BHARATHIDASAN UNIVERSITY
CONSTITUENT COLLEGE OF ARTS AND SCIENCE
NAGAPATTINAM-611106

M.B.A

CORE COURSE - XII

RESEARCH METHODS IN MANAGEMENT

AUTHORS

Dr. J.JEYANTHI, M.B.A., M.Phil., NET, Ph.D
Head, Department of Business Administration,
Bharathidasan University Constituent College,
Nagapattinam – 611106

RESEARCH METHODS IN MANAGEMENT

Unit-I

Research – Qualities of Researcher – Components of Research Problem – Various Steps In Scientific Research – Types of Research – Hypotheses Research Purposes - Research Design – Survey Research – Case Study Research.

Unit-II

Data Collection – Sources of Data – Primary Data – Secondary DataProcedure Questionnaire – Sampling Methods – Merits and Demerits – Experiments – Observation Method – Sampling Errors - Type-I Error & Type-II Error.

Unit-III

Statistical Analysis – Introduction To Statistics – Probability Theories – Conditional Probability, Poisson Distribution, Binomial Distribution and Properties of Normal Distributions – Hypothesis Tests– One Sample Test – Two Sample Tests / Chi-Square Test, Association of Attributes - Standard Deviation – Co-Efficient of Variations .

Unit-IV

Statistical Applications – Correlation and Regression Analysis – Analysis of Variance – Partial and Multiple Correlation – Factor Analysis and Conjoint Analysis – Multifactor Evaluation – Two-Factor Evaluation Approaches.

Unit-V

Research Reports – Structure and Components of Research Report– Types of Report, Characteristics of Good Research Report, Pictures and Graphs, Introduction To SPSS.

[Note: Distribution of Questions between Problems and Theory of thispaper must be 40:60 i.e., Problem Questions: 40 % & Theory Questions:60 %]

REFERENCES

Panneerselvam, R., RESEARCH METHODOLOGY,*Prentice hall ofIndia, New Delhi, 2004.*

Kothari CR, RESEARCH METHODOLOGY-METHODS ANDTECHNIQUES, *New Wiley Eastern ltd., Delhi, 2009.*

CONTENTS

UNIT – I

Lesson – 1

RESEARCH

Introduction

The introduction leads the reader from a general subject area to a particular topic of inquiry. It establishes the scope, context, and significance of the research being conducted by summarizing current understanding and background information about the topic, stating the purpose of the work in the form of the research problem supported by a hypothesis or a set of questions, explaining briefly the methodological approach used to examine the research problem, highlighting the potential outcomes your study can reveal, and outlining the remaining structure and organization of the paper.

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 or issue. It is also known as the art of scientific investigation. Several social scientists have defined research in different ways.

In the *Encyclopedia of Social Sciences*, D. Slesinger and M. Stephension (1930) defined research as "the manipulation of things, 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 pursue truth through the methods of study, observation, comparison and experiment. In sum, research is the search for knowledge, using objective and systematic methods to find solution to a problem.

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 research is 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:

- To gain familiarity with new insights into a phenomenon (i.e., formative research studies);
- To accurately portray the characteristics of a particular individual, group, or a situation (i.e., descriptive research studies);
- To analyse the frequency with which something occurs (i.e., diagnostic research studies); and
- 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 subjective assessment of opinions, behaviour and attitudes. Research in such a situation is a function of the researcher's impressions and insights. The results generated by this type of research are either in non-quantitative form or in the form which cannot be put to rigorous quantitative analysis. Usually, this approach uses techniques like in-depth interviews, focus group interviews, and projective techniques.

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:

First of all, the nature of a researcher must be of the temperament that vibrates in unison with the theme which he is searching. Hence, the seeker of knowledge must be truthful with truthfulness of nature, which is much more important, much more exacting than what is sometimes known as truthfulness. The truthfulness relates to the desire for accuracy of observation and precision of statement. Ensuring facts is the principle rule of science, which is not an easy matter. The difficulty may arise due to untrained eye, which fails to see anything beyond what it has the power of seeing and sometimes even less than that. This may also be due to the lack of discipline in the method of science. An unscientific individual often remains satisfied with 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.

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 not make 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. Researchdemands 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".

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 easy task. 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 making statements cautiously. According to Huxley, the assertion that outstrips the evidence is not only a blunder but a crime (Thompson, 1975). A researcher should cultivate the habit of reserving judgment when the required data are insufficient.

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 theanalysis 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, although in 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 is engaged by the government these days to undertake this work. Thus, research as a tool of government economic policy formulation involves three distinct stages of operation:

(i) investigation of economic structure through continual compilation of 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 research are 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 application of logical, mathematical, and analytical techniques to find solution to business problems, such as cost minimization or profit maximization, or the optimization problems.

Motivational research helps to determine why people behave in the manner they do with respect to market characteristics. More specifically, it is concerned with the 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 sake of 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.

Research Process

Research process consists of a series of steps or actions required for effectively conducting research. The following are the steps that provide useful procedural guidelines regarding the conduct of research:

- Formulating the research problem;
- Extensive literature survey;
- Developing hypothesis;
- Preparing the research design;
- Determining sample design;
- Collecting data;
- Execution of the project;
- Analysis of data;
- Hypothesis testing;
- Generalization and interpretation, and
- Preparation of the report or presentation of the results.

In other words, it involves the formal write-up of conclusions.

Research Problem:

The first and foremost stage in the research process is to select and properly define the research problem. A researcher should first identify a problem and formulate it, so as to make it amenable or susceptible to research. In general, a research problem refers to an unanswered question that a researcher might encounter in the context of either a theoretical or practical situation, which he/she would like to answer or find a solution to. A research problem is generally said to exist if the following conditions emerge (Kothari, 1988):

There should be an individual or an organisation, say X, to whom the Problem can be attributed. The individual or the organization is situated in an environment Y, which is governed by certain uncontrolled variables Z;

There should be at least two courses of action to be pursued, say A1 and A2. These courses of action are defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

There should be atleast two alternative possible outcomes of the said courses of action, say B1 and B2. Of them, one alternative should be preferable to the other. That is, atleast one outcome should be what the researcher wants, which becomes an objective.

The courses of possible action available must offer a chance to theresearcher to achieve the objective, but not the equal chance. Therefore, if $P(B_j / X, A, Y)$ represents the probability of the occurrence of an outcome B_j when X selects A_j in Y, then $P(B1 / X, A1,Y) \neq P(B1 / X, A2, Y)$. Putting it in simple words, it means that the choices must not have equal efficiencies for the desired outcome.

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

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

Thus, identification of a research problem is the pre-condition to conducting research. A research problem is said to be the one which requires a researcher to find the best available solution to the given problem. That is, the researcher needs to find out the best course of action through which the research objective may be achieved optimally in the context of a given situation. Several factors may contribute to making the problem complicated. For example, the environment may alter, thus affecting the efficiencies of the alternative courses of action taken or the quality of the outcomes. The number of alternative courses of action might be very large and the individual not involved in making the decision may be affected by the change in environment and may react to it favorably or unfavorably. Other similar factors are also likely to cause such changes in the context of research, all of which may be considered from the point of view of a research problem.

5 MOST IMPORTANT COMPONENTS OF RESEARCH PROBLEM

The Most Important Components of research problem as discussed by R. L. Ackoff are listed below:

(1) Research Consumer:

There must be individuals or groups which have some difficulty or problem. The individuals or the groups theAll are affected by the decision on the part of the research consumer.

(2) Research-Consumer’s Objective:

There must be some objectives to be attained as the research consumer must have something he wants to get it. It one wants nothing, one cannot have a problem.

(3) Alternative Means to Meet the Objective:

There must be alternative means or the courses of action for attaining an objective one wishes to obtain. Means are courses of action. A course of action may involve the use of objects. Objects are the instruments. This means that there must be at least two means available to a researcher or if he has no choice or means, he cannot have a problem.

(4) Doubt in Regard to Selection of Alternatives:

The existence of alternative courses of action is not enough. To experience a problem the researcher must have some doubt as to which alternative to select. Without such a doubt there can be no problem. This means that research must answer the question concerning the relative efficiency of the possible alternative.

(5) There must be one or More Environments:

There must be some environments to which the difficulty or problem pertains. A change in the environment may produce or remove a problem. A researcher may have doubts as to which will be the most efficient means in one environment but may entertain no such doubt in another. Some problems are quite general.

Thus, a research problem is one which requires a researcher to find out the best solution for the given problem so that the objective can be attained optimally in the context of a given environment.

Conclusion

The specific length of the conclusive section to a paper may vary. Normally, the appropriate length is dependent upon the general length of the paper. A research paper that is very long, such as a dissertation or a graduate thesis, may need a conclusion that extends for several pages. However, the conclusion to paper that is written for an ordinary research assignment may need only a few paragraphs in order to generate an effective conclusion. A shorter paper of a few pages may only need a single paragraph. It is the responsibility of the writer to use their best judgment concerning how long their conclusion needs to be. Being able to make judgment calls of this kind is an essential aspect of becoming a good writer.

VARIOUS STEPS IN SCIENTIFIC METHODS

The ‘scientific method’ merely refers to a broad framework for studying and learning more about the world around us in a scientific manner. It is not so much a series of absolute, unchangeable steps as a guideline to the method that must be used when trying to reach a scientifically acceptable theory about a subject matter. Therefore, it is not possible to provide a finite number of steps or an exact procedure for following the scientific method. However, the scientific method steps detailed below describe the main steps that scientists commonly take when conducting a scientific inquiry.

Steps of the Scientific Method

1. **Make an Observation**

   Scientists are naturally curious about the world. While many people may pass by a curious phenomenon without sparing much thought for it, a scientific mind will take note of it as something worth further thought and investigation.

2. **Form a Question**

   After making an interesting observation, a scientific mind itches to find out more about it. This is in fact a natural phenomenon. If you have ever wondered why or how something occurs, you have been listening to the scientist in you. In the scientific method, a question converts general wonder and interest to a channeled line of thinking and inquiry.

3. **Form a Hypothesis**

   A hypothesis is an informed guess as to the possible answer of the question. The hypothesis may be formed as soon as the question is posed, or it may require a great deal of background research and inquiry. The purpose of the hypothesis is not to arrive at the perfect answer to the question but to provide a direction to further scientific investigation.

4. **Conduct an Experiment**

   Once a hypothesis has been formed, it must be tested. This is done by conducting a carefully designed and controlled experiment. The experiment is one of the most important steps in the scientific method, as it is used to prove a hypothesis right or wrong, and to formulate scientific theories. In order to be accepted as scientific proof for a theory, an experiment must meet certain conditions – it must be controlled, i.e. it must test a single variable by keeping all other variables under control. The experiment must also be reproducible so that it can be tested for errors.

5. **Analyse the Data and Draw a Conclusion**

   As the experiment is conducted, it is important to note down the results. In any experiment, it is necessary to conduct several trials to ensure that the results are constant. The experimenter then analyses all the data and uses it to draw a conclusion regarding the strength of the hypothesis. If the data proves the hypothesis correct, the original question is answered. On the other hand, if the data disproves the hypothesis, the scientific inquiry continues by doing research to form a new hypothesis and then conducting an experiment to test it. This process goes on until a hypothesis can be proven correct by a scientific experiment.

Conclusion

The whole process is collaborative and is conducted in a clearly documented manner to help other scientists who are doing research in the same field. Throughout history, there are instances where scientists have stopped their research before completing all the steps of the scientific method**,** only to have the inquiry taken up and solved by another scientist interested in answering the same question.

Questions

1. What is research?
2. What are the objectives of research?
3. Distinguish between Research methods Vs Methodology?
4. Explain research approaches
5. What are the qualities of a researcher?
6. State the significance of research
7. State the research process.
8. Explain the research problem.
9. State the components of research problem.
10. What are the various steps in scientific methods?

Lesson – 2

TYPES OF RESEARCH

Introduction

Classification of research can be based on different considerations. Thus we can base our classification on the nature of the dominant data (qualitative or quantitative), the purpose of the research (applied or basic) or the type of analysis that will be carried out (descriptive or analytical). The attempt to classify research into these categories is somewhat misleading since most research has elements of all the categories. It should be said that it is only an aid to broad understanding of the different types of research rather than distinct categories.

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 post facto research' is quite often used for descriptive research studies in social sciences and business research. The most distinguishing feature of this method is that the researcher has no control over the variables here. He/she has to only report what is happening or what has happened. Majority of the ex post facto research projects are used for descriptive studies in which the researcher attempts to examine phenomena, such as the consumers' preferences, frequency of purchases, shopping, etc. Despite the inability of the researchers to control the variables, ex post facto studies may also comprise attempts by them to discover the causes of the selected problem. The methods of research adopted in conducting descriptive research are survey methods of all kinds, including correlational and comparative methods.

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

2. **Applied 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, an industry, a business organisation, or the society is known as applied research. Researchers engaged in such researches aim at drawing certain conclusions confronting a concrete social or business problem.

On the other hand, fundamental research mainly concerns generalizations and formulation of a theory. In other words, "Gathering knowledge for knowledge's sake is termed 'pure' or 'basic' research" (Young in Kothari, 1988). Researches relating to pure mathematics or concerning some natural phenomenon are instances of Fundamental Research. Likewise, studies focusing on human behaviour also fall under the category of fundamental research.

Thus, while the principal objective of applied research is to find a solution to some pressing practical problem, the objective of basic research is to find information with a broad base of application and add to the already existing organized body of scientific knowledge.

3. Quantitative Versus Qualitative:

Quantitative research relates to aspects that can be quantified or can be expressed in terms of quantity. It involves the measurement of quantity or amount. 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 is discovering the underlying motives and desires of human beings by usingn-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 research is 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 prove that 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.

Other Types Of Research:

The remaining types of research are variations of one or more of the 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 causes of 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 specific hypotheses to be verified. As regards to historical research, sources like historical documents, remains, etc. Are utilized to study past events or ideas. It also includes philosophy of persons and groups of the past or any remote point of time.

Research 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 has no 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:

- ✓ 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;
- ✓ The knowledge of doing research inculcates the ability to evaluate and utilize the research findings with confidence;

- ✓ The knowledge of research methodology equips the researcher with the tools that help him/her to make the observations objectively; and
- ✓ The knowledge of methodology helps the research consumers to evaluate research and make rational decisions.

