnosis of events that are taking place and analysis of the forces underlying them; and (iii) the prognosis i.e., the prediction of future developments (Wilkinson and Bhandarkar, 1979).

Research also assumes significance in solving various operational and planning problems associated with business and industry. In several ways, operations research, market research and motivational researchare vital and their results assist in taking business decisions. Market research 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 applicationof 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 analysis of the motivations underlying consumer behaviour. All these researches are very useful for business and industry, and are responsible for business decision-making.

Research is equally important to social scientists for analyzing the social relationships and seeking explanations to various social problems. It gives intellectual satisfaction of knowing things for the sakeof knowledge. It also possesses the practical utility for the social scientistto gain knowledge so as to be able to do something better or in a more

efficient manner. The 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. 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.

Hypothesis:

"Hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified groupof phenomena either asserted merely as a provisional conjecture to guide some investigation in the light of established facts" (Kothari, 1988). A research hypothesis is quite often a predictive statement, which is capableof 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 do 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 examineits validity.

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 canbe confirmed or disproved by observation" (Kothari, 1988).

iii. A hypothesis must state relationship between two variables, in thecase 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 researcher. And therefore, he/she must formulate such hypotheses.

v. As far as possible, a hypothesis must be stated in the simplest 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 knownfacts. In other words, it should be consistent with a substantial bodyof established facts. That is, it must be in the form of a statementwhich is 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 it if it cannot be tested within a givenperiod of time, as no one can afford to spend a life-time on collecting data to test it.

viii. A hypothesis should state the facts that give 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.

Concepts Relating To Testing Of Hypotheses:

Testing of hypotheses requires a researcher to be familiar with various concepts concerned with it such as:

1) Null Hypothesis And Alternative Hypothesis:

In the context of statistical analysis, hypotheses are 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 thatboth the methods are equally good, then such a statement is called as thenull 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 $\mathbf{H_0}$, whilethe 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. Symbolicallyit may be written as:-

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

If sample results do not support this null hypothesis, then it shouldbe concluded that something else is true. The conclusion of rejecting thenull hypothesis is called as alternative hypothesis H_1. 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:

Alternative hypothesis	To be read as follows
H_1: $\mu \neq \mu H_0$	The alternative hypothesis is that the population mean is not equal to 100, i.e., it could be greater than or less than 100
$H_1 : \mu > \mu H_0$	The alternative hypothesis is that the population mean is greater than 100
$H_1 : \mu < \mu H_0$	The alternative hypothesis is that the population mean is less than 100

Before the sample is drawn, the researcher has to state the nullhypothesis 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 wishesto prove, whereas the null hypothesis is the one which he/she wishesto disprove. Thus, a null hypothesis is usually the one which a researcher tries to reject, while an alternative hypothesis is the onethat 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 thenthe 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 shouldnot 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 insight. 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 otherwords, 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 whichis usually determined prior to testing the hypothesis.

3) Test Of Hypothesis Or Decision Rule:

Suppose the given hypothesis is H_0 and the alternative hypothesis H_1, 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_1 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 researcher 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 Ierror is denoted by α (alpha) and is known as α error, while Type II erroris 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 valuethat is not equal to the specified value of the null hypothesis.

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:

1) 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 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, the hypotheses may be stated as follows:

Null Hypothesis H_0 : = 100 Alternative Hypothesis H_1 : > 100

2) 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 levelsof 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 inthe context of the nature and purpose of enquiry.

3) Deciding The Distribution To Use:

After making decision on the level of significance for hypothesis testing, the researcher 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.

4) Selection Of A Random Sample And Computing An AppropriateValue:

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.

5) 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.

6) Comparing The Probability:

Another step involved consists of making a comparison of the probability calculated with the specified value of α, i.e. The significance level. If the calculated probability works out to be equal to or smallerthan 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.

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 numberof 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. Someof 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:

1) Type Of Universe:

The first step involved in developing sample design is to clearly define the number of cases, technically known as the universe. 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 comprisefinite universes, the number of stars in the sky, or throwing of a dicerepresent infinite universe.

2) 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.

3) 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 itemsof 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 representativeof the population as possible.

4) 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 required. The size of the population should be 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 make an estimate of 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 consistof 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 amajor influence on the decisions related to not only the sample size, butalso 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.

Introduction:

It is important for a researcher to know the sources of data whichhe requires for different purposes. Data are nothing but the information. There are two sources of information or data they are - Primary and Secondary data. The data are name after the source. Primary data refersto the data collected for the first time, whereas secondary data refers tothe data that have already been collected and used earlier by somebodyor some agency. For example, the statistics collected by the Government

of India relating to the population is primary data for the Government of India since it has been

collected for the first time. Later when the samedata are used by a researcher for his study of a particular problem, thenthe 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 the

(a) purpose and scope of enquiry,
(b) availability of time,
(c) availability of finance,
(d) accuracy required,
(e) statistical tools to be used,
(f) sources of information (data), and (g) method of data collection.

(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 for 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. 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 totest 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 stepis 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 wholeor a state or an industrial town wherein 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, failing which the information collected may become outdated, and would have no meaning at all. For instance, if a producer wants toknow the expected demand for a product newly launched by him and the result of the enquiry that the demand would be meager takes two yearsto reach

him, then the whole purpose of enquiry would become useless because by that time he would have already incurred a huge loss. Thus, in this respect the information is quickly required and hence the researcherhas to choose the type of enquiry accordingly.

(c) Availability Of Resources:

The investigation will greatly depend on the resources available like number of skilled personnel, the financial position etc. If the number of skilled personnel who will carry out the enquiry is quite sufficient and the availability of funds is not a problem, then enquiry can be conducted over a big area covering a good number of samples, otherwise a small samplesize will do.

(d) The Degree Of Accuracy Desired:

Deciding the degree of accuracy required is a must for the investigator, because absolute accuracy in statistical work is seldom achieved. This is so because (i) statistics are based on estimates, (ii) toolsof measurement are not always perfect and (iii) there may be unintentional bias on the part of the investigator, enumerator or informant. Therefore, a desire of 100% accuracy is bound to remain unfulfilled. Degree of accuracy desired primarily depends upon the object of enquiry. For example, whenwe 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 Tools 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, a family, a household or a block of locality.A very important step before the collection of data begins is to defineclearly 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 meters, kilometers, quintals, hours, days, weeks 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 a 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 deciding 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 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 research. However, if primary data are to be collected, a decision has to be taken whether (i) census method or (ii) sampling technique is to be used for data collection. In census method, we go for total enumeration i.e., allthe units of a universe have to be investigated. But in sampling technique,we inspect or study only a selected representative and adequate fraction of the population and after analyzing the results of the sample data wedraw conclusions about the characteristics of the population. Selection of a particular technique becomes difficult because where populationor 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 it is very expensive. Therefore, for a single researcher or for a small institution it proves to be unsuitable. Onthe 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. Schedule 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 dataabout the working conditions of the workers of the Tata Iron and Steel Company, Jamshedpur, he would go to the factory, contact the workersand 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 are discussed as under:

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 interviewer can clear doubts of the informants about certain questions and thus obtain correct information. In casethe 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 supplementary information from the informant, because while interviewing it is possible to ask some supplementary questions which may be of greater use later.
4. There might be some questions which the interviewer would find difficult to ask directly, but with some tactfulness, he can mingle such

questions with others 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:

1. This method can prove to be expensive if the number of informants is large and the area is widely spread.
2. There is a greater chance of personal bias and prejudice under thismethod as compared to other methods.
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 to be obtained from the working members of households, interviews will have to be held in the evening or onweek 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 longtime, or a large number of people may have to be employed which may involve huge expenses.