Conclusion

Many of us have experienced research writing projects as a way to "prove" what we already believe. An essay assignment may ask us to take a position on a matter, and then support that position with evidence found in research. You will likely encounter projects like this in several classes in college. Because you enter a project like this with a thesis in hand, it is very tempting to look for and use only those sources that agree with you and to discard or overlook the others.

Questions

1. Explain the types of research.
2. What is descriptive research?
3. What do you meant by analytical research?
4. What is the importance of knowing how to conduct research?

Lesson – 3

HYPOTHESIS

Introduction

A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and "theory" are often used synonymously, a scientific hypothesis is not the same as a scientific theory. A working hypothesis is a provisionally accepted hypothesis proposed for further research.

Hypothesis meaning

"Hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of 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 capable of being tested using scientific methods that involve an independent and some dependent variables. For instance, the following statements may be considered:

- "Students who take tuitions perform better than the others who do not receive tuitions" or,
- "The female students perform as well as the male students".

These two statements are hypotheses that can be objectively verified and tested. Thus, they indicate that a hypothesis states what one is looking for. Besides, it is a proposition that can be put to test in order to examine its validity.

Characteristics of Hypothesis:

A hypothesis should have the following characteristic features:-

- 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.
- A hypothesis must be capable of being put to test. Quite often, the research programmes fail owing to its incapability of being subject to testing for validity. Therefore, some prior study may be conducted by the researcher in order to make a hypothesis testable. A hypothesis "is tested if other deductions can be made from it, which in turn can be confirmed or disproved by observation" (Kothari, 1988).
- A hypothesis must state relationship between two variables, in the case of relational hypotheses.
- 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.
- 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.
- A hypothesis must be consistent and derived from the most known facts. In other words, it should be consistent with a substantial body of established facts. That is, it must be in the form of a statement which is most likely to occur.
- 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 given period of time, as no one can afford to spend a life-time on collecting data to test it.

- 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:

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 that both the methods are equally good, then such a statement is called as the null hypothesis. On the other hand, if method A is considered relatively superior to method B, or vice-versa, then such a statement is known as an alternative hypothesis. The null hypothesis is expressed as **H_0**, while the alternative hypothesis is expressed as H_a. For example, if a researcher wants to test the hypothesis that the population mean (μ) is equal to the hypothesized mean (H_0) = 100, then the null hypothesis should be stated as the population mean is equal to the hypothesized mean 100. Symbolically it may be written as:-

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

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

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 null hypothesis and the alternative hypothesis. While formulating the null hypothesis, the following aspects need to be considered:

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

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

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

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 other words, the five per cent level of significance implies that the researcher is willing to take a risk of five per cent of rejecting the null hypothesis, when (H_0) is actually true. In sum, the significance level reflects the maximum value of the probability of rejecting H_0 when it is actually true, and which is usually determined prior to testing the hypothesis.

PURPOSES OF RESEARCH AND OBJECTIVES

The School of Engineering embraces "Research First" and "Open-door" as the basic principles, and advocates "practice-oriented research and education" based on *creative* researches, and always opens its doors to the world as a research-intensive university. By establishing a *"global center of excellence for intellectual creativity"* backed by vision and professionalism, the School of Engineering aims to solve various difficult problems faced by modern society and to fulfill its obligations to the future of humanity and the earth.

The research purposes are to foresee future problems through pursuit of truth as a *"global center of excellence for intellectual creativity"*, to respond to current social demands, and to contribute to the creation and development of scientific technologies with the aim of realizing an affluent society and natural environment for humanity. At the same time, the School of Engineering aims to create excellent educational resources and an excellent educational environmentthrough frontline researches.

To achieve the above-mentioned purposes, the following *objectives* are set:

1. Lead the academic world and conduct internationally high-level researches in each engineering field
2. Aim to discover new phenomena and create new technologies, based on principles and rules in natural phenomena and pursuit of truth in each engineering field
3. Conduct cutting-edge researches to lead the academic and industrial worlds at home and abroad and create and develop new academic and technological fields
4. Aim to globalize research and education
5. Provide suggestions for the future of humanity and the earth, based on high-level academic foundations and vision
6. Conduct researches that contribute to the development of human resources who can play a leading and core role in society and researchers who can conduct cutting-edge researches

To achieve the above-mentioned goals, the following *subjects* are emphasized:

- Raise research levels to global standards and aim to activate and advance researches further
- Promote a wide range of researches, from basic researches to cutting-edge researches
- Actively disclose results of not only individual researches but also other researches to return research results to society, promote industry-academic-government research projects and regional partnership research projects, and promote corporatization and commercialization
- Establish an organizational structure in which flexible responses to cross-cutting research systems, including cutting-edge fields and new fields, are possible, despite the basic principle that classes and majors in the same field should be emphasized
- Establish a research system corresponding to the promotion of academic researches
- Endeavor to develop a research performance evaluation system to encourage the teaching staff to carry out research activities
- Aim to develop a research environment in which young researchers can become active
- Conduct self-assessment and actively incorporate results of external evaluations with a view to improving the quality of researches and extending researches to interdisciplinary domains
- Aim to manage research space effectively

Conclusion

A working hypothesis is a hypothesis that is provisionally accepted as a basis for further research in the hope that a tenable theory will be produced, even if the hypothesis ultimately fails. Like all hypotheses, a working hypothesis is constructed as a statement of expectations, which can be linked to the exploratory research purpose in empirical investigation. Working hypotheses are often used as a conceptual framework in qualitative research.

Questions

1. What do you meant by Hypothesis?
2. What are the concepts relating to hypothesis?
3. State the characteristics of hypothesis.
4. What is Null hypothesis?
5. What is Alternative hypothesis?
6. What do you meant by level of significance?
7. State the purposes and objectives of research.

Lesson - 4

RESEARCH DESIGN

Introduction

Before examining types of research designs it is important to be clear about the role and purpose of research design. We need to understand what research design is and what it is not. We need to know where design fits into the whole research process from framing a question to finally analysing and reporting data.

Research design meaning

The most important step after defining the research problem is preparing the design of the research project, which is popularly known as the ‘research design’. A research design helps to decide upon issues like what, when, where, how much, by what means etc. With regard to an enquiry or a research study. A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure. Infact, research design is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data

(Selltiz et al, 1962). Thus, research design provides an outline of what the researcher is going to do in terms of framing the hypothesis, its operational implications and the final data analysis. Specifically, the research design highlights decisions which include:

- ➢ The nature of the study
- ➢ The purpose of the study
- ➢ The location where the study would be conducted
- ➢ The nature of data required
- ➢ From where the required data can be collected
- ➢ What time period the study would cover
- ➢ The type of sample design that would be used
- ➢ The techniques of data collection that would be used
- ➢ The methods of data analysis that would be adopted and
- ➢ The manner in which the report would be prepared

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

- The sampling design that deals with the method of selecting items to be observed for the selected study;
- The observational design that relates to the conditions under which the observations are to be made;
- The statistical design that concerns with the question of how many items are to be observed, and how the information and data gathered are to be analysed; and
- The operational design that deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

Features of Research Design:

The important features of Research Design may be outlined as follows:

- ❖ It constitutes a plan that identifies the types and sources of information required for the research problem;
- ❖ It constitutes a strategy that specifies the methods of data collection and analysis which would be adopted; and
- ❖ It also specifies the time period of research and monetary budget involved in conducting the study, which comprise the two major constraints of undertaking any research

Concepts Relating To Research Design:

Some of the important concepts relating to Research Design are discussed below:

1. Dependent and Independent Variables:

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

2. Extraneous Variables:

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

3. Control:

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

4. Confounded Relationship:

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

5. Research Hypothesis:

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

6. Experimental and Non-experimental Hypothesis Testing Research:

When the objective of a research is to test a research hypothesis, it is known as hypothesis-testing research. Such research may be in the nature of experimental design or non-experimental design. The research in which the independent variable is manipulated is known as 'experimental hypothesis-testing research', whereas the research in which the independent variable is not manipulated is termed as 'non-experimental hypothesis-testing research'.

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

7. Experimental And Control Groups:

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

8. Treatments:

Treatments refer to the different conditions to which the experimental and control groups are subject to. In the example considered, the two treatments are the parents with regular earnings and those with no regular earnings. Likewise, if a research study attempts to examine through an experiment the comparative effect of three different types of fertilizers on the yield of rice crop, then the three types of fertilizers would be treated as the three treatments.

9. Experiment:

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

10. Experimental Unit(s):

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

Characteristics of a good research design:

A good research design often possesses the qualities of being flexible, suitable, efficient, economical and so on. Generally, a research design which minimizes bias and maximizes the reliability of the data collected and analysed is considered a good design (Kothari 1988). A research design which does not allow even the smallest experimental error is said to be the best design for investigation. Further, a research design that yields maximum information and provides an opportunity of viewing the various dimensions of a research problem is considered to be the most appropriate and efficient design. Thus, the question of a good design relates to the purpose or objective and nature of the research problem studied. While a research design may be good, it may not be equally suitable to all studies. In other words, it may be lacking in one aspect or the other in the case of some other research problems. Therefore, no single research design can be applied to all types of research problems.

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

- The methods of gathering the information;
- The skills and availability of the researcher and his/her staff, if any;
- The objectives of the research problem being studied;

- The nature of the research problem being studied; and
- The available monetary support and duration of time for the research work.

Conclusion

This chapter described the research methodology. The purpose of a research design is to maximize valid answers to a research question. This was achieved by using a non-experimental, qualitative, exploratory-descriptive approach that was contextual. The researcher was main data collection instrument. Data was collected by means of interviewing. The researcher made sense of data by using a descriptive method to analyse it and also ensured that the data was trustworthy. Observing the principles of beneficence, human dignity as well as justice ensured that the participants were morally and ethically protected. Chapter 4 discusses the data analysis and findings.

Questions

1. What do you mean by research design ?
2. What are the features of research design?
3. State the concepts of research design.
4. What are the characteristics of good research design?

Lesson – 5

SURVEY RESEARCH AND CASE STUDY RESEARCH

Introduction

A method of sociological investigation that uses question based or statistical surveys to collect information about how people think and act. For example, a possible application of survey research to a business context might involve looking at how effective mass media is in helping form and shift public opinion.

Survey Research Meaning

Survey research is a commonly used method of collecting information about a population of interest. There are many different types of surveys, several ways to administer them, and many methods of sampling. There are two key features of survey research:

- **Questionnaires** -- a predefined series of questions used to collect information from individuals
- **Sampling** -- a technique in which a subgroup of the population is selected to answer the survey questions; the information collected can be generalized to the entire population of interest

I. Questionnaire Design

The two most common types of survey questions are closed-ended questions and open-ended questions.

Closed-Ended Questions

- The respondents are given a list of predetermined responses from which to choose their answer
- The list of responses should include every possible response and the meaning of the responses should not overlap
- An example of a close-ended survey question would be, "Please rate how strongly you agree or disagree with the following statement: 'I feel good about my work on the job.' Do you

strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, or strongly disagree?"

- A Likert scale, which is used in the example above, is a commonly used set of responses for closed-ended questions
- Closed-ended questions are usually preferred in survey research because of the ease of counting the frequency of each response

Open-Ended Questions

- Survey respondents are asked to answer each question in their own words
- Responses are usually categorized into a smaller list of responses that can be counted by the study team for statistical analysis

Considerations for Designing a Questionnaire

- It is important to consider the order in which questions are presented. Sensitive questions, such as questions about income, drug use, or sexual activity, should be put at the end of the survey. This allows the researcher to establish trust before asking questions that might embarrass respondents. Researchers also recommend putting routine questions, such as age, gender, and marital status, at the end of the questionnaire
- Double-barreled questions, which ask two questions in one, should never be used in a survey. An example of a double barreled question is, "Please rate how strongly you agree or disagree with the following statement: 'I feel good about my work on the job, and I get along well with others at work.'" This question is problematic because survey respondents are asked to give one response for two questions
- Researchers should avoid using emotionally loaded or biased words and phrases

Survey Administration

Surveys can be administered in three ways:

- **Through the mail**

  Advantage: Low cost

  Disadvantage: Low response rate
- **By telephone**

  Advantages: Higher response rates; responses can be gathered more quickly

  Disadvantage: More expensive than mail surveys
- **Face-to-face**

  Advantages: Highest response rates; better suited to collecting complex information

  Disadvantage: Very expensive

II. Sampling Procedures

One of the primary strengths of sampling is that accurate estimates of a population's characteristics can be obtained by surveying a small proportion of the population. Four sampling techniques are described here:

1. Simple Random Sampling

- Simple random sampling is the most basic form of sampling
- Every member of the population has an equal chance of being selected
- This sampling process is similar to a lottery: the entire population of interest could be selected for the survey, but only a few are chosen at random
- Researchers often use random-digit dialing to perform simple random sampling. In this procedure, telephone numbers are generated by a computer at random and called to identify individuals to participate in the survey

Cluster Sampling

- Cluster sampling is generally used when it is geographically impossible to undertake a simple random sample
- Cluster sampling requires that adjustments be made in statistical analyses

For example, in a face-to-face interview, it is difficult and expensive to survey households across the nation. Instead, researchers will randomly select geographic areas (for example, counties), then randomly select households within these areas. This creates a cluster sample, in which respondents are clustered together geographically.

2. Stratified Sampling

- Stratified samples are used when a researcher wants to ensure that there are enough respondents with certain characteristics in the sample
- The researcher first identifies the people in the population who have the desired characteristics, then randomly selects a sample of them
- Stratified sampling requires that adjustments be made in statistical analyses

For example, a researcher may want to compare survey responses of African-Americans and Caucasians. To ensure that there are enough Afrian-Americans in the survey, the researcher will first identify the African-Americans in the population and then randomly select a sample of African-Americans.

3. Nonrandom Sampling

- Common nonrandom sampling techniques include convenience sampling and snowball sampling
- Nonrandom samples cannot be generalized to the population of interest. Consequently, it is problematic to make inferences about the population
- In survey research, random, cluster, or stratified samples are preferable

CASE STUDY RESEARCH

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

Criteria for Evaluating Adequacy of Case Study:

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

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

The organic motors of action should be socially relevant. This is to say that the action of the individual cases should be viewed as a series of reactions to social stimuli or situations. To put in simple words, the social meaning of behaviour should be taken into consideration.

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

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

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

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

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

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

Dollard (1935) attempted to express the diverse events depicted in the life-histories of persons during the course of repeated interviews by utilizing psycho-analytical techniques in a given situational context. His criteria of life-history originated directly from this experience. While the life-histories possess independent significance as research documents, the interviews recorded by the investigators can afford, as Dollard observed, "rich insights into the nature of the social situations experienced by them".

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

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

If the life-history is written in first person, it should be as comprehensive and coherent as possible.

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

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

Efforts must be made to verify the reliability of life-history data by examining the internal consistency of the collected material, and by repeating the interviews with the concerned person. Besides this, personal interviews with the persons who are well-acquainted with him/her, belonging to his/her own group should be conducted.

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

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

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

Case histories are filled with valuable information of a personal or private nature. Such information not only helps the researcher to portray the personality of the individual, but also the social background that contributed to it. Besides, it also helps in the formulation of relevant hypotheses.

In general, although Blummer (in Wilkinson and Bhandarkar, 1979) was critical of documentary material, he gave due credit to case histories by acknowledging the fact that the personal documents offer an opportunity to the researcher to develop his/her spirit of enquiry. The analysis of a particular subject would be more effective if the researcher acquires close acquaintance with it through personal documents. However, Blummer also acknowledges the limitations of the personal documents. According to him, such documents do not entirely fulfill the criteria of adequacy, reliability, and representativeness. Despite these shortcomings, avoiding their use in any scientific study of personal life would be wrong, as these documents become necessary and significant for both theory-building and practice.

In spite of these formidable limitations, case study data are used by anthropologists, sociologists, economists and industrial psychiatrists. Gordon Allport (Kothari, 1988) strongly recommends the use of case study data for in-depth analysis of a subject. For, it is one's acquaintance with an individual that instills a desire to know his/her nature and understand them. The first stage involves understanding the individual and all the complexity of his/her nature. Any haste in analyzing and classifying the individual would create the risk of reducing his/her emotional world into artificial bits. As a consequence, the important emotional organizations, anchorages and natural identifications characterizing the personal life of the individual might not yield adequate representation. Hence, the researcher should understand the life of the subject.

Therefore, the totality of life-processes reflected in the well-ordered life-history documents become invaluable source of stimulating insights. Such life-history documents provide the basis for comparisons that contribute to statistical generalizations and help to draw inferences regarding the uniformities in human behaviour, which are of great value. Even if some personal documents do not provide ordered data about personal lives of people, which is the basis of psychological science, they should not be ignored. This is because the final aim of science is to understand, control and make predictions about human life. Once they are satisfied, the theoretical and practical importance of personal documents must be recognized as significant. Thus, a case study may be considered as the beginning and the final destination of abstract knowledge.

Conclusion

When evaluating a case, it is important to be systematic. Analyze the case in a logical fashion, beginning with the identification of operating and financial strengths and weaknesses and environmental opportunities and threats. Move on to assess the value of a company's current strategies only when you are fully conversant with the SWOT analysis of the company. Ask yourself whether the company's current strategies make sense, given its SWOT analysis. If they do not, what changes need to be made? What are your recommendations? Above all, link any strategic recommendations you may make to the SWOT analysis. State explicitly how the strategies you identify take advantage of the company's strengths to exploit environmental opportunities, how they rectify the company's weaknesses, and how they counter environmental threats. Also, do not forget to outline what needs to be done to implement your recommendations.

Questions

1. What is survey research?
2. What are the criteria evaluating the adequacy of case study?
3. What is case study?
4. What are the key factors of survey research?
5. How do you design a questionnaire?

UNIT – II

Lesson – 6

DATA COLLECTION

Introduction

Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a component of research in all fields of study including physical and social sciences, humanities, and business. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.

Meaning

It is important for a researcher to know the sources of data which he 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 refers to the data collected for the first time, whereas secondary data refers to the 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 same data are used by a researcher for his study of a particular problem, then the same data become the secondary data for the researcher. Both the sources of information have their merits and demerits. The selection of a particular source depends upon the (a) purpose and scope of enquiry,availability of time, (c) availability of finance, (d) accuracy required,statistical tools to be used, (f) sources of information (data), and (g) method of data collection.

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

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

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 to know the expected demand for a product newly launched by him and the result of the enquiry that the demand would be meager takes two years to reach him, then the whole purpose of enquiry would become uselessbecause by that time he would have already incurred a huge loss. Thus, in this respect the information is quickly required and hence the researcher has to choose the type of enquiry accordingly.

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 sample size will do.

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) tools of 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, when we buy gold, even a difference of 1/10th 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.

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 define clearly the statistical units on which the data are to be collected.

In number of situations the units are conventionally fixed like the physical units of measurement, such as 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.

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.

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., all the 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 we draw conclusions about the characteristics of the population. Selection of a particular technique becomes difficult because where population or census method is more scientific and 100% accuracy can be attained through this method, choosing this becomes difficult because it is time taking, it requires more labor and it is very expensive. Therefore, for a single researcher or for a small institution it proves to be unsuitable. On the other hand, sample method is less time taking, less laborious and less expensive but a 100% accuracy cannot be attained through this method because of sampling and non-sampling errors attached to this method. Hence, a researcher has to be very cautious and careful while choosing a particular method.

Methods of Collecting Primary Data:

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

- ❖ Direct Personal Interviews.
- ❖ Indirect Oral Interviews.
- ❖ Information from Correspondents.
- ❖ Mailed Questionnaire Methods.
- ❖ Schedule Sent Through Enumerators.

1. Direct Personal Interviews:

A face to face contact is made with the informants (persons fromwhom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information. Thus, if a person wants to collect data about the working conditions of the workers of the Tata Iron and Steel Company, Jamshedpur, he would go to the factory, contact the workers and obtain the desired information. The information collected in this manner is first hand and also original in character. There are many merits and demerits of this method, which are discussed as under:

Merits:

- Most often respondents are happy to pass on the information required from them when contacted personally and thus response is encouraging.
- 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 case the interviewer apprehends that the informant is not giving accurate information, he may cross-examine him and thereby try to obtain the information.
- 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.
- There might be some questions which the interviewer would finddifficult 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.

- 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:

- This method can prove to be expensive if the number of informants is large and the area is widely spread.
- There is a greater chance of personal bias and prejudice under this method as compared to other methods.
- 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.
- 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 on week end. Even during evening only an hour or two can be used for interviews and hence, the work may have to be continued for a long time, or a large number of people may have to be employed which may involve huge expenses.

In the present time of extreme advancement in the communication system, the investigator instead of going personally and conducting a face to face interview may also obtain information 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. Thismethod 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 have to be very short.

Indirect Oral Interviews:

Under this method of data collection, the investigator contacts third parties generally called ‘witnesses’ who are capable of supplying necessary information. This method is generally adopted when the information to be obtained is of a complex nature and informants are not inclined to respond if approached directly. For example, when the researcher is trying to obtain data on drug addiction or the habit of taking liquor, there is high probability that the addicted person will not provide the desired data 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:

- 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 research will become useless.
- The ability of the interviewers to draw information from witnesses by means of appropriate questions and cross-examination.
- 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 theevidence of one person alone is not relied upon. Views from other persons and related agencies should also be ascertained to find the real position .Utmost care must be exercised in the selection of these persons because it is on their views that the final conclusions are reached.

Information from Correspondents:

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

Mailed Questionnaire Method:

Under this method, a list of questions pertaining to the survey which is known as ‘Questionnaire’ is prepared and sent to the various informants by post. Sometimes the researcher himself too contacts the respondents and gets the responses 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:

- The degree to which the questionnaire is formalized or structured.
- The disguise or lack of disguise of the questionnaire and
- 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, when no 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 known as *non- disguised*; on the other hand, when the objective is not clear, the questionnaire is a *disguised one*. On the basis of these two classifications, four types of studies can be distinguished:

- Non-disguised structured,
- Non-disguised non-structured,
- Disguised structured and
- Disguised non-structured.

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

Merits:

- Questionnaire method of data collection can be easily adopted where the field of investigation is very vast and the informants are spread over a wide geographical area.
- This method is relatively cheap and expeditious provided the informants respond in time.
- 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:

This method can be adopted only where the informants are literates so that they can understand written questions and lend the answers in writing.

- It involves some uncertainty about the response. Co-operation on the part of informants may be difficult to presume.
- 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.
    - Prepaid postage stamp should be affixed
    - The sample should be large
    - 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.
    - It should be preferred in such enquiries where there could be a legal compulsion to provide the information.

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 their own handwriting in the questionnaire form. There is difference between questionnaire and schedule. Questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in him self, whereas schedule is the name usually applied to a set of questions which are asked 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:

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

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

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

Demerits:

- In comparison to other methods of collecting primary data, this method is quite costly as enumerators are generally paid persons.
- The success of the method depends largely upon the training imparted to the enumerators.
- Interviewing is a very skilled work and it requires experience and training. Many statisticians have the tendency to neglect this extremely important part of the data collecting process and this result in bad interviews. Without good interviewing most of the information collected may be of doubtful value.
- Interviewing is not only a skilled work but it also requires a great degree of politeness and thus the way the enumerators conduct the interview would affect the data collected. When questions are asked by a number of different interviewers, it is possible that variations in the personalities of the interviewers will cause variation in the answers obtained. This variation will not be obvious. Hence, every effort must be made to remove as much of variation as possible due to different interviewers.

SECONDARY DATA:

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

Sources of Secondary Data:

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

Published sources, and

Unpublished sources.

Published Sources:

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

(a) International publications:

There are some international institutions and bodies like I.M.F, I.B.R.D, I.C.A.F.E and U.N.O who publish regular and occasional reports on economic and statistical matters.

(b) Official Publications of Central and State Governments:

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

(c) Semi-Official Publications:

Semi-Government institutions like Municipal Corporations, District Boards, Panchayats, etc. Publish reports relating to different matters of public concern.

(d) Publications of Research Institutions:

Indian Statistical 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.

Publications of various Commercial and Financial Institutions

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.

Journals and News Papers:

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

Unpublished Sources:

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

Precautions in the use of Secondary Data:

Since secondary data have already been obtained, it is highly desirable that a proper scrutiny of such data is made before they are used by the investigator. In fact the user has to be extra-cautious while using secondary data. In this context Prof. Bowley rightly points out that "Secondary data should not be accepted at their face value." The reason being that data may be erroneous in many respects due to bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors etc. Even if there is no error such data may not be suitable and adequate for the purpose of the enquiry. Prof. 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:

The Suitability of Data:

The investigator must satisfy himself that the data available are suitable for the purpose of enquiry. It can be judged by the nature and scope of the present enquiry with the original enquiry. For example, if the object of the present enquiry is to study the trend in retail prices, and if the data provide only wholesale prices, such data are unsuitable.

(A) Adequacy of Data:

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

(b) Reliability of Data:

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

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

While editing primary data, the following considerations should be borne in mind:

- The data should be complete in every respect
- The data should be accurate

- The data should be consistent, and
- The data should be homogeneous.

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

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 answers to 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. In such questions, the editor should mark 'No answer' in the space provided for answers and if the questions are of vital importance then the schedule or questionnaire should be dropped.

(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 correctness of information. If the information supplied is wrong, conclusions can never be valid. It is, therefore, necessary for the editor to see that the information is accurate in all respects. If the inaccuracy is due to arithmetical errors, it can be easily detected and corrected. But if the cause of inaccuracy is faulty information supplied, it may be difficult to verify it 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 the definitions of various terms used, units to be employed, type of enquiry to 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 of the agency (whether government- state or central-or private institution of an individual).

In using the secondary data, it is best to obtain the data from the primary source as far as possible. By doing so, we would at least save ourselves from the errors of transcription which might have inadvertently crept in the secondary source. Moreover, the primary source will also provide us with detailed discussion about the terminology used, statistical units employed, size of the sample and the technique of sampling (if sampling method was used), methods of data collection and analysis of results and we can ascertain ourselves if these 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 existany secondary data suited to the investigation under study. In some of the investigations both primary as well as secondary data may be used.

Conclusion

Regardless of data type and how they are collected, surveillance systems can be successful if the implementers and end users understand the limitations of both the data and the collection methodology and incorporate that knowledge into their interpretation procedures.

Questions

1. What do you mean by data collection?
2. What is the purpose and scope of enquiry of data collection?
3. What are the methods of data collection?
4. What are the methods of collecting primary data?
5. What is mailed questionnaire?
6. What is secondary data?
7. What are the sources of secondary data?
8. What are the precautions used in the secondary data?
9. Explain the suitability of data in secondary data.

Lesson – 7

QUESTIONNAIRE

Introduction

A questionnaire is a research instrument consisting of a series of questions (or other types of prompts) for the purpose of gathering information from respondents. The questionnaire was invented by the Statistical Society of London in 1838.

Although questionnaires are often designed for statistical analysis of the responses, this is not always the case.Questionnaires have advantages over some other types of surveysin that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Thus, for some demographic groups conducting a survey by questionnaire may not be concrete.

Meaning

Questionnaire is a systematic, data collection technique consists of a series of questions required to be answered by the respondents to identify their attitude, experience, and behavior towards the subject of research.

One of the most critical parts of the survey is the creation of questions that must be framed in such a way that it results in obtaining the desired information from the respondents. There are no scientific principles that assure an ideal questionnaire and in fact, the questionnaire design is the skill which is learned through experience.

Types

A distinction can be made between questionnaires with questions that measure separate variables, and questionnaires with questions that are aggregated into either a scale or index. Questionnaires with questions that measure separate variables, could for instance include questions on:

preferences (e.g. political party)

behaviors (e.g. food consumption)

facts (e.g. gender)

Questionnaires with questions that are aggregated into either a scale or index, include for instance questions that measure:

latent traits

attitudes (e.g. towards immigration)

an index (e.g. Social Economic Status)

Examples

A food frequency questionnaire (FFQ) is a questionnaire the type of diet consumed in people, and may be used as a research instrument. Examples of usages include assessment of intake of vitamins or toxins such as acryl amide.