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 limitedfield is desired.

In the present time of extreme advancement in the communication system, the investigator instead of going personally and conducting a faceto face interview may also obtain information over 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, such as (a) those who own a phone or a television only 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 haveto 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 informationto be obtained is of a complex nature and informants are not inclinedto respond if approached directly. For example, when the researcher istrying to obtain data on drug addiction or the habit of taking liquor, there is high probability that the addicted person will not provide the desireddata and hence will disturb the whole research process. In this situation taking the help of such persons or agencies or the neighbours who know them well becomes necessary. Since these people know the person well,they can provide 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 the facts related to the enquiry.

Though this method is very popular, its correctness depends upon a number of factors such as

1. The person or persons or agency whose help is solicited must be of proven integrity; otherwise any bias or prejudice on their part will not bring out the correct information and the whole process of researchwill become useless.
2. The ability of the interviewers to draw 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 not 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 itis 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 behonorary 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 ofa 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. Buta word of caution is that it may not always ensure accurate results because of the personal prejudice and bias of the correspondents. As 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.

4. Mailed Questionnaire Method:

Under this method, a list of questions pertaining to the surveywhich 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 related to 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. They might even try to 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, whenno prescribed sequence of questions exists, the study is *non-structured*.

When questionnaires are constructed in such a way that the objective is clear to the respondents then these questionnaires are knownas *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 be distinguished:

1. Non-disguised structured,
2. Non-disguised non-structured,
3. Disguised structured and
4. Disguised non-structured.

There are certain merits and demerits of this method of data collectionwhich are discussed below:

Merits:

1. Questionnaire method of data collection can be easily adopted wherethe field of investigation is very vast and the informants are spreadover a wide geographical area.
2. This method is relatively cheap and expeditious provided the informants respond in time.
3. This method has proved to be superior when compared to other methods like personal interviews or telephone method. This is because when questions pertaining to personal nature or the ones requiring reaction by the family are put forth to the informants, there is a chance for them to be embarrassed in answering them.

Demerits:

1. This method can be adopted only where the informants are literatesso that they can understand written questions and lend the answers inwriting.

2. It involves some uncertainty about the response. Co-operation on the part of informants may be difficult to presume.
3. The information provided by the informants may not be correct and it may be difficult to verify the accuracy.

However, by following the guidelines given below, this method can be made more effective:

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

i. Prepaid postage stamp should be affixed
ii. The sample should be large
iii. 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.
iv. It should be preferred in such enquiries where there could be a legal compulsion to provide the information.

5. Schedules Sent Through Enumerators:

Another method of data collection is 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 theirown 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 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:

1. It can be adopted in those cases where informants are illiterate.
2. There is very little scope of non-response as the enumerators go personally to obtain the information.

3. 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. Themain limitations are listed below:

Demerits:

1. In comparison to other methods of collecting primary data, this method is quite costly as enumerators are generally paid persons.
2. The success of the method depends largely upon the training imparted to the enumerators.
3. Interviewing is a very skilled work and it requires experienceand training. Many statisticians have the tendency to neglect this extremely important part of the data collecting process and this resultin bad interviews. Without good interviewing most of the information collected may be of doubtful value.
4. Interviewing is not only a skilled work but it also requires a greatdegree of politeness and thus the way the enumerators conduct theinterview would affect the data collected. When questions are askedby 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 mustbe made to remove as much of variation as possible due to different interviewers.

Secondary Data:

As stated earlier, secondary data are those data which have already been collected and analyzed by some earlier agency for its ownuse, 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. Asecondary source is a publication, reporting the data which was gatheredby other authorities and for which others are responsible."

Sources Of Secondary Data:

The various sources of secondary data can be divided into twobroad categories:

1. Published sources, and
2. Unpublished sources.

1. Published Sources:

The governmental, international and local agencies publishstatistical data, and chief among them are explained below:

(a) International Bublications:

There are some international institutions and bodies like I.M.F, I.B.R.D, I.C.A.F.E and U.N.O who publish regular and occasional reportson 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 Institute (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 sourceof secondary data. Current and important materials on statistics and socio- economic problems can be obtained from journals and newspapers likeEconomic 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, thetheses 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 whileusing 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 errorsetc. Even if there is no error such data may not be suitable and adequatefor the purpose of the enquiry. Prof. SimonKuznet'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:

4. The Suitability Of Data:

The investigator must satisfy himself that the data available aresuitable for the purpose of enquiry. It can be judged by the nature andscope 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 theabove example, if our object is to study the retail price trend of india, andif 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 of data, statistics may be relied upon.

It is not enough to have baskets of data in hand. In fact, data ina 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 editingis to detect possible errors and irregularities. The task of editing is a highly specialized one and requires great care and attention. Negligencein this respect may render useless the findings of an otherwise valuable study. Editing data collected from internal records and published sourcesis relatively simple but the data collected from a survey need excessive editing.

While editing primary data, the following considerations should be borne 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 thesame type of editing which is discussed below:

5. Editing for Completeness:

while editing, the editor should see that each schedule and questionnaire is complete in all respects. He should see to it that the answersto each and every question have been furnished. If some questions are not answered and if 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. Insuch questions, the editor should mark 'No answer' in the space provided for answers and if the questions are of vital importance then the scheduleor questionnaire should be dropped.

(a) 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 they 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.

(b) Editing for Accuracy:

The reliability of conclusions depends basically on the correctnessof 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 and an example of this kind is information relating to income, age etc.

(c) 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 a lot of differences in the methods of collecting Primary and Secondary data. Primary data which is to be collected originally involves an entire scheme of plan starting with thedefinitions of various terms used, units to be employed, type of enquiryto be conducted, extent of accuracy aimed at etc. For the collection of secondary data, a mere compilation of the existing data would be sufficient. 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 ofthe 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 would 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 the publications 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 theinvestigations both primary as well as secondary data may be used.

Questionnaire

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 ofpersons 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 elicitinformation upon a subject or sequence of subjects from 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 job 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 questionsas small as possible, keeping in view the nature, objectives and scope ofthe enquiry. Respondent's time should not be wasted by asking irrelevant and unimportant questions. A large number of questions would involvemore 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 are 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 easy, 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 also 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? Whatis 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 sincethese may be interpreted differently by different persons and as such might give unreliable and misleading information. Similarly the use of wordshaving double meaning like price, assets, capital income etc., should also be avoided.

5. Questions Should Be Comprehensive & Easily Answerable:

Questions should be designed in such a way 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 & Sensitive Nature Should Not BeAsked:

There are some questions which disturb the respondents and he/ she may be shy or irritated by hearing such questions. Therefore, everyeffort 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 bythe framers of the questionnaire and the respondent is required to tickone 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 as *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 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 forthe 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. Thishas been found extremely useful in practice. The given questionnaire canbe 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 shouldbe enclosed along with the questionnaire for the purposes 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, 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 testsa small quantity of rice to see whether it has been well-cooked and givesthe generalized result about the whole rice boiling in the vessel. The result arrived at is most of the times 100% correct. In another example, whena 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 mostof the times correct and represent the whole amount of blood available inthe 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 ofour 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 expenor too time consuming to attempt either a complete or a nearly complete coverage in a statistical study. Further to arrive at valid conclusions, itmay not be necessary to enumerate all or nearly all of a population. Wemay 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 notonly considerably cheaper but may give results which are just accurateand 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:

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.

It reduces cost

Since only a few and selected items are studied in sampling, there is reduction in cost of money and reduction in terms of man hours.

More reliable results can be obtained:

a) there are fewer chances of sampling statistical errors. If there is sampling error, it is 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.

1. **It provides more detailed information:**

As it saves time, money and labor, more detail information can be collected in a sample survey.

2. **Sometimes only sampling 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 someone's 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.

3. **Administrative convenience:**

The organization and administration of sample survey are easy for the reasons which have been discussed earlier.

4. **More scientific:**

Since the methods used to collect data are based on scientific theory and results obtained

can be tested, sampling is a more scientific method of collecting 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 arepresentative sample is taken from the universe, the result is applicableto the whole population. If the sample is not representative of the universe the result may be false and misleading.

Lack Of Experts:

As there are lack of experts to plan and conduct a sample survey, itsexecution and analysis, and its results would be Unsatisfactory and not trustworthy.

Sometimes More Difficult Than Census Method:

Sometimes the sampling plan may be complicated and requiresmore money, labor and time than a census method.

Personal Bias:

There may be personal biases and prejudices with regard to thechoice of technique and drawing of sampling units.

Choice Of Sample Size:

If the size of the sample is not appropriate then it may lead to untruecharacteristics of the population.

Conditions Of Complete Coverage:

If the information is required for each and every item of theuniverse, then a complete enumeration survey is better.

Essentials of sampling:

In order to reach a clear conclusion, the sampling should possessthe following essentials:

1. **It must be representative:**

The sample selected should possess the similar characteristics ofthe original universe from which it has been drawn.

2. **Homogeneity:**

Selected samples from the universe should have similar nature and should 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 termsof cost as well as efficiency. If the size of the sample is larger, there isbetter 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 costand efficiency.

Test Of Significance For Small Samples

If the sample size is less than 30, then those samples may be regardedas small samples. As a rule, the methods and the theory of large samplesare not applicable to the small samples. The small samples are used intesting 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 30or less and the population standard deviation is unknown.
The formula is

where,

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

$$\sigma = \sqrt{\frac{\sum(X - X)^2/n - 1}{}}$$

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.

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

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

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

Df = 20 – 1 = 19 = (10/9) √20 = (10/9) x 4.47 = 44.7/9 = 4.96Df = 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 are 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 middleof the nineteenth century, it was only after 1945 that non-parametric tests 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 understandthan parametric tests; and
iii. They can be used with type of measurements that prohibit the use ofparametric tests.

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

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

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 frequenciesand on degrees of freedom v. It is a non-parametric test.

iii. χ^2 distribution is a limiting approximation of the multinomialdistribution.

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:

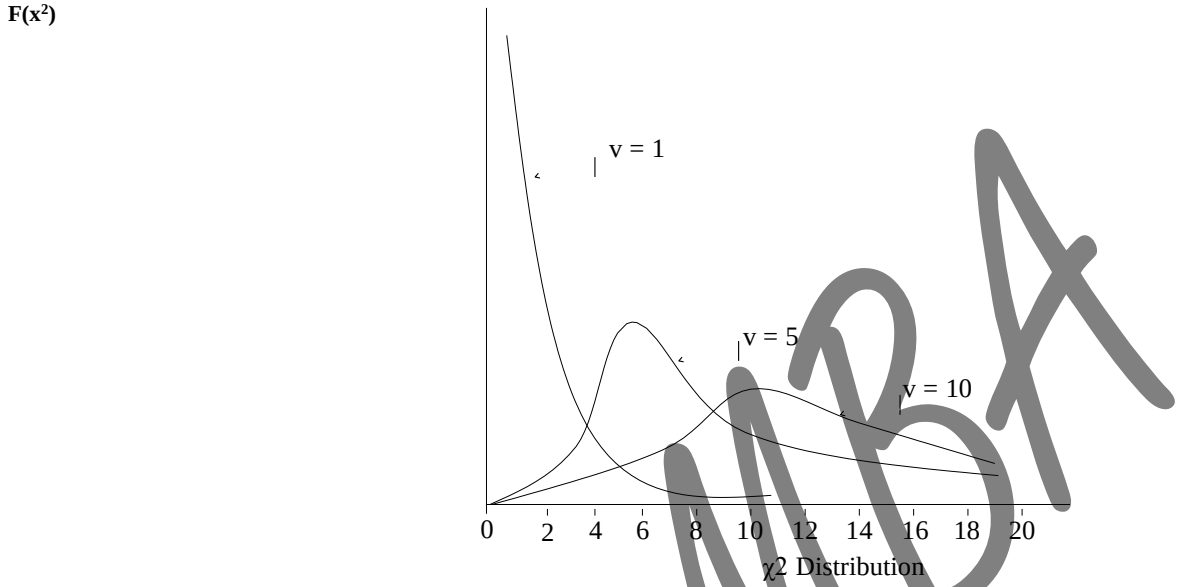
$$F(\chi^2) = C\,(\chi^2)^{(v/2-1)} e - x^2/2$$

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 distributionfor each different number of degrees of freedom. For very small numberof 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 is closely approximated by the normal curve.