Usually, a questionnaire consists of a number of questions that the respondent has to answer in a set format. A distinction is made between open-ended and closed-ended questions. An open-ended question asks the respondent to formulate his own answer, whereas a closed-ended question has the respondent pick an answer from a given number of options. The response options for a closed-ended question should be exhaustive and mutually exclusive. Four types of response scales for closed-ended questions are distinguished:

Dichotomous, where the respondent has two options

Nominal-polytomous, where the respondent has more than two unordered options

Ordinal-polytomous, where the respondent has more than two ordered options

(Bounded)Continuous, where the respondent is presented with a continuous scale

A respondent's answer to an open-ended question is coded into a response scale afterwards. An example of an open-ended question is a question where the testie has to complete a sentence (sentence completion item).

Question sequence

In general, questions should flow logically from one to the next. To achieve the best response rates, questions should flow from the least sensitive to the most sensitive, from the factual and behavioural to the attitudinal, and from the more general to the more specific.

There typically is a flow that should be followed when constructing a questionnaire in regards to the order that the questions are asked. The order is as follows:

- Screens
- Warm-ups
- Transitions
- Skips
- Difficult
- Classification

Screens are used as a screening method to find out early whether or not someone should complete the questionnaire. Warm-ups are simple to answer, help capture interest in the survey, and may not even pertain to research objectives. Transition questions are used to make different areas flow well together. Skips include questions similar to "If yes, then answer question 3.

If no, then continue to question ." Difficult questions are towards the end because the respondent is in "response mode." Also, when completing an online questionnaire, the progress bars lets the respondent know that they are almost done so they are more willing to answer more difficult questions. Classification, or demographic question should be at the end because typically they can feel like personal questions which will make respondents uncomfortable and not willing to finish survey.

Basic rules for questionnaire item construction

- Use statements which are interpreted in the same way by members of different subpopulations of the population of interest.
- Use statements where persons that have different opinions or traits will give different answers.
- Think of having an "open" answer category after a list of possible answers.
- Use only one aspect of the construct you are interested in per item.
- Use positive statements and avoid negatives or double negatives.
- Do not make assumptions about the respondent.
- Use clear and comprehensible wording, easily understandable for all educational levels
- Use correct spelling, grammar and punctuation.
- Avoid items that contain more than one question per item (e.g. Do you like strawberries and potatoes?).
- Question should not be biased or even leading the participant towards an answer.

Questionnaire administration modes

Main modes of questionnaire administration include:

- Face-to-face questionnaire administration, where an interviewer presents the items orally.
- Paper-and-pencil questionnaire administration, where the items are presented on paper.
- Computerized questionnaire administration, where the items are presented on the computer.
- Adaptive computerized questionnaire administration, where a selection of items is presented on the computer, and based on the answers on those items, the computer selects following items optimized for the testee's estimated ability or trait.

Concerns with questionnaires

While questionnaires are inexpensive, quick, and easy to analyze, often the questionnaire can have more problems than benefits.

For example, unlike interviews, the people conducting the research may never know if the respondent understood the question that was being asked. Also, because the questions are so specific to what the researchers are asking, the information gained can be minimal.Often, questionnaires such as the Myers-Briggs Type Indicator, give too few options to answer; respondents can answer either option but must choose only one response. Questionnaires also produce very low return rates, whether they are mail or online questionnaires. The other problem associated with return rates is that often the people who do return the questionnaire are those who have a really positive or a really negative viewpoint and want their opinion heard. The people who are most likely unbiased either way typically don't respond because it is not worth their time.

Some questionnaires have questions addressing the participants gender. Seeing someone as male or female is something we all do unconsciously, we don’t give much important to one’s sex or gender as most people use the terms ‘sex’ and ‘gender’ interchangeably, unaware that they are not synonyms. Gender is a term to exemplify the attributes that a society or culture constitutes as masculine or feminine. Although your sex as male or female stands at a biological fact that is identical in any culture, what that specific sex means in reference to your gender role as a ‘woman’ or ‘man’ in society varies cross culturally according to what things are considered to be masculine or feminine. The survey question should really be what is your sex. Sex is traditionally split into two categories, which we typically don’t have control over, you were either born a girl or born a boy and that’s decided by nature. There's also the intersex population which is disregarded in the North American society as a sex. Not many questionnaires have a box for people who fall under Intersex. These are some small things that can be misinterpreted or ignored in questionnaires.

More generally, one key concern with questionnaires is that there may contain quite large measurement errors. These errors can be random or systematic. Random errors are caused by unintended mistakes by respondents, interviewers and/or coders. Systematic error can occur if there is a systematic reaction of the respondents to the scale used to formulate the survey question. Thus, the exact formulation of a survey question and its scale are crucial, since they affect the level of measurement error. Different tools are available for the researchers to help them decide about this exact formulation of their questions, for instance estimating the quality of a question using MTMM experiments or predicting this quality using the Survey Quality Predictor software (SQP). This information about the quality can also be used in order to correct for measurement errors.

Further, if the questionnaires are not collected using sound sampling techniques, often the results can be non-representative of the population—as such a good sample is critical to getting representative results based on questionnaires.

QUESTIONNAIRE DESIGN PROCESS

The following steps are involved in the questionnaire design process:

1. **Specify the Information Needed:** The first and the foremost step in designing the questionnaire is to specify the information needed from the respondents such that the objective of the survey is fulfilled. The researcher must completely review the components of the problem, particularly the hypothesis, research questions, and the information needed.

2. **Define the Target Respondent:** At the very outset, the researcher must identify the target respondent from whom the information is to be collected. The questions must be designed keeping in mind the type of respondents under study. Such as, the questions that are appropriate for serviceman might not be appropriate for a businessman. The less diversified respondent group shall be selected because the more diversified the group is, the more difficult it will be to design a single questionnaire that is appropriate for the entire group.
3. **Specify the type of Interviewing Method:** The next step is to identify the way in which the respondents are reached**.** In personal interviews, the respondent is presented with a questionnaire and interacts face-to-face with the interviewer. Thus, lengthy, complex and varied questions can be asked using the personal interview method. In telephone interviews, the respondent is required to give answers to the questions over the telephone. Here the respondent cannot see the questionnaire and hence this method restricts the use of small, simple and precise questions.

   The questionnaire can be sent through mail or post. It should be self-explanatory and contain all the important information such that the respondent is able to understand every question and gives a complete response. The electronic questionnaires are sent directly to the mail ids of the respondents and are required to give answers online.
4. **Determine the Content of Individual Questions:** Once the information needed is specified and the interviewing methods are determined, the next step is to decide the content of the question. The researcher must decide on what should be included in the question such that it contribute to the information needed or serve some specific purpose.

   In some situations, the indirect questions which are not directly related to the information needed may be asked. It is useful to ask neutral questions at the beginning of a questionnaire with intent to establish respondent's involvement and rapport. This is mainly done when the subject of a questionnaire is sensitive or controversial. The researcher must try to avoid the use of **double-barreled questions**. A question that talks about two issues simultaneously, such as Is the Real juice tasty and a refreshing health drink?
5. **Overcome Respondent's Inability and Unwillingness to Answer:** The researcher should not presume that the respondent can provide accurate responses to all the questions. He must attempt to overcome the respondent's inability to answer.

   The questions must be designed in a simple and easy language such that it is easily understood by each respondent. In situations, where the respondent is not at all informed about the topic of interest, then the researcher may ask the **filter questions**, an initial question asked in the questionnaire to identify the prospective respondents to ensure that they fulfil the requirements of the sample.

Despite being able to answer the question, the respondent is unwilling to devote time in providing information. The researcher must attempt to understand the reason behind such unwillingness and design the questionnaire in such a way that it helps in retaining the respondent's attention.

6. **Decide on the Question Structure:** The researcher must decide on the structure of questions to be included in the questionnaire. The question can be structured or unstructured. The **unstructured questions are the open-ended questions** which are answered by the respondents in their own words. These questions are also called as a **free-response** or **free-answer question**s.

   While, the **structured questions are called as closed-ended questions** that pre-specify the response alternatives. These questions could be a multiple choice question, dichotomous (yes or no) or a scale.

7. **Determine the Question Wording:** The desired question content and structure must be translated into **words which are easily understood** by the respondents. At this step, the researcher must translate the questions in easy words such that the information received from the respondents is similar to what was intended.

   In case the question is written poorly, then the respondent might refuse to answer it or might give a wrong answer. In case, the respondent is reluctant to give answers, then "**nonresponse"** arises which increases the complexity of data analysis. On the other hand, if the wrong information is given, then **"response error"** arises due to which the result is biassed.

8. **Determine the Order of Questions:** At this step, the researcher must decide the **sequence in which the questions are to be asked**. The opening questions are crucial in establishing respondent's involvement and rapport, and therefore, these questions must be interesting, non-threatening and easy. Usually, the **open-ended questions** which ask respondents for their opinions are considered as good opening questions, because people like to express their opinions.

9. **Identify the Form and Layout:** The **format, positioning and spacing** of questions has a significant effect on the results. The layout of a questionnaire is specifically important for the self-administered questionnaires. The questionnaires must be divided into several parts, and each part shall be numbered accurately to clearly define the branches of a question.

10. **Reproduction of Questionnaire:** Here, we talk about the appearance of the questionnaire, i.e. the quality of paper on which the questionnaire is either written or printed. In case, the questionnaire is reproduced on a poor-quality paper; then the respondent might feel the research is unimportant due to which the quality of response gets adversely affected.

Thus, it is recommended to reproduce the questionnaire on a good-quality paper having a professional appearance. In case, the questionnaire has several pages, then it should be presented in the form of a booklet rather than the sheets clipped or stapled together.

11. **Pretesting:** Pretesting means testing the questionnaires on a few selected respondents or a small sample of actual respondents with a purpose of improving the questionnaire by identifying and eliminating the potential problems. All the aspects of the questionnaire must be tested such as question content, structure, wording, sequence, form and layout, instructions, and question difficulty. The researcher must ensure that the respondents in the pretest should be similar to those who are to be finally surveyed.

Conclusion

A well designed questionnaire is essential to a successful survey. However, the researcher must develop his/her own intuition with respect to what constitutes 'good design' since there is no theory of questionnaires to guide him/her.

A good questionnaire is one which help directly achieve the research objectives, provides complete and accurate information; is easy for both interviewers and respondents to complete, is so designed as to make sound analysis and interpretation possible and is brief.

There are at least nine distinct steps: decide on the information required; define the target respondents, select the method(s) of reaching the respondents; determine question content; word the questions; sequence the questions; check questionnaire length; pre-test the questionnaire and develop the final questionnaire.

Questions

1. What do you mean by questionnaire?
2. What are the types of questionnaire?
3. What are the sequence of questionnaire?
4. Explain the basic rules of questionnaire.
5. What are the process of questionnaire?

Lesson -8

SAMPLING

Introduction

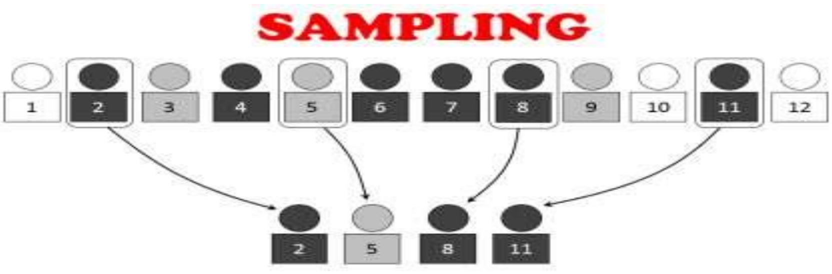
Though sampling is not new, the sampling theory has been developed recently. People knew or not but they have been using the sampling technique in their day to day life. For example a house wife tests a small quantity of rice to see whether it has been well-cooked and gives the generalized result about the whole rice boiling in the vessel. The result arrived at is most of the times 100% correct. In another example, when a doctor wants to examine the blood for any deficiency, takes only a few drops of blood of the patient and examines. The result arrived at is most of the times correct and represent the whole amount of blood available in the body of the patient. In all these cases, by inspecting a few, they simply believe that the samples give a correct idea about the population. Most of our decision are based on the examination of a few items only i.e. Samplestudies. In the words of Croxton and Cowdon, "It may be too expensive or too time consuming to attempt either a complete or a nearly complete coverage in a statistical study. Further to arrive at valid conclusions, it may not be necessary to enumerate all or nearly all of a population. We may study a sample drawn from the large population and if that sample is adequately representative of the population, we should be able to arrive at valid conclusions."

Definition

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

Meaning

Sampling may be defined as the procedure in which a sample is selected from an individual or a group of people of certain kind for research purpose. In sampling, the population is divided into a number of parts called sampling units.



Merits:

1. **It saves time:**

Sampling method of data collection saves time because fewer items are collected and processed. When the results are urgently required, this method is very helpful.

2. It reduces cost:

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

3. More reliable results can be obtained:

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

4. It provides more detailed information:

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

5. 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.

6. Administrative convenience:

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

7. More scientific:

Since the methods used to collect data are based on scientific theory and results obtained can be tested, sampling is 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:

Illusory conclusion:

If a sample enquiry is not carefully planned and executed, the conclusions may be inaccurate and misleading.

Sample Not Representative:

To make the sample representative is a difficult task. If a representative sample is taken from the universe, the result is applicable to the whole population. If the sample is not representative of the universe the result may be false and misleading.

Lack of Experts:

As there are lack of experts to plan and conduct a sample survey, its execution 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 requires more money, labor and time than a census method.

Personal Bias:

There may be personal biases and prejudices with regard to the choice 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 untrue characteristics of the population.

Conditions of Complete Coverage:

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

Essentials of sampling:

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

1. It must be representative:

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

2. Homogeneity:

Selected samples from the universe should have similar nature and should most have any difference when compared with the universe.

3. Adequate samples:

In order to have a more reliable and representative result, a good number of items are to be included in the sample.

4.Optimization:

All efforts should be made to get maximum results both in terms of cost as well as efficiency. If the size of the sample is larger, there is better efficiency and at the same time the cost is more. A proper size of sample is maintained in order to have optimized results in terms of cost and efficiency.

STATISTICAL LAWS

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

THE LAW OF STATISTICAL REGULARITY

The Law of Statistical Regularity is derived from the mathematical theory of probability. According to W.I.King, “the Law of Statistical Regularity formulated in the mathematical theory of probability lays down that a moderately large number of items chosen at random from a very large group are almost sure to have the characteristics of the large group.”