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



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 degreesof freedom,

i.e.,
$\bar{X} = v$

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

(iii) $\mu_1 = 0$,

(iv) $\mu_2 = 2v$,

(v) $\mu_3 = 8v$,

(vi) $\mu_4 = 48v + 12v^2.\mu_3^2 \quad 64v^2 \quad 8$

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

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

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 χ^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 ofthe test is simple, for deviation of $\sqrt{2\chi^2}$ from $\sqrt{2v - 1}$ may be interpreted asa 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 valueof 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 χ^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. Let's assume that 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 colourof husband and wife. To examine the attributes that are associated,we formulate the null hypothesis that there is no association againstan alternative hypothesis and that there is an association between the attributes under study. If the calculated value of χ^2 is less than thetable value at a certain level of significance, we say that the result of the experiment provides no evidence for doubting the hypothesis. On the other hand, if the calculated value of χ^2 is greater than the table valueat 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 tothe data, we are interested in finding out how well this curve fits withthe observed facts. A test of the concordance of the two can be made just by inspection, but such a test is obviously inadequate. Precisioncan 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 particularmedicine, say x was administered to 812 persons out of a total populationof 3248. The number of diabetes cases 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 :

$$\text{Expectation of (AB)} = \frac{\text{(A) X (B)}}{N} = \frac{240 \times 812}{3248} = 60$$

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 hypothesisis rejected. Hence medicine x is useful in checking malaria.

Illustration:

In an experiment on immunization of cattle from tuberculosis thefollowing results were obtained:

	Affected	not affected
Inoculated	10	20
Not inoculated	15	5

Calculate χ^2 and discuss the effect of vaccine in controlling susceptibilityto 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 incontrolling 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.

UNIT-IV
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 involvestwo variables only.

TYPES OF CORRELATION

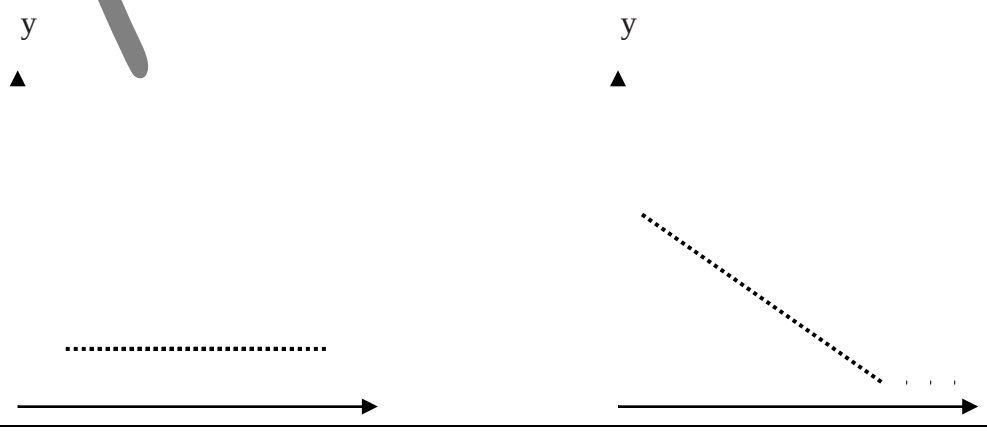
Positive correlation

If two variables x and y move in the same direction, we say thatthere is a positive correlation between them. In this case, when the valueof one variable increases, the value of the other variable also increases and when the value of one variable decreases, the value of the other variablealso 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 correlationsbetween x and y.



x x

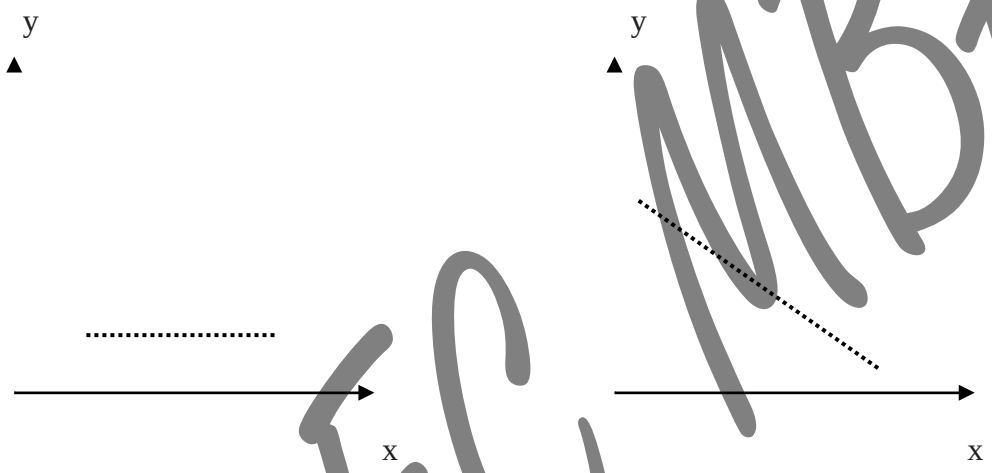
Positive Correlation Negative Correlation

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

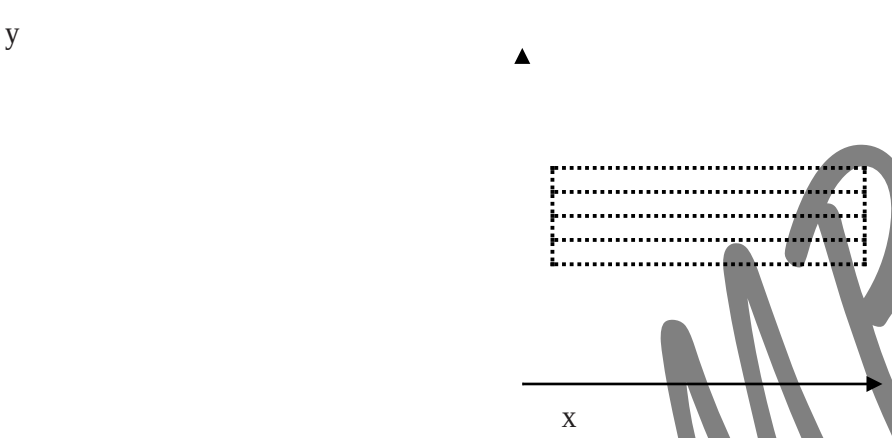
If changes in two variables are in opposite directions and the absolute values of changes are in equal proportion, we say that there is a perfect negative correlation between them.



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 betweenthe two variables is zero.