For example, if we want to find out the average income of 10,000 people,we take a sample of 100 people and find the average. Suppose another person takes another sample of 100 people from the same population and finds the average, the average income found out by both the persons will have the least difference. On the other hand if the average income of the same 10,000 people is found out by the census method, the result will be more or less the same.

Characteristics

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

LAW OF INERTIA OF LARGE NUMBERS

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

SAMPLING METHODS

Sampling method refers to the way that observations are selected from apopulation to be in the sample for a sample survey.

Population Parameter vs. Sample Statistic

The reason for conducting a sample survey is to estimate the value of some attribute of a population.

- **Population parameter**. A population parameter is the true value of a population attribute.
- **Sample statistic**. A sample statistic is an estimate, based on sample data, of a population parameter.

Consider this example. A public opinion pollster wants to know the percentage of voters that favor a flat-rate income tax. The *actual* percentage of all the voters is a population parameter. The *estimate* of that percentage, based on sample data, is a sample statistic.

The quality of a sample statistic (i.e., accuracy, precision, representativeness) is strongly affected by the way that sample observations are chosen; that is., by the sampling method.

Probability vs. Non-Probability Samples

As a group, sampling methods fall into one of **two categories.**

- **Probability samples**. With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.
- **Non-probability samples**. With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

Non-probability sampling methods offer two potential advantages - convenience and cost. The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from population parameters. Only probability sampling methods permit that kind of analysis.

Non-Probability Sampling Methods

Two of the main types of non-probability sampling methods are voluntary samples and convenience samples.

- **Voluntary sample**. A voluntary sample is made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey. Suppose, for example, that a news show asks viewers to participate in an on-line poll. This would be a volunteer sample. The sample is chosen by the viewers, not by the survey administrator.

- **Convenience sample**. A convenience sample is made up of people who are easy to reach.

Consider the following example. A pollster interviews shoppers at a local mall. If the mall was chosen because it was a convenient site from which to solicit survey participants and/or because it was close to the pollster's home or business, this would be a convenience sample.

Probability Sampling Methods

The main types of probability sampling methods are simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic random sampling. The key benefit of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid.

- **Simple random sampling**. Simple random sampling refers to any sampling method that has the following properties.
  - The population consists of N objects.
  - The sample consists of n objects.
  - If all possible samples of n objects are equally likely to occur, the sampling method is called simple random sampling.

  There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the N population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.
- **Stratified sampling**. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called **strata**. As a example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.
- **Cluster sampling**. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.

Note the difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters.

- **Multistage sampling**. With multistage sampling, we select a sample by using combinations of different sampling methods.

For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

- **Systematic random sampling**. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* elements on the population list. Thereafter, we select every *kth* element on the list.

Advantages of sampling

Sampling ensures convenience, collection of intensive and exhaustive data, suitability in limited resources and better rapport. In addition to this, sampling has the following advantages also.

1. Low cost of sampling

If data were to be collected for the entire population, the cost will be quite high. A sample is a small proportion of a population. So, the cost will be lower if data is collected for a sample of population which is a big advantage.

2. Less time consuming in sampling

Use of sampling takes less time also. It consumes less time than census technique. Tabulation, analysis etc., take much less time in the case of a sample than in the case of a population.

3. Scope of sampling is high

The investigator is concerned with the generalization of data. To study a whole population in order to arrive at generalizations would be impractical.

Some populations are so large that their characteristics could not be measured. Before the measurement has been completed, the population would have changed. But the process of sampling makes it possible to arrive at generalizations by studying the variables within a relatively small proportion of the population.

4. Accuracy of data is high

Having drawn a sample and computed the desired descriptive statistics, it is possible to determine the stability of the obtained sample value. A sample represents the population from which its is drawn. It permits a high degree of accuracy due to a limited area of operations. Moreover, careful execution of field work is possible. Ultimately, the results of sampling studies turn out to be sufficiently accurate.

5. Organization of convenience

Organizational problems involved in sampling are very few. Since sample is of a small size, vast facilities are not required. Sampling is therefore economical in respect of resources. Study of samples involves less space and equipment.

6. Intensive and exhaustive data

In sample studies, measurements or observations are made of a limited number. So, intensive and exhaustive data are collected.

7. Suitable in limited resources

The resources available within an organization may be limited. Studying the entire universe is not viable. The population can be satisfactorily covered through sampling. Where limited resources exist, use of sampling is an appropriate strategy while conducting marketing research.

8. Better rapport

An effective research study requires a good rapport between the researcher and the respondents. When the population of the study is large, the problem of rapport arises. But manageable samples permit the researcher to establish adequate rapport with the respondents.

Disadvantages of sampling

The reliability of the sample depends upon the appropriateness of the sampling method used. The purpose of sampling theory is to make sampling more efficient. But the real difficulties lie in selection, estimation and administration of samples.

Disadvantages of sampling may be discussed under the heads:

- Chances of bias
- Difficulties in selecting truly a representative sample
- Need for subject specific knowledge
- changeability of sampling units
- impossibility of sampling.

1. Chances of bias

The serious limitation of the sampling method is that it involves biased selection and thereby leads us to draw erroneous conclusions. Bias arises when the method of selection of sample employed is faulty. Relative small samples properly selected may be much more reliable than large samples poorly selected.

2. Difficulties in selecting a truly representative sample

Difficulties in selecting a truly representative sample produces reliable and accurate results only when they are representative of the whole group. Selection of a truly representative sample is difficult when the phenomena under study are of a complex nature. Selecting good samples is difficult.

3. In adequate knowledge in the subject

Use of sampling method requires adequate subject specific knowledge in **sampling technique**. Sampling involves statistical analysis and calculation of probable error. When the researcher lacks specialized knowledge in sampling, he may commit serious mistakes. Consequently, the results of the study will be misleading.

4. Changeability of units

When the units of the population are not in homogeneous, the sampling technique will be unscientific. In sampling, though the number of cases is small, it is not always easy to stick to the, selected cases. The units of sample may be widely dispersed.

Some of the cases of sample may not cooperate with the researcher and some others may be inaccessible. Because of these problems, all the cases may not be taken up. The selected cases may have to be replaced by other cases. Changeability of units stands in the way of results of the study.

5. Impossibility of sampling

Deriving a representative sample is di6icult, when the universe is too small or too heterogeneous. In this case, census study is the only alternative. Moreover, in studies requiring a very high standard of accuracy, the sampling method may be unsuitable. There will be chances of errors even if samples are drawn most carefully.

Meaning ofExperiments

In the scientific method, an experiment is an empirical procedure that arbitrates competing models or hypotheses.Researchers also use experimentation to test existing theories or new hypotheses to support or disprove them.

An experiment usually tests a hypothesis, which is an expectation about how a particular process or phenomenon works. However, an experiment may also aim to answer a "what-if" question, without a specific expectation about what the experiment reveals, or to confirm prior results. If an experiment is carefully conducted, the results usually either support or disprove the hypothesis. According to some philosophies of science, an experiment can never "prove" a hypothesis, it can only add support. On the other hand, an experiment that provides a counterexample can disprove a theory or hypothesis. An experiment must also control the possible confounding factors—any factors that would mar the accuracy or repeatability of the experiment or the ability to interpret the results. Confounding is commonly eliminated through scientific controlsand/or, in randomized experiments, through random assignment.

Types of experiment

Experiments might be categorized according to a number of dimensions, depending upon professional norms and standards in different fields of study. In some disciplines (e.g., psychology or political science), a 'true experiment' is a method of social research in which there are two kinds of variables. The independent variable is manipulated by the experimenter, and the dependent variable is measured. The signifying characteristic of a true experiment is that it randomly allocates the subjects to neutralize experimenter bias, and ensures, over a large number of iterations of the experiment, that it controls for all confounding factors.

Controlled experiments

A controlled experiment often compares the results obtained from experimental samples against control samples, which are practically identical to the experimental sample except for the one aspect whose effect is being tested (the independent variable). A good example would be a drug trial. The sample or group receiving the drug would be the experimental group (treatment group); and the one receiving the placebo or regular treatment would be the control one. In many laboratory experiments it is good practice to have several replicate samples for the test being performed and have both a positive control and a negative control. The results from replicate samples can often be averaged, or if one of the replicates is obviously inconsistent with the results from the other samples, it can be discarded as being the result of an experimental error (some step of the test procedure may have been mistakenly omitted for that sample). Most often, tests are done in duplicate or triplicate. A positive control is a procedure similar to the actual experimental test but is known from previous experience to give a positive result. A negative control is known to give a negative result. The positive control confirms that the basic conditions of the experiment were able to produce a positive result, even if none of the actual experimental samples produce a positive result. The negative control demonstrates the base-line result obtained when a test does not produce a measurable positive result. Most often the value of the negative control is treated as a "background" value to subtract from the test sample results. Sometimes the positive control takes the quadrant of a standard curve.

An example that is often used in teaching laboratories is a controlled protein assay. Students might be given a fluid sample containing an unknown (to the student) amount of protein. It is their job to correctly perform a controlled experiment in which they determine the concentration of protein in the fluid sample (usually called the "unknown sample").

The teaching lab would be equipped with a protein standard solution with a known protein concentration. Students could make several positive control samples containing various dilutions of the protein standard. Negative control samples would contain all of the reagents for the protein assay but no protein. In this example, all samples are performed in duplicate. The assay is a colorimetric assay in which a spectrophotometer can measure the amount of protein in samples by detecting a colored complex formed by the interaction of protein molecules and molecules of an added dye. In the illustration, the results for the diluted test samples can be compared to the results of the standard curve (the blue line in the illustration) to estimate the amount of protein in the unknown sample.

Controlled experiments can be performed when it is difficult to exactly control all the conditions in an experiment. In this case, the experiment begins by creating two or more sample groups that are probabilistically equivalent, which means that measurements of traits should be similar among the groups and that the groups should respond in the same manner if given the same treatment. This equivalency is determined by statistical methods that take into account the amount of variation between individuals and the number of individuals in each group. In fields such as microbiology and chemistry, where there is very little variation between individuals and the group size is easily in the millions, these statistical methods are often bypassed and simply splitting a solution into equal parts is assumed to produce identical sample groups.

Once equivalent groups have been formed, the experimenter tries to treat them identically except for the one variable that he or she wishes to isolate. Human experimentation requires special safeguards against outside variables such as the placebo effect. Such experiments are generally double blind, meaning that neither the volunteer nor the researcher knows which individuals are in the control group or the experimental group until after all of the data have been collected. This ensures that any effects on the volunteer are due to the treatment itself and are not a response to the knowledge that he is being treated.

In human experiments, researchers may give a subject (person) a stimulus that the subject responds to. The goal of the experiment is to measure the response to the stimulus by a test method.

In the design of experiments, two or more "treatments" are applied to estimate the difference between the mean responses for the treatments. For example, an experiment on baking bread could estimate the difference in the responses associated with quantitative variables, such as the ratio of water to flour, and with qualitative variables, such as strains of yeast.

Experimentation is the step in the scientific method that helps people decide between two or more competing explanations – or hypotheses. These hypotheses suggest reasons to explain a phenomenon, or predict the results of an action. An example might be the hypothesis that "if I release this ball, it will fall to the floor": this suggestion can then be tested by carrying out the experiment of letting go of the ball, and observing the results. Formally, a hypothesis is compared against its opposite or null hypothesis ("if I release this ball, it will not fall to the floor"). The null hypothesis is that there is no explanation or predictive power of the phenomenon through the reasoning that is being investigated. Once hypotheses are defined, an experiment can be carried out and the results analysed to confirm, refute, or define the accuracy of the hypotheses.

Natural experiments

The term "experiment" usually implies a controlled experiment, but sometimes controlled experiments are prohibitively difficult or impossible. In this case researchers resort to natural experiments or quasi-experiments. Natural experiments rely solely on observations of the variables of the system under study, rather than manipulation of just one or a few variables as occurs in controlled experiments. To the degree possible, they attempt to collect data for the system in such a way that contribution from all variables can be determined, and where the effects of variation in certain variables remain approximately constant so that the effects of other variables can be discerned. The degree to which this is possible depends on the observed correlation between explanatory variables in the observed data. When these variables are not well correlated, natural experiments can approach the power of controlled experiments. Usually, however, there is some correlation between these variables, which reduces the reliability of natural experiments relative to what could be concluded if a controlled experiment were performed. Also, because natural experiments usually take place in uncontrolled environments, variables from undetected sources are neither measured nor held constant, and these may produce illusory correlations in variables under study.

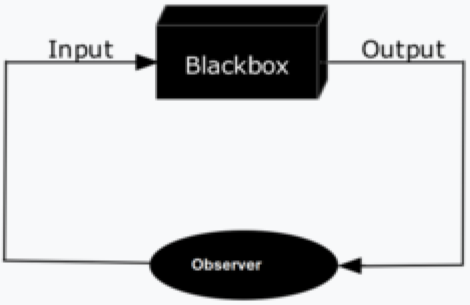
Much research in several science disciplines, including economics, political science, geology, paleontology, ecology, meteorology, and astronomy, relies on quasi-experiments. For example, in astronomy it is clearly impossible, when testing the hypothesis "Stars are collapsed clouds of hydrogen", to start out with a giant cloud of hydrogen, and then perform the experiment of waiting a few billion years for it to form a star. However, by observing various clouds of hydrogen in various states of collapse, and other implications of the hypothesis (for example, the presence of various spectral emissions from the light of stars), we can collect data we require to support the hypothesis.

An early example of this type of experiment was the first verification in the 17th century that light does not travel from place to place instantaneously, but instead has a measurable speed. Observation of the appearance of the moons of Jupiter were slightly delayed when Jupiter was farther from Earth, as opposed to when Jupiter was closer to Earth; and this phenomenon was used to demonstrate that the difference in the time of appearance of the moons was consistent with a measurable speed.

Field experiments

Field experiments are so named to distinguish them from laboratory experiments, which enforce scientific control by testing a hypothesis in the artificial and highly controlled setting of a laboratory. Often used in the social sciences, and especially in economic analyses of education and health interventions, field experiments have the advantage that outcomes are observed in a natural setting rather than in a contrived laboratory environment. For this reason, field experiments are sometimes seen as having higher external validity than laboratory experiments. However, like natural experiments, field experiments suffer from the possibility of contamination: experimental conditions can be controlled with more precision and certainty in the lab. Yet some phenomena (e.g., voter turnout in an election) cannot be easily studied in a laboratory.

Contrast with observational study



The black box model for observation (input and output are *observables*). When there are a feedback with some observer's control, as illustred, the observation is also an experiment.

An observational study is used when it is impractical, unethical, cost-prohibitive (or otherwise inefficient) to fit a physical or social system into a laboratory setting, to completely control confounding factors, or to apply random assignment. It can also be used when confounding factors are either limited or known well enough to analyze the data in light of them (though this may be rare when social phenomena are under examination). For an observational science to be valid, the experimenter must know and account for confounding factors.

In these situations, observational studies have value because they often suggest hypotheses that can be tested with randomized experiments or by collecting fresh data.

Fundamentally, however, observational studies are not experiments. By definition, observational studies lack the manipulation required for Baconian experiments. In addition, observational studies (e.g., in biological or social systems) often involve variables that are difficult to quantify or control. Observational studies are limited because they lack the statistical properties of randomized experiments. In a randomized experiment, the method of randomization specified in the experimental protocol guides the statistical analysis, which is usually specified also by the experimental protocol.Without a statistical model that reflects an objective randomization, the statistical analysis relies on a subjective model.Inferences from subjective models are unreliable in theory and practice. In fact, there are several cases where carefully conducted observational studies consistently give wrong results, that is, where the results of the observational studies are inconsistent and also differ from the results of experiments. For example, epidemiological studies of colon cancer consistently show beneficial correlations with broccoli consumption, while experiments find no benefit.

A particular problem with observational studies involving human subjects is the great difficulty attaining fair comparisons between treatments (or exposures), because such studies are prone to selection bias, and groups receiving different treatments (exposures) may differ greatly according to their covariates (age, height, weight, medications, exercise, nutritional status, ethnicity, family medical history, etc.). In contrast, randomization implies that for each covariate, the mean for each group is expected to be the same. For any randomized trial, some variation from the mean is expected, of course, but the randomization ensures that the experimental groups have mean values that are close, due to the central limit theorem and Markov's inequality. With inadequate randomization or low sample size, the systematic variation in covariates between the treatment groups (or exposure groups) makes it difficult to separate the effect of the treatment (exposure) from the effects of the other covariates, most of which have not been measured. The mathematical models used to analyze such data must consider each differing covariate (if measured), and results are not meaningful if a covariate is neither randomized nor included in the model.

To avoid conditions that render an experiment far less useful, physicians conducting medical trials – say for U.S. Food and Drug Administration approval – quantify and randomize the covariates that can be identified. Researchers attempt to reduce the biases of observational studies with complicated statistical methods such as propensity score matching methods, which require large populations of subjects and extensive information on covariates.

Outcomes are also quantified when possible (bone density, the amount of some cell or substance in the blood, physical strength or endurance, etc.) and not based on a subject's or a professional observer's opinion. In this way, the design of an observational study can render the results more objective and therefore, more convincing.

Ethics

By placing the distribution of the independent variable(s) under the control of the researcher, an experiment – particularly when it involves human subjects – introduces potential ethical considerations, such as balancing benefit and harm, fairly distributing interventions (e.g., treatments for a disease), and informed consent. For example, in psychology or health care, it is unethical to provide a substandard treatment to patients.

Therefore, ethical review boards are supposed to stop clinical trials and other experiments unless a new treatment is believed to offer benefits as good as current best practice. It is also generally unethical (and often illegal) to conduct randomized experiments on the effects of substandard or harmful treatments, such as the effects of ingesting arsenic on human health. To understand the effects of such exposures, scientists sometimes use observational studies to understand the effects of those factors.

Even when experimental research does not directly involve human subjects, it may still present ethical concerns. For example, the nuclear bomb experiments conducted by the Manhattan Project implied the use of nuclear reactions to harm human beings even though the experiments did not directly involve any human subjects.

Conclusion

In conclusion, it can be said that using a sample in research saves mainly on money and time, if a suitable sampling strategy is used, appropriate sample size selected and necessary precautions taken to reduce on sampling and measurement errors, then a sample should yield valid and reliable information. Details on sampling can be obtained from the references included below and many other books on statistics or qualitative research which can be found in libraries.

Questions

1. Define sampling. What are the advantages and disadvantages of sampling?
2. Explain the statistical laws
3. Sampling methods – Explain.
4. What do you mean by non – probability sampling.
5. What is probability sampling
6. What is simple random sampling?
7. What do you mean by cluster sampling?
8. What is multistage sampling?
9. What do you mean by experiment?
10. What are the types of experiment?

Lesson – 9

SAMPLING ERRORS

Introduction

In statistics, sampling error is the error caused by observing a sample instead of the whole population. The sampling error is the difference between asample statistic used to estimate a population parameter and the actual but unknown value of the parameter

Meaning

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

Sampling errors occur primarily due to the following reasons:

Faulty selection of the sample:

Some of the bias is introduced by the use of defective sampling technique for the selection of a sample e.g. Purposive or judgment sampling in which the investigator deliberately selects a representative sample to obtain certain results. This bias can be easily overcome by adopting the technique of simple random sampling.

Substitution:

When difficulties arise in enumerating a particular sampling unit included in the random sample, the investigators usually substitute a convenient member of the population. This obviously leads to some bias since the characteristics possessed by the substituted unit will usually be different from those possessed by the unit originally included in the sample.

Faulty demarcation of sampling units:

Bias due to defective demarcation of sampling units is particularly significant in area surveys such as agricultural experiments in the field of crop cutting surveys etc. In such surveys, while dealing with border line cases, it depends more or less on the discretion of the investigator whether to include them in the sample or not.

Error due to bias in the estimation method:

Sampling method consists in estimating the parameters of the population by appropriate statistics computed from the sample. Improper choice of the estimation techniques might introduce the error.

Variability of the population:

Sampling error also depends on the variability or heterogeneity of the population to be sampled.

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

Biased Errors:

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

Unbiased Errors:

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

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

Reducing Sampling Errors:

Errors in sampling can be reduced if the size of sample is increased.

TYPE I & TYPE II ERROR

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (also known as a "false positive" finding), while a type II error is incorrectly retaining a false null hypothesis (also known as a "false negative" finding).[1] More simply stated, a type I error is to falsely infer the existence of something that is not there, while a type II error is to falsely infer the absence of something that is.

Definition

In statistics, a null hypothesis is a statement that one seeks to nullify with evidence to the contrary. Most commonly it is a statement that the phenomenon being studied produces no effect or makes no difference. An example of a null hypothesis is the statement "This diet has no effect on people's weight." Usually, an experimenter frames a null hypothesis with the intent of rejecting it: that is, intending to run an experiment which produces data that shows that the phenomenon under study does make a difference.

In some cases there is a specific alternative hypothesis that is opposed to the null hypothesis, in other cases the alternative hypothesis is not explicitly stated, or is simply "the null hypothesis is false" – in either event, this is a binary judgment, but the interpretation differs and is a matter of significant dispute in statistics.

A **type I error** (or **error of the first kind**) is the incorrect rejection of a true null hypothesis. Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't. Examples of type I errors include a test that shows a patient to have a disease when in fact the patient does not have the disease, a fire alarm going on indicating a fire when in fact there is no fire, or an experiment indicating that a medical treatment should cure a disease when in fact it does not.

A **type II error** (or **error of the second kind**) is the failure to reject a false null hypothesis. Examples of type II errors would be a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease; a fire breaking out and the fire alarm does not ring; or a clinical trial of a medical treatment failing to show that the treatment works when really it does.

In terms of false positives and false negatives, a positive result corresponds to rejecting the null hypothesis, while a negative result corresponds to failing to reject the null hypothesis; "false" means the conclusion drawn is incorrect. Thus a type I error is a false positive, and a type II error is a false negative.

Differentiation between type I error and type II error.

Type I error

A **type I error** occurs when the null hypothesis (H_0) is true, but is rejected. It is **asserting something that is absent**, a **false hit**. A type I error may be likened to a so-called false positive (a result that indicates that a given condition is present when it actually is not present).

The type I error rate or **significance level** is the probability of rejecting the null hypothesis given that it is true.It is denoted by the Greek letter α (alpha) and is also called the alpha level. Often, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

Type II error

A **type II error** occurs when the null hypothesis is false, but erroneously fails to be rejected. It is **failing to assert what is present**, a **miss**. A type II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss') in a test checking for a single condition with a definitive result of true or false. A Type II error is committed when we fail to believe a true alternative hypothesis. In terms of folk tales, an investigator may fail to see the wolf ("failing to raise an alarm"). Again, H_0: no wolf.

The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1−β).

Table of error types

Tabularised relations between truth/falseness of the null hypothesis and outcomes of the test:[2]

Table of error types		**Null hypothesis (*H_0*) is**	
		True	**False**
Decision About Null Hypothesis (*H_0*)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

Example 1

Hypothesis: "Adding water to toothpaste protects against cavities."

Null hypothesis (H_0): "Adding water to toothpaste has no effect on cavities."

This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.

A type I error occurs when detecting an effect (adding water to toothpaste protects against cavities) that is not present. The null hypothesis is true (i.e., it is true that adding water to toothpaste has no effect on cavities), but this null hypothesis is rejected based on bad experimental data.

Example 2

Hypothesis: "Adding fluoride to toothpaste protects against cavities."

Null hypothesis (H_0): "Adding fluoride to toothpaste has no effect on cavities."

This null hypothesis is tested against experimental data with a view to nullifying it with evidence to the contrary.

A type II error occurs when failing to detect an effect (adding fluoride to toothpaste protects against cavities) that is present. The null hypothesis is false (i.e., adding fluoride is actually effective against cavities), but the experimental data is such that the null hypothesis cannot be rejected.

Example 3

Hypothesis: "The evidence produced before the court proves that this man is guilty."

Null hypothesis (H_0): "This man is innocent."

A type I error occurs when convicting an innocent person (a miscarriage of justice). A type II error occurs when letting a guilty person go free (an error of impunity).

A positive correct outcome occurs when convicting a guilty person. A negative correct outcome occurs when letting an innocent person go free.

Example 4

Hypothesis: "A patient's symptoms improve after treatment A more rapidly than after a placebo treatment."

Null hypothesis (H_0): "A patient's symptoms after treatment A are indistinguishable from a placebo."

A Type I error would falsely indicate that treatment A is more effective than the placebo, whereas a Type II error would be a failure to demonstrate that treatment A is more effective than placebo even though it actually is more effective.

Conclusion

In statistics, sampling error is the error caused by observing a sample instead of the whole population. Random sampling is used precisely to ensure a truly representative sample from which to draw conclusions, in which the same results would be arrived at if one had included the entirety of the population instead.

Questions

1. Define Sampling error.
2. What are the reason for sampling error?
3. What are the types of sampling error?

4. What is Biased sampling?
5. What do you mean by unbiased sampling?
6. Differentiate Type I and Type II error.
7. Define Type I error.
8. What is Type II error

UNIT – III

Lesson - 10

STATISTICAL ANALYSIS

Introduction

With Applications in the Biological and Life Sciences develops a conceptual foundation in statistical analysis while providing readers with opportunities to practice these skills via research-based data sets in biology, kinesiology, and physical anthropology.

PROBABILITY

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

Random Experiment

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

Sample Space

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

An Event

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

Complementary Event

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

Mutually Exclusive Events

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

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

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

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

Independent Event

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

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

Addition Theorem of Probability

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

Multiplication Theorem of Probability

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

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

Both of them speak the truth,

Any one of them speak the truth and

Truth may not be told.

Solution:

The probability of X speaks the truth = 1/5

The probability that X speaks lie = 4/5

The probability that Y speaks the truth = 1/4

The probability that Y speaks lie = ¼

(i) Both of them speak truth = P(X) * P(Y)

= 1/5 * 1/4

= 1/20

ii) (independent event)

Any one of them speak truth = P(X) + P(Y) - P(X*Y)

1/5 + 1/4 - 1/5*1/4

8/20

2/5 (not mutually exclusive events)

(iii)Truth may not be told

1 – p(any one of them speak truth)(complementary event)

1 – 2/5

3/5.

PROBABILITY DISTRIBUTION

If X is discrete random variable which takes the values of x_1,x_2,x_3….. X_n and the corresponding probabilities are p_1, p_2, ……….p_n,then, X follows the probability distribution. The two main properties of probability distribution are: (i) P(Xi) is always greater than or equal to zero and less than or equal to one, and (ii) the summation of probability distribution is always equal to one. For example, tossing of an unbiased coin twice.

Then the probability distribution is

X(Probability of obtaining head)	: 0	1	2
P(xi)	: ¼	½	¼

Expectation of probability

Let X be the discrete random variable which takes the value of x_1, x_2,…… x_n then the respective probability is p_1, p_2, ………… p_n, then the expectation of probability distribution is $p_1x_1 + p_2x_2 +$ ………….. + p_nx_n.

In the above example, the expectation of probability distribution is (0* ¼ 1*1/2+2*¼) =1.

BINOMIAL DISTRIBUTION

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

If an experiment is repeated under the same conditions for a fixed number of trials, say, n.

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

S = [failure, success]

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

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

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

Based on the above two theorems, the probability of success or failure is

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

n !

--------------- . Pxqn-x

x ! (n – x) !

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

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

$$N(p + q)^{n} = Np^{n} + {}^{n}C_{1}p^{n-1}q + {}^{n}C2p^{n-2}q^{2} + \ldots\ldots + {}^{n}C_{r}p^{n-r}q^{r} + \ldots\ldots q^{n}.$$

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

Obtaining Coefficient of the Binomial Distribution:

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

The first term is qn.,

The second term is nC$_1$qn-1p,

In each succeeding term the power of q is reduced by 1 and the power of p is increased by 1.

The coefficient of any term is found by multiplying the coefficient of the preceding term by the power of q in that preceding term, and dividing the products so obtained by one more than the power of p in that proceeding term.

Thus, when we expand (q + p)n, we will obtain the following:-

$(p + q)^n = p^n + {}^nC_1p^{n-1}q + {}^nC_2p^{n-2}q^2 + \ldots\ldots + {}^nC_rp^{n-r}q^r + \ldots\ldots q^n$.

Where, 1, nC_1, nC_2 ……. are called the binomial coefficient. Thus inthe expansion of (p + q)4 we will have (p + q)4 = p4 +4p3q +6p2q2 + 4p1q3 + q4 and the coefficients will be 1, 4, 6, 4, 1.