Zero Correlation

Linear Correlation'

If the quantum of change in one variable always bears a constantratio 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. Toapply 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≥ isinterpreted using the following properties:

1. The value of r ranges from – 1.0 to 0.0 or from 0.0 to 1.0
2. A value of r = 1.0 indicates that there exists perfect positive correlation between the two variables.
3. A value of r = - 1.0 indicates that there exists perfect negativecorrelation between the two variables.
4. A value r = 0.0 indicates zero correlation i.e., it shows that there is nocorrelation at all between the two variables.
5. A positive value of r shows a positive correlation between the twovariables.
6. A negative value of r shows a negative correlation between the twovariables.
7. A value of r = 0.9 and above indicates a very high degree of positivecorrelation between the two variables.
8. A value of - 0.9 ≥ r > - 1.0 shows a very high degree of negativecorrelation between the two variables.
9. For a reasonably high degree of positive correlation, we require r tobe from 0.75 to 1.0.
10. A value of r from 0.6 to 0.75 may be taken as a moderate degree ofpositive correlation.

Problem 1

The following are data on Advertising Expenditure (in Rupees Thousand)and Sales (Rupees 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 theresult.

Solution:

We have N = 6. Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum Y^2$, $\sum Y^2$ 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 asfollows:

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

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

= (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 andinterpret 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 x 2174 – (118 x 108)} / { √(6 x 2498 - 1182) √(6 x 2262 - 1082) }

= (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 andinterpret 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 x 2997 – (159 x 166)} / { √(8 x 3423 - 1592) √(8 x 3860 - 1662) }

= (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 andinterpret 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 x 8396 – (575 x 144)} / {√(10 x 33355 - 5752) √(10 x 2126 - 1442)}

= (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 thetwo variables move in the same direction. The value of r is very high. Therefore we conclude that there is high positive correlation betweenthe 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 oralcommunication skills of the short listed candidates.

Solution:

Take X = Written Communication Skill and Y = Oral CommunicationSkill.

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 $\rho = 1 - \{6 \sum D^2 / (N^3 – N)\}$

$= 1 – \{6 \times 30 / (1000 – 10)\}$

$= 1 – (180 / 990)$

$= 1 – 0.18$

$= 0.82$

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 perefficiency	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 $\rho = 1 - \{6 \sum D^2 / (N^3 - N)\}$

$= 1 - \{ 6 \times 82 / (1000 - 10) \}$

$= 1 - (492 / 990)$

$= 1 - 0.497$

$= 0.503$

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 dataon 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 equityshares 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

$\rho = 1 - \{ 6 \sum D2 / (N3 – N)\}$

$= 1 – \{ 6 \times 108 / (512 – 8) \}$

$= 1 – (648 / 504)$

$= 1 – 1.29$

$= - 0.29$

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 Manager III	4	5	1	8	9	10	6	7	3	2

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 andII is

$$\rho = 1 - \{ 6 \sum D^2 / (N^3 - N)\}$$
$$= 1 - \{ 6 \times 44 / (1000 - 10) \}$$
$$= 1 - (264 / 990)$$
$$= 1 - 0.27$$
$$= 0.73$$

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 are awarded to candidates, it is possible that certain candidates obtain equal ranks. For example, if two or three, or four

candidates secure equal ranks, a procedure that can be followed to resolve 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 fractionsalso.

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 the three items.A similar procedure is followed if four or more number of items has 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 $(m^3 – m) / 12$ asthere are ties.

Let us calculate the correction terms for certain values of m. These areprovided 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 is0.5 + 0.5 = 1

If there are 3 ties with 2 items each, then the correction term is0.5 + 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.

SchemeI 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 scoresin 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 scoresin Scheme I.

The revised ranks for Insurance Scheme I are as follows:

SchemeI 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.

SchemeII 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

$\rho = 1 - [\{ 6 \sum D^2 + (1/2) + (1/2) + 2 \} / (N^3 - N)]$

$= 1 - \{ 6 (4.+0.5+0.5+2) / (512 - 8) \} = 1 - (6 \times 7 / 504) = 1 - (42/504)$

$= 1 - 0.083 = 0.917$

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 relationshipbetween the variables X and Y, then the average relationship between these two variables is called regression, which 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 line of the best fit**. The regression of X on Y is different from the regression of Y on X. Thus, there are 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. Wehave the following result.

b_{yx} $= r\frac{\sigma_Y}{\sigma}$ *and* $b_{xy} = r\frac{\sigma_X}{\sigma}$

$$\therefore r^2 = b_{yx}b_{xy} \quad \text{and so} \quad r = \sqrt{b_{yx}b_{xy}}$$

Result:

The coefficient of correlation r between X and Y is the square rootof the product of the b values in the two regression equations. We can findr by 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 xbe 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.

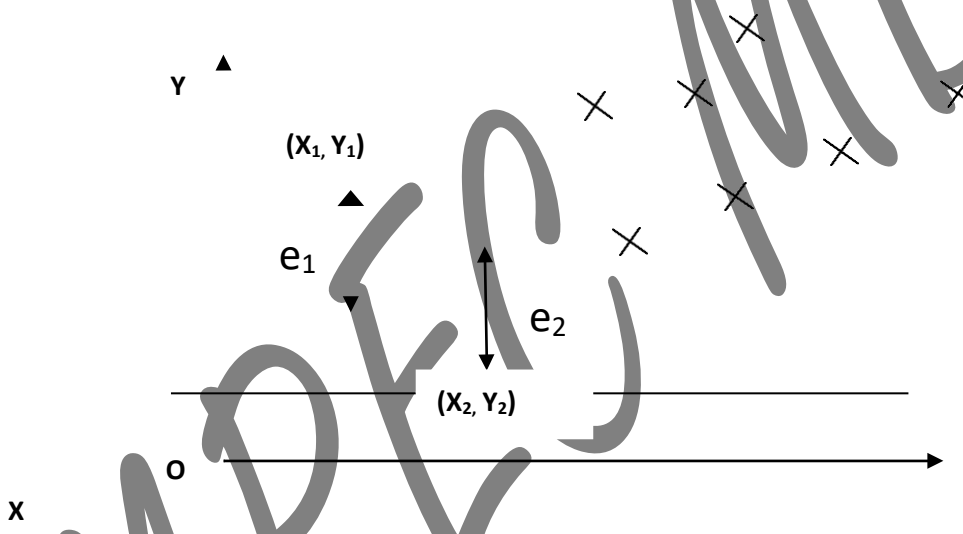
Geometrically speaking, if we plot the corresponding values of x and y in a 2-dimensional plane and join such points, we shall obtain a straight line. However, hardly we can expect all the pairs (x, y) to lie ona straight line. We can consider several straight lines which are, to some extent, near all 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 $E_1 = y_1 - y_1$. 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:

$$E_1 = y_1 - y_1 , E_2 = y_2 - y_2 ,$$

$$E_n = y_n - y_n .$$

Some of these quantities are positive while the remaining ones arenegative. However, the squares of all these quantities are positive.