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

The number of terms in a binomial expansion is always n + 1,

The exponents of p and q, for any single term, when added together, always sum to n.

The exponents of p are n, (n – 1), (n – 2),…….1, 0, respectively and the exponents of q are 0, 1, 2, ……(n – 1), n, respectively.

The coefficients for the n + 1 terms of the distribution are always symmetrical in nature.

Properties of Binomial Distribution

The main properties of binomial distribution are:-

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

The mode of the binomial distribution is equal to the value of xwhich has the largest probability. The mean and mode are equal if np is an integer.

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

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

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

The various constants of binomial distribution are:

Mean	=	np
Standard Deviation	=	√npq
μ_1	=	0
μ_2	=	npq

$$\mu_3 = npq(q - p)$$

$$\mu_4 = 3n2p2q2 + npq(1 - 6pq).$$

$$\text{Skewness} = \frac{(q - p)2}{npq}$$

$$\text{Kurtosis} = 3 + \frac{1 - 6pq}{npq}$$

Illustrations:

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

Solution:

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

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

$(p + q)4 = p4 + 4p3q + 6p2q2 + 4p1q3 + q4$

Therefore, the probability of obtaining 2 heads is

$6p2q2 = 6 \times (½)2(½)2$

$= 3/8$

The probability of obtaining 3 heads is $6p3q1 = 4 \times (½)3(½)1$

$=1/4$

The probability of obtaining 4 heads is $(q)4 = (½)4$

$=1/16$

Therefore, the probability of obtaining 2 or more heads is

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

Illustration:

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

Solution:

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

Using binomial distribution, we have P(r) = $nc_r qn-rpr$

Putting the various values, we have

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

The probability that in a sample of 10, three or less people are vegetarian shall be given by:

$$P(0) + p(1) + p(2) + p(3)$$

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

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

Hence out of 1000 investigators, the number of investigators who will Report 3 or less vegetarians in a sample of 10 is $1000 \times \frac{11}{64} = 172.$

POISSON DISTRIBUTION

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

P, successes, approach zero ($p \to 0$)

np = m is finite.

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

0	e-m

1	me-m/1!
2	m2e-m/ 2!
r	mre-m/ r!
n	mne-m/ n!

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

The poisson distribution is a discrete distribution with a parameter m.

The various constants are:

i. Mean	=	m = p
ii. Standard Deviation	=	√m
iii. Skewness β1	=	1/m
iv. Kurtosis, β2	=	3 + 1/m
v. Variance	=	m

Illustration:

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

Solution:

Total number of misprints 100
M = ------------------------------- = ----- = 1
Total number of pages 100

Probability that a page contains at least two misprints:

P(r≥2) = 1 – [p(0) + p (1)]

mre-m
P(r) = -------
r!

10e-1 1 1
P(0) = ------ = e-1 = ---- = ---------
0! e 2.7183

11e-1 1 1
P(1) = ------ = e-1 = ---- = ---------
1! e 2.7183

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

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

Illustration:

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

Solution:

m = 16

1. S.D. = √m = √16 = 4
2. $^{\beta}1$ = 1/m = 1/16 = 0.625
3. $^{\beta}2$ = 3 + 1/m = 3 + 0.625 = 3.0625
4. $^{\mu}3$ = m = 16
5. $^{\mu}4$ = m + 3m2 = 16 + 3(16)2 = 784

NORMAL DISTRIBUTION

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

The probability function of a Normal Distribution is defined as:

-(x - µ)2 / 2σ2

P(X) = ------------ e

σ√2π

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

Relation Between Binomial, Poisson And Normal Distributions

Binomial, Poisson and Normal Distribution are closely related to one other. When N is large while the probability P of the occurrence of an event is close to zero so that q = (1-p) the binomial distribution is very closely approximated by the Poisson distribution with m = np. The Poisson distribution approaches a normal distribution with standardized variable (x – m)/ √m as m increases to infinity.

Normal distribution and its properties

The important properties of the normal distribution are:-

The normal curve is “bell shaped” and symmetrical in nature. The distribution of the frequencies on either side of the maximum ordinate of the curve is similar with each other.

The maximum ordinate of the normal curve is at x = μ. Hence the mean, median and mode of the normal distribution coincide.

It ranges between - ∞ to + ∞

The value of the maximum ordinate is 1/ σ√2π.

The points where the curve change from convex to concave or vice

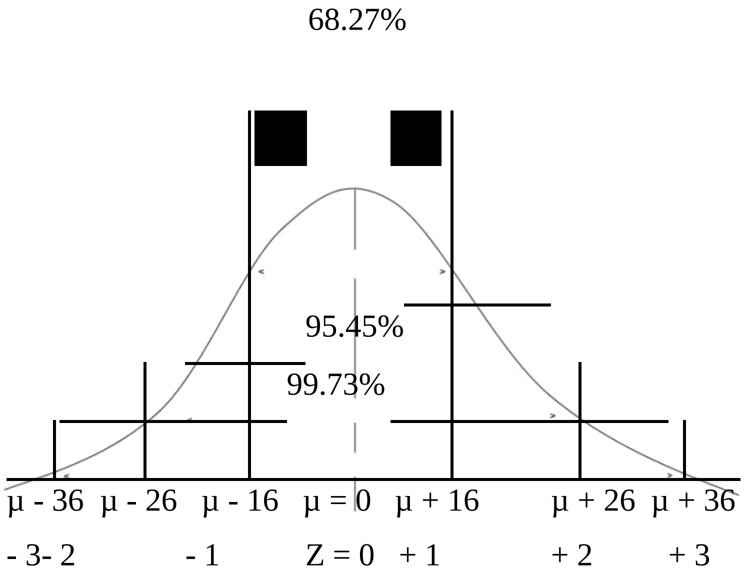
The first and third quartiles are equidistant from median.

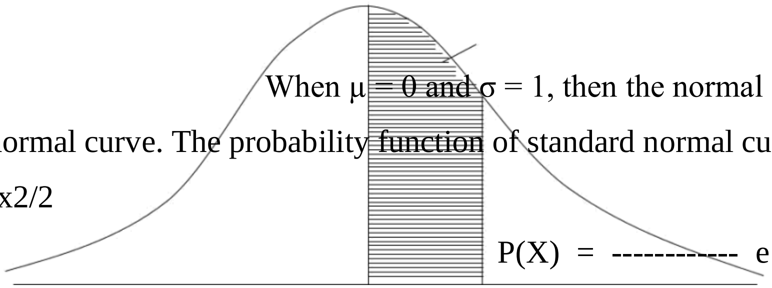
The area under the normal curve distribution are:

μ ± 1σ covers 68.27% area;

μ ± 2σ covers 95.45% area.

μ ± 3σ covers 99.73% area.





When μ = 0 and σ = 1, then the normal distribution will be a standard normal curve. The probability function of standard normal curve is

-x2/2

P(X) = ------------ e

√2π

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

Distance from the mean ordinate in Terms of ± σ	Area under the curve
Z = ± 0.6745	0.50
Z = ± 1.0	0.6826
Z = ± 1.96	0.95
Z = ± 2.00	0.9544
Z = ± 2.58	0.99
Z = ± 3.0	0.9973

All odd moments are equal to zero.

Skewness = 0 and Kurtosis = 3 in normal distribution.

Illustration:

Find the probability that the standard normal value lies between 0 and 1.5

0.4332 (43.32%)

Z = 0 Z = 1.5

As the mean, Z = 0.

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

Illustration:

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

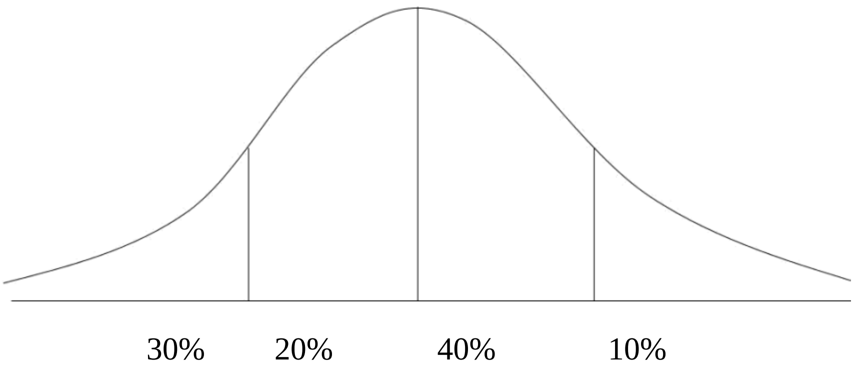
Result	**Percentage of candidates**
Passed with distinction	10
Passed	60
Failed	30

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

Solution:

30% students get marks less than 40.

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



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

10% students get more than 75

40% area = $75 - X = 1.28$ ------------ (ii)

$$= 75 - X = 1.28\sigma$$

Subtract (ii) from (i)

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

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

$$-35 \qquad = -1.8\ \sigma$$

$$35 = 1.8\ \sigma$$

$$1.80\ \sigma = 35$$

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

Mean

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

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

Illustration:

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

Solution:

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

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

For X = 800

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

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

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

Therefore, the percentage = 0.15866 x 100 = 15.86%

ii) when, X = 1200,

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

200

Area between Z = 0 and Z = 1 is 0.34134

Area between X = 400 to X = 600

i.e.,

Z = -1 and Z = 1 is 0.34134 + 0.34134 = 0.6826 = 68.26%

Conclusion

Today, in the information age, an immense amount of data is being accumulated in every aspect of society. These data will be useful only when we can pick out effective information from them. In the near future, statistical methods, which aim at putting data and information to practical use, will become increasingly valuable not only in the areas of science and industry, but also in public administration. The Institute of Statistical Mathematics is making efforts to both develop theory for this and extend the range of application. We are advancing the development of statistics, the basis of science and culture.

Questions

1. Define probability
2. What is random experiment?
3. What is event?
4. What do you mean by complementary event?
5. What is mutually exclusive event?
6. What is independent event?
7. What do you meant by addition and multiplication theorem?
8. What is probability distribution?
9. What is binomial distribution?
10. Explain the properties of binomial distribution
11. What is Poisson distribution?
12. What is normal distribution?
13. Relation between binomial, Poisson and normal distribution – Explain.

Lesson - 11

TESTING OF HYPOTHESIS

Introduction

Previously we used confidence intervals to estimate unknown population parameters. We compared confidence intervals to specified parameter values and when the specific value was contained in the interval, we concluded that there was not evidence of a difference between the population parameter and the specified value. In other words, any values within the confidence intervals were reasonable estimates of the population parameter and any values outside of the confidence intervals were not reasonable estimates. Here, we are going to look at a more formal method for testing whether a given value is a reasonable value of a population parameter. To do this we need to have a hypothesized value of the population parameter.

In this lesson we will compare data from a sample to a hypothesized parameter. In each case, we will compute the probability that a population with the specified parameter would produce a sample statistic as extreme or more extreme to the one we observed in our sample. This probability is known as the p-value and it is used to evaluate statistical significance.

Test of Significance for Large Samples

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

The random sampling distribution of statistics is approximately normal.

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

The Standard Error of Mean.

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

when standard deviation of the population is known, the formula is

$$\text{S.E. } \bar{X} = \frac{\sigma p}{\sqrt{n}}$$

Where,

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

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

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

Where,

σ = Standard deviation of the sample, and n = Sample size

Illustration:

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

Solution:

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

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

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

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

Test of Significance For Small Samples

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

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

The formula is

$$(\overline{X} - \mu)$$

$$t = \frac{\quad}{\sigma} \text{ x } \sqrt{n}$$

where,

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

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 μ = 510 gms.

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

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

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

Df = 20 – 1 = 19 = (10/9) √20 = (10/9) x 4.47 = 44.7/9 = 4.96

Df = 19, $t_{0.01}$ = 3.25

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

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 middle of 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:-

These statistical tests are distribution-free

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

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

The $\chi2$ test is one of the simplest and most widely used non-parametric tests in statistical work. It is defined as:

$\sum(O- E)2$

$X2$ = ------------

E

Where,

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

Steps:

The steps required to determine the value of $\chi2$are:

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

$R X C$

E = ------------

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 ofobservations.

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

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 $\chi2$ which can range from zero to infinity. If $\chi2$ 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 $\chi2$.

The computed value of $\chi2$ is compared with the table value of $\chi2$ for given degrees of freedom at a certain specified level of significance. If at the stated level, the calculated value of $\chi 2$ 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 $\chi 2$ distribution:-i.

The sum of the observed and expected frequencies is always zero.

Symbolically, $\sum(O - E) = \sum O - \sum E \qquad = N - N = 0$

The $\chi 2$ test depends only on the set of observed and expected frequencies and on degrees of freedom v. It is a non-parametric test.