i.e.,

$$E^2_1 = (y_1 - y_1)^2 \geq 0, E^2_2 = (y_2 - y_2)^2 \geq 0, \ldots, E^2_n = (y_n - y_n)^2 \geq 0.$$

Hence the sum of squares of errors (SSE) $= E^2_1 + E^2_2 + \ldots + E^2_n$

$$= (y_1 - y_2)^2 + (y_2 - y_2)^2 + \ldots + (y_n - y_n)^2 \geq 0.$$

Among all those straight lines which are somewhat near to thegiven 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. 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 finding 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 satisfythe relation

(1). Consider the summation of all such terms. We get

$\sum y = \sum (a + b x) = \sum (a .1 + b x) = (\sum a.1) + (\sum b x) = a (\sum 1) + b (\sum x)$. 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.,

$$\sum x y = a (\sum x) + b (\sum x^2) \ldots\ldots\ldots\ldots\ldots (3)$$

Equations (2) and (3) are referred to as the normal equations associatedwith the regression of y on x. Solving these two equations, we obtain

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n \sum X^2 - (\sum X)}$$

and

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

Note:

For calculating the coefficient of correlation,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, tabular column is same in both the cases with the difference that $\sum Y^2$ is also required for the coefficient of correlation.

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 interchangingthe roles of X and Y in the previous discussion. Thus, we obtain

$$a = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{n \sum Y^2 - (\sum Y)}$$

And

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum Y^2 - (\sum Y)^2}$$

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 10households.

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 theexpenditure when the income is 65.

Solution:

We have N = 10. Take X = Income, Y = ExpenditureCalculate $\sum X$, $\sum y$, $\sum Xy$, $\sum X^2$ 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

$$a = \{(\sum x^2)(\sum y) - (\sum x)(\sum xy)\} / \{n(\sum x^2) - (\sum x)^2\}$$

$$= (38400 \times 395 - 600 \times 25100) / (10 \times 38400 - 600^2)$$

$$= (15168000 - 15060000) / (384000 - 360000)$$

$$= 108000 / 24000$$

$$= 4.5$$

$$b = \{n(\sum xy) - (\sum x)(\sum y)\} / \{n(\sum x^2) - (\sum x)^2\}$$

$$= (10 \times 25100 - 600 \times 395) / 24000$$

$$= (251000 - 237000) / 24000$$

$$= 14000 / 24000$$

$$= 0.58$$

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 getY = 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 regressionline and so we did not take $\sum Y^2$. Now we require two regression lines. Therefore,

Calculate $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, $\sum Y^2$.

X	Y	XY	X2	Y2
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

$$a = (51075 \times 885 - 695 \times 65450) / (10 \times 51075 - 695^2)$$

$$= (45201375 - 45487750) / (510750 - 483025)$$

$$= -286375 / 27725$$

$$= -10.329$$

$$b = (10 \times 65450 - 695 \times 885) / 27725$$

$$= (654500 - 615075) / 27725$$

$$= 39425 / 27725$$

= 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

$$a = \{(\sum y^2)(\sum x) - (\sum y)(\sum xy)\} / \{n(\sum y^2) - (\sum y)^2\}$$

And

$$b = \{n(\sum xy) - (\sum x)(\sum y)\} / \{n(\sum y^2) - (\sum y)^2\}.$$

We get

$a = (84925 \times 695 - 885 \times 65450) / (10 \times 84925 - 885^2)$

$= (59022875 - 57923250) / (849250 - 783225)$

$= 1099625 / 66025$

$= 16.655,$

$b = (10 \times 65450 - 695 \times 885) / 66025$

$= (654500 - 615075) / 66025$

$= 39425 / 66025$

$= 0.597$

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 - 695^2) √ (10 x 84925 - 885^2) }

= (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 = √ (1.422 x 0.597)
= √ 0.848934
= 0.9214.

Note that this agrees with the above value of r.

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

According to 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 andconstant 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.

One-way classified data

When the set of observations is distributed over different levels ofa single factor, then it gives one-way classified data.

ANOVA for One-way classified data

Let y denote the j^{th} observation corresponding to the i^{th} 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

$$y_{ij} = \mu + a_i + e_{ij} \qquad \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

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 i^{th} level of the factor A (i=1,2,…,k) andhence 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

$$y_{ij} = \mu + a_i + e_{ij} \qquad \begin{pmatrix} i = 1,2,...,k \\ j = 1,2,...,n_i \end{pmatrix}$$

The expected value of the general observation y_{ij} in the experimentalunits 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 touncontrolled 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 modelreduces to

$$y_{ij} = \mu + a_i + e_{ij} \quad \begin{pmatrix} i = 1, 2, ..., k \\ j = 1, 2, ..., n \end{pmatrix} \qquad (1)$$

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 μ 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

$$E = \sum_{ij} e_{ij}^2$$
$$= \sum_{ij} (y_{ij} - \mu - a_i)^2,$$

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 $T_i = n_i \mu + n_i a_i,\ i = 1,2,\ldots,k$ (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 linear equations, it follows that unique solution for this systemis not possible.

However, by making the assumption that $\sum_i n_i a_i = 0$, we can get a unique solution μ for 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 $\hat{\mu} = \frac{G}{N}$ (1)

Again from equation (2), we have

$$\frac{T_i}{n} = \mu + a_i$$

Hence, $a_i = \frac{T_i}{n_i} - \mu$

Therefore, the estimate of a_i is given by

$$\hat{a}_i = \frac{T_i}{n_i} - \hat{\mu}$$

i.e.,

$$\hat{a}_i = \frac{T_i}{n_i} - \frac{G}{N} \quad (2)$$

Substituting the least square estimates of μ and μ_i in the residualsum 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

$$E = \left(\sum_{ij} y_{ij}^2 - \frac{G^2}{N} \right) - \left(\sum_{i} \frac{T^2}{n} - \frac{G^2}{N} \right) \qquad (3)$$

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.,

$$H_o : \mu_1 = \mu_2 = ... = \mu_k = \mu$$

$$i.e.,\ H_o : \mu_i = \mu \quad for\ all\ i = 1, 2, ..., k$$

$$i.e.,\ H_o : \mu_i - \mu = 0 \quad for\ all\ i = 1, 2, ..., k$$

$$i.e.,\ H_o : a_i = 0$$

Under H_o , the linear model reduces to

$$y_{ij} = \mu + e_{ij} \qquad \begin{matrix} i = 1, 2, ..., k \\ j = 1, 2, ..., n \end{matrix}$$

Proceeding as before, we get the residual sum of squares for thishypothetical model as

$$E_1 = \left(\sum_{ij} y_{ij}^2 \right) - \frac{G^2}{N} \qquad (4)$$

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

$$E_1 - E = \sum_{i=1}^{k} \frac{T^2}{n_i} - \frac{G^2}{N} \qquad (5)$$

The expression in (8) is usually called the **corrected treatment sum of squares** while the term $\sum_{i=1}^{k} \frac{T^2}{n_i}$ 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 denoted by **α**. The testing of the null hypothesis H_o may be carried out by F test. For given **α**, we have

$$F = \frac{TrMSS}{EMSS} = \frac{\frac{Trss}{dF}}{\frac{Ess}{dF}} : F_{k-1, N-k}$$

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

Sourceof Variation	Degrees of Freedom	Sumof Squares(SS)	Mean Squares (MS)	variance ratio F
Between the level of the factor (Treatment)	k-1	$E_1 - E = Q_T$ $\sum_i^k \frac{T^2}{n_i} - \frac{G^2}{N}$	$M_T = \frac{Q_T}{k-1}$	$F_T = \frac{M_T}{M_E}$: $F_{k-1, N-k}$
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 toas 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.,) 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 personsin 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 anysignificant 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 itemsfor 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 - X = sales – 7	$(Sales-7)^2$
A	8	1	1
A	9	2	4
A	5	-2	4
A	10	3	9
B	7	0	0
B	6	-1	1
B	6	-1	1
B	9	2	4
C	6	-1	1
C	6	-1	1
C	7	0	0
C	5	2	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) = 2Degrees of freedom for smaller variance (df_1) = 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 ofF. 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

$$\frac{(\sum A)^2}{N_1} \quad \frac{(\sum B)^2}{N_2} \quad \frac{(\sum C)^2}{N_3} = \_\_\_\_\_ + \_\_\_\_\_ + \_\_\_\_\_ - CF$$

$$= \frac{32^2}{4} + \frac{28^2}{4} + \frac{24^2}{4} - 588$$

$$= \frac{1024}{4} + \frac{784}{4} + \frac{576}{4} - 588$$

$$= 256 + 196 + 144 - 588$$

$$= 8$$

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 IIare 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 1420to 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

$$\text{Correction factor} = \frac{T^2}{N} = \frac{96^2}{12} = 768$$

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 deviation	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:

$$F = \frac{Greater\ Variance}{Smaller\ Variance} = \frac{27.38}{10.75} = 2.5470$$

Degrees of freedom for greater variance (df_1) = 9

Degrees of freedom for smaller variance (df_2) = 2

The table value of f at 5% level of significance = 19.38Inference:

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

$$(df_1) = 5$$

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 Fwe accept the null-hypothesis and conclude that there is no significantdifference in the performance of the three financial schemes.

I. PARTIAL CORRELATION

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 partialcorrelation coefficient $r_{12.3}$

Solution:

We have

$$= \frac{0.6 - 0.58\,x\,0.70}{\sqrt{(1-(0.58)^2)\,(1-(0.70)^2)}}$$

$$= \frac{0.6 - 0.406}{\sqrt{(1-0.3364)\,(1-0.49)}}$$

$$= \frac{0.194}{\sqrt{0.6636\,x\,0.51}}$$

$$= \frac{0.194}{0.8146\,x\,0.7141}$$

$$= \frac{0.194}{0.5817}$$

$$= 0.3335$$

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

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1 - r^2_{12})\ (1 - r^2_{32})}}$$

$$= \frac{0.8 - 0.75\,X\,0.70}{\sqrt{(1-(0.75)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.8 - 0.525}{\sqrt{(1-0.5625)\ (1-0.49)}}$$

$$= \frac{0.275}{\sqrt{(0.4375)\ (0.51)}}$$

$$= \frac{0.275}{0.6614\,X\,0.7141}$$

$$= \frac{0.275}{0.4723}$$

= 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. Multiplecorrelations arise 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 coefficients are denoted by the letter R.The dependent variable is denoted by X_1. The independent variables are denoted by $X_2, X_3, X_4, \ldots$, 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 Correlations

The coefficient of multiple linear correlation is given in terms ofthe partial correlation coefficients as follows:

$$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$$

$R_{3.12} = \dfrac{\sqrt{r^2_{31} + r^2_{32} - 2\, r_{31}\, r_{32}\, r_{12}}}{\sqrt{1 - r^2_{12}}}$

Properties Of The Coefficient Of Multiple Linear Correlations:

1. The coefficient of multiple linear correlations 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

$R_{1.23} = \dfrac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$

$= \dfrac{\sqrt{(0.6)^2 + (0.65)^2 - 2x0.6x0.65x0.8}}{\sqrt{1 - (0.8)^2}}$

$$= \frac{\sqrt{0.36+ 0.4225- 0.624}}{\sqrt{1 -0.64}}$$

$$= \frac{\sqrt{0.7825- 0.624}}{\sqrt{0.36}}$$

$$= \frac{\sqrt{0.1585}}{\sqrt{0.36}}$$

$$= \sqrt{0.4403}$$

$$= 0.6636$$

Problem 4

Given that $r_{21} = 0.7$, $r_{23} = 0.85$ and $r_{13} = 0.75$, determine $R_{2.13}$

Solution:

We have $R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$

$$= \frac{\sqrt{(0.7)^2 + (0.85)^2 - 2\ x0.7x0.85x0.75}}{\sqrt{1 - (0.75)^2}}$$

$$= \frac{\sqrt{0.49+ 0.7225- 0.8925}}{\sqrt{1 - 0.5625}}$$

$$= \frac{\sqrt{1.2125- 0.8925}}{\sqrt{0.4375}}$$

$$= \frac{\sqrt{0.32}}{\sqrt{0.4375}}$$

$$= \sqrt{0.7314}$$

=0.8552

INDEX NUMBERS

Introduction:

Index numbers are meant to study the change in the effects of such factors which cannot be measured directly. According to Bowley, –Index numbers are used to measure the changes in some quantity which we cannot observe directly‖. For example, changes in business activity in a country are not capable of direct measurement but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers are commonly used statistical device for measuring the combined fluctuations in a group related variables. If we wish to compare the price level of consumer items today with that prevalent ten years ago, we are not interested in comparing the prices of only one item, but in comparing some sort of average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers which may be defined as a device for combining the variations that have come in group of related variables over a period of time, with a view to obtain a figure that represents the _net_ result of the change in the constitute variables.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv) business activity. Thus, we have index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

Index Numbers:

Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

Simple Index Number:

A simple index number is a number that measures a relative change in a single variable with respect to a base.

Composite Index Number:

A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base.

Types of Index Numbers:

Following types of index numbers are usually used:

Price index Numbers:

Price index numbers measure the relative changes in prices of a commodities between two periods. Prices can be either retail or wholesale.

Quantity Index Numbers:

These index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items.

Uses

This index number is a useful number that helps us quantify changes in our field. It is easier to see one value than a thousand different values for each item in our field.

Take the stock market, for example. It is comprised of thousands of different public companies. We could, of course, look at the stock value of each of these companies to see how the companies are doing as a whole, or we can look at just one number, the stock index, to get a general feel for how the companies are doing.

The same goes for the cost of goods. We could look at the cost of each item and compare it to its cost from last year. But that would mean looking at the cost of millions of items. Or we could look at the cost of goods index, just one number, to see whether prices have increased or decreased over the past year.

We can say that the index number is one simple number that we can look at to give us a general overview of what is happening in our field. Let's take a look at two real world index numbers.

Line of Best Fit (Least Square Method)

A **line of best fit** is a straight line that is the best approximation of the given set of data.
It is used to study the nature of the relation between two variables. (We're only considering the two-dimensional case, here.)

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible).

A more accurate way of finding the line of best fit is the **least square method**.
Use the following steps to find the equation of line of best fit for a set of ordered pairs $(x1,y1),(x2,y2),...(x_n,y_n)$(x1,y1),(x2,y2),...(xn,yn).

Step 1: Calculate the mean of the xx-values and the mean of the yy-values.
$\overline{X}=\sum i=1nx_inY=\sum i=1ny_inX^-=\sum i=1nxinY^-=\sum i=1nyin$

Step 2: The following formula gives the slope of the line of best fit:
$m=\sum i=1n(x_i-\overline{X})(y_i-\overline{Y})\sum i=1n(x_i-\overline{X})_2m=\sum i=1n(xi-X^-)(yi-Y^-)\sum i=1n(xi-X^-)2$

Step 3: Compute the yy-intercept of the line by using the formula:
$b=\overline{Y}-m\overline{X}b=Y^--mX^-$

Step 4: Use the slope mm and the yy-intercept bb to form the equation of the line.

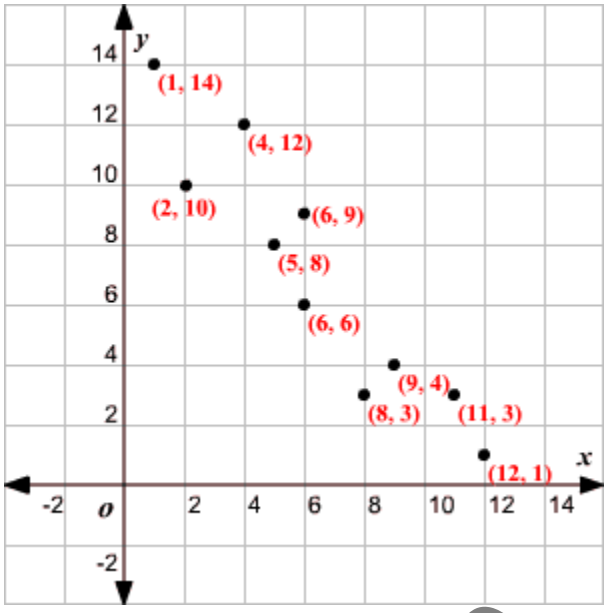
Example:

Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

x x	8 8	22	111 1	6 6	5 5	44	121 2	9 9	6 6	11
y y	3 3	101 0	33	6 6	8 8	121 2	11	4 4	9 9	141 4

Solution:

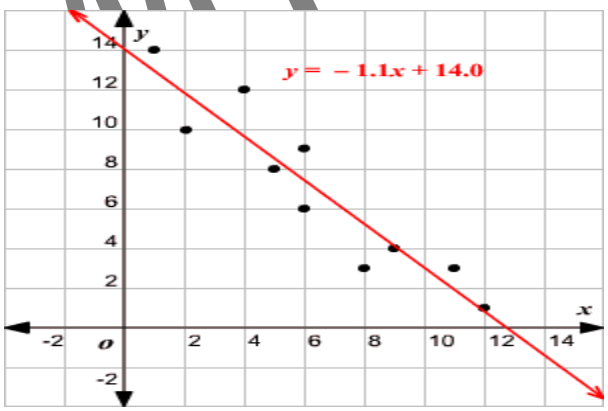
Plot the points on a coordinate plane.



Calculate the means of the *x*x-values and the *y*y-values.

$X^{\overline{\quad}}=8+2+11+6+5+4+12+9+6+110=6.4Y^{\overline{\quad}}=3+10+3+6+8+12+1+4+9+1410=7X^{-}=8+2+11+6+5+4+12+9+6+11\ 0=6.4Y^{-}=3+10+3+6+8+12+1+4+9+1410=7$

Now calculate $x_i-X^{\overline{\quad}}$ xi−X¯, $y_i-Y^{\overline{\quad}}$ yi−Y¯, $(x_i-X^{\overline{\quad}})(y_i-Y^{\overline{\quad}})$(xi−X¯)(yi−Y¯), and $(x_i-X^{\overline{\quad}})_2$(xi−X¯)2 for each *i*i.



*i*i	*xi*x i	*yi*y i	x i−X¯	y i−Y¯	(xi−X¯)(yi−Y¯)	(*xi*−*X*¯¯¯)2
11	88	33	1.61.6	−4−4	−6.4−6.4	2.562.56
22	22	10	−4.4−4.4	33	−13.2−13.2	19.3619.3
33	11	33	4.64.6	−4−4	−18.4−18.4	21.1621.1
44	66	66	−0.4−0.4	−1−1	0.40.4	0.160.16
55	55	88	−1.4−1.4	11	−1.4−1.4	1.961.96
66	44	12	−2.4−2.4	55	−12−12	5.765.76
77	12	11	5.65.6	−6−6	−33.6−33.6	31.3631.3
88	99	44	2.62.6	−3−3	−7.8−7.8	6.766.76
99	66	99	−0.4−0.4	22	−0.8−0.8	0.160.16
10	11	14	−5.4−5.4	77	−37.8−37.8	29.1629.1

Calculate the slope.

$m=\sum i=1n(x_i-\bar{X})(y_i-\bar{Y})\sum i=1n(x_i-\bar{X})_2=-131118.4\approx-1.1$m=∑i=1n(xi−X¯)(yi −Y¯)∑i=1n(xi−X¯)2=−131 118.4≈−1.1

Calculate the *y*y-intercept.

Use the formula to compute the *y*y-intercept.

$b=\bar{Y}-m\bar{X}$ =7−(−1.1×6.4) =7+7.04 ≈14.0b=Y¯−mX¯ =7−(−1.1×6.4) = 7+7.04 ≈14.0 Use the slope and *y*y-intercept to form the equation of the line of best fit.

The slope of the line is −1.1−1.1 and the *y*y-intercept is 14.014.0.Therefore, the equation is

$y=-1.1x+14.0$y=−1.1x+14.0.

Draw the line on the scatter plot.