$\chi 2$ distribution is a limiting approximation of the multinomial distribution.

iv. Even though $\chi 2$ 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 $\chi 2$ can be closely approximated by a continuous curve known as the Chi-square distribution. The probability function of $\chi 2$ distribution is:

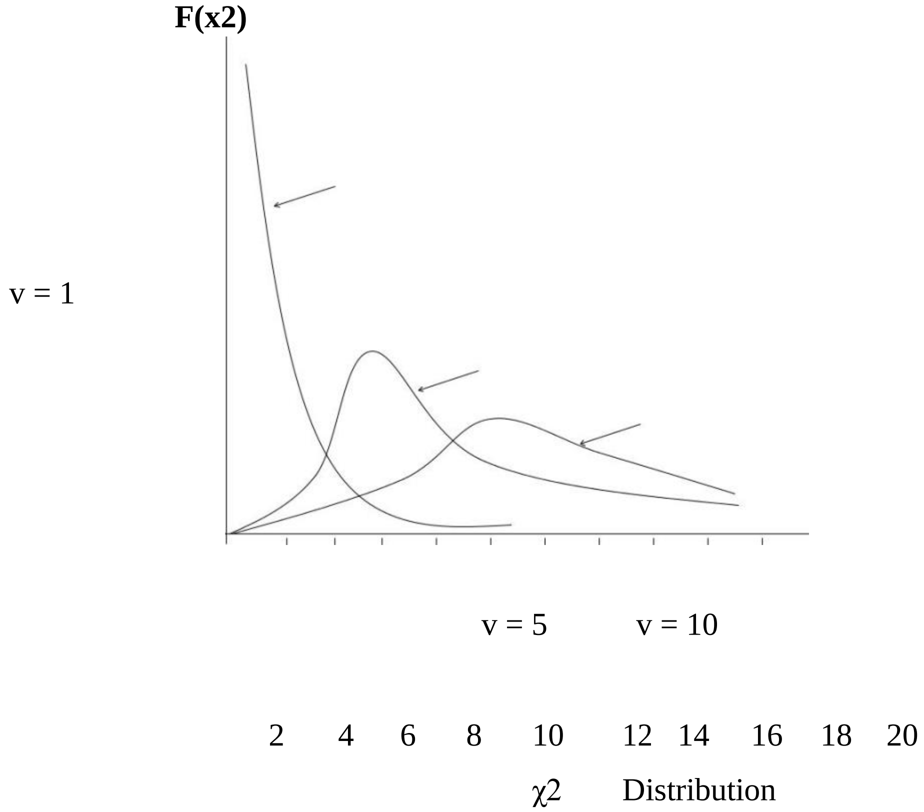
$$F(\chi 2) = C\ (\chi 2)(v/2 - 1)e - x2/2$$

Where

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

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

The following diagram gives χ2 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 χ2 Distribution

The main properties of χ2 distribution are:-

The mean of the χ2 distribution is equal to the number of degrees of freedom,

i.e.,

$$X = \overline{v}$$

The variance of the χ2 distribution is twice the degrees of freedom, Variance = 2v

(iii) $\mu_1 = 0,$

(iv) $\mu_2 = 2v,$

(v) $\mu_3 = 8v,$

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

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

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

The table values of χ2 are available only up to 30 degrees of freedom. For degrees of freedom greater than 30, the distribution of χ2 approximates the normal distribution. For degrees of freedom greater than 30, the approximation is acceptable close. The mean of the distribution √2χ2 is √2v – 1, and the standard deviation is equal to 1. Thus the application of the test is simple, for deviation of √2χ2 from √2v – 1 may be interpreted as a normal deviate with units standard deviation. That is,

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

Alternative Method Of Obtaining The Value of χ2

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

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

N is the total frequency and ad the larger cross-product, the value of χ2 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

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

Conditions for Applying χ2 Test:

The main conditions considered for employing the χ2 test are:

N must be to ensure the similarity between theoretically correct distribution and our sampling distribution of χ2.

No theoretical cell frequency should be small when the expected frequencies are too small. If it is so, then the value of χ2 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.

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 ∑O = ∑E = N.

Uses of χ2 test:

The main uses of χ2 test are:

χ2 test as a test of independence. With the help of χ2test, we can findout 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 colour of husband and wife. To examine the attributes that are associated, we formulate the null hypothesis that there is no association against an alternative hypothesis and that there is an association between the attributes under study. If the calculated value of χ2 is less than the table 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 χ2 is greater than the table value at a certain level of significance, the results of the experiment do not support the hypothesis.

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

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

Illustration:

In an anti-diabetes campaign in a certain area, a particular medicine, say x was administered to 812 persons out of a total population of 3248. The number of diabetes 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 $\chi 2$ test :

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

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
			[∑(O – E)2/E] = 38.593

$\chi2$ = [∑(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 $\chi2$ is greater than the table value. The hypothesis is rejected. Hence medicine x is useful in checking malaria.

Illustration:

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

	Affected	not affected
Inoculated	10	20
Not inoculated	15	5

Calculate $\chi2$ and discuss the effect of vaccine in controlling susceptibility to tuberculosis (5% value of $\chi2$ for one degree of freedom = 3.84).

Solution:

Let us take the hypothesis that the vaccine is not effective in controlling susceptibility to tuberculosis. Applying $\chi 2$ test:

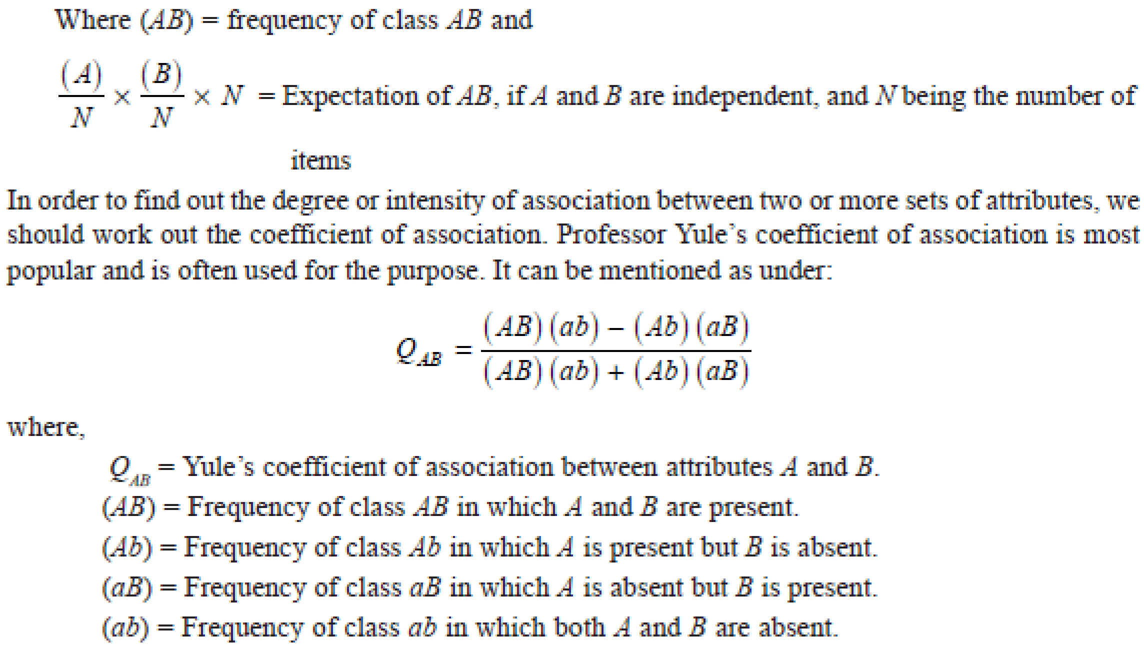
$$\chi 2 = \frac{N(ad - bc)2}{(a+b)\,(c+d)(a+c)(b+d)} = \frac{50\,(11x5 - 20x15)2}{30x20x25x25} = 8.3$$

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

ASSOCIATION IN CASE OF ATTRIBUTES

When data is collected on the basis of some attribute or attributes, we have statistics commonly termed as statistics of attributes. It is not necessary that the objects may process only one attribute; rather it would be found that the objects possess more than one attribute. In such a situation our interest may remain in knowing whether the attributes are associated with each other or not. For example, among a group of people we may find that some of them are inoculated against small-pox and among the inoculated we may observe that some of them suffered from small-pox after inoculation. The important question which may arise for the observation is regarding the efficiency of inoculation for its popularity will depend upon the immunity which it provides against small-pox. In other words, we may be interested in knowing whether inoculation and immunity from small-pox are associated.

Technically, we say that the two attributes are associated if they appear together in a greater number of cases than is to be expected if they are independent and not simply on the basis that they are appearing together in a number of cases as is done in ordinary life. The association may be positive or negative (negative association is also known as disassociation). If class frequency of AB, symbolically written as (AB), is greater than the expectation of AB being together if they are independent, then we say the two attributes are positively associated; but if the class frequency of AB is less than this expectation, the two attributes are said to be negatively associated. In case the class frequency of AB is equal to expectation, the two attributes are considered as independent i.e., are said to have no association. It can be put symbolically as shown hereunder:



Where (AB) = frequency of class AB and

$\frac{(A)}{N} \times \frac{(B)}{N} \times N$ = Expectation of AB, if A and B are independent, and N being the number of items

In order to find out the degree or intensity of association between two or more sets of attributes, we should work out the coefficient of association. Professor Yule's coefficient of association is most popular and is often used for the purpose. It can be mentioned as under:

$$Q_{AB} = \frac{(AB)(ab) - (Ab)(aB)}{(AB)(ab) + (Ab)(aB)}$$

where,

Q_{AB} = Yule's coefficient of association between attributes A and B.
(AB) = Frequency of class AB in which A and B are present.
(Ab) = Frequency of class Ab in which A is present but B is absent.
(aB) = Frequency of class aB in which A is absent but B is present.
(ab) = Frequency of class ab in which both A and B are absent.

The value of this coefficient will be somewhere between +1 and –1. If the attributes are completely associated (perfect positive association) with each other, the coefficient will be +1, and if they are completely disassociated (perfect negative association), the coefficient will be –1. If the attributes are completely independent of each other, the coefficient of association will be 0. The varying degrees of the coefficients of association are to be read and understood according to their positive and negative nature between +1 and –1. Sometimes the association between two attributes, A and B, may be regarded as unwarranted when we find that the observed association between A and B is due to the association of both A and B with another attribute C. For example, we may observe positive association between inoculation and exemption for small-pox, but such association may be the result of the fact that there is positive association between inoculation and richer section of society and also that there is positive association between exemption from small-pox and richer section of society. The sort of association between A and B in the population of C is described as partial association as distinguished from total association between A and B in the overall universe. We can workout the coefficient of partial association between A and B in the population of C by just modifying the above stated formula for finding association between A and B as shown below:

$$Q_{AB.C} = \frac{(ABC)(abC) - (AbC)(aBC)}{(ABC)(abC) + (AbC)(aBC)}$$

where, QAB.C = Coefficient of partial association between A and B in the population of C; and all other values are the class frequencies of the respective classes (A, B, C denotes the presence of concerning attributes and a, b, c denotes the absence of concerning attributes).

At times, we may come across cases of illusory association, wherein association between two attributes does not correspond to any real relationship.

This sort of association may be the result of some attribute, say C with which attributes A and B are associated (but in reality there is no association between A and B). Such association may also be the result of the fact that the at=tributes A and B might not have been properly defined or might not have been correctly recorded. Researcher must remain alert and must not conclude association between A and B when in fact there is no such association in reality. In order to judge the significance of association between two attributes, we make use of Chisquare test by finding the value of Chi-square (c2) and using Chi-square distribution the value of c2 can be worked out as under:

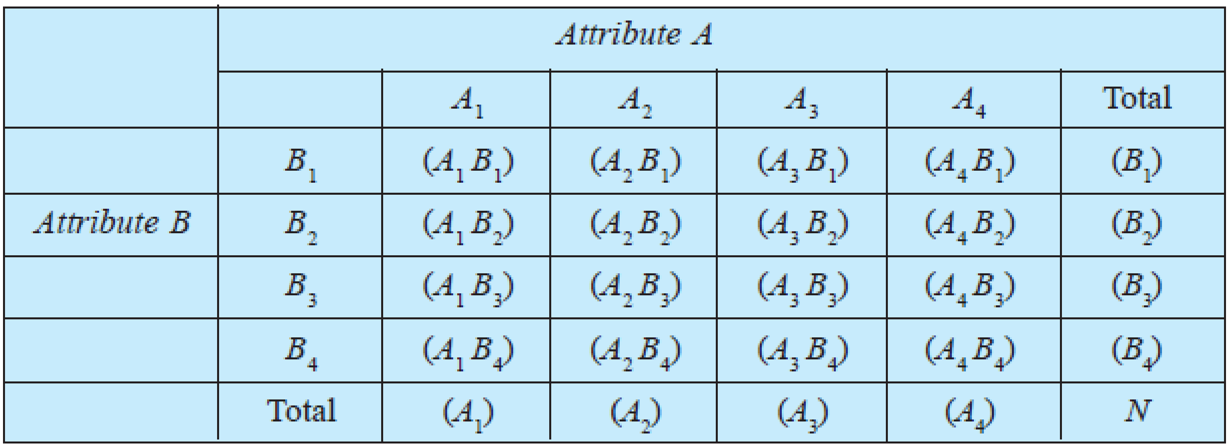
$$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad i = 1, 2, 3 \ldots$$

where $\qquad j = 1, 2, 3 \ldots$

O_{ij} = observed frequencies
E_{ij} = expected frequencies.

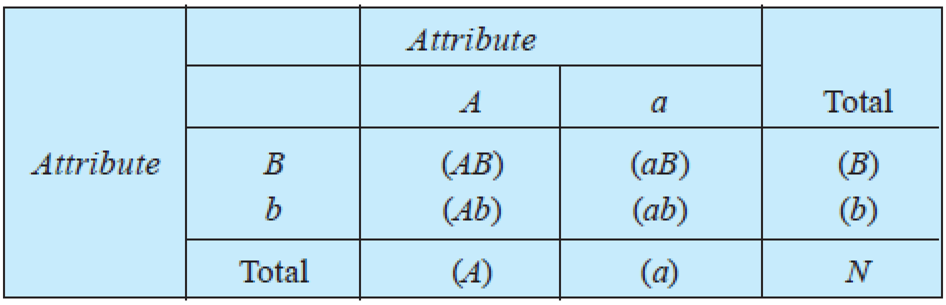
Association between two attributes in case of manifold classification and the resulting contingency table can be studied as explained below:

We can have manifold classification of the two attributes in which case each of the two attributes are first observed and then each one is classified into two or more subclasses, resulting into what is called as contingency table. The following is an example of 4 × 4 contingency table with two attributes A and B, each one of which has been further classified into four sub-categories.



	Attribute A					
		A_1	A_2	A_3	A_4	Total
	B_1	$(A_1 B_1)$	$(A_2 B_1)$	$(A_3 B_1)$	$(A_4 B_1)$	(B_1)
Attribute B	B_2	$(A_1 B_2)$	$(A_2 B_2)$	$(A_3 B_2)$	$(A_4 B_2)$	(B_2)
	B_3	$(A_1 B_3)$	$(A_2 B_3)$	$(A_3 B_3)$	$(A_4 B_3)$	(B_3)
	B_4	$(A_1 B_4)$	$(A_2 B_4)$	$(A_3 B_4)$	$(A_4 B_4)$	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

Association can be studied in a contingency table through Yule's coefficient of association as stated above, but for this purpose we have to reduce the contingency table into 2 × 2 table by combining some classes. For instance, if we combine (A1) + (A2) to form (A) and (A3) + (A4) to form (a) and similarly if we combine (B1) + (B2) to form (B) and (B3) + (B4) to form (b) in the above contingency table, then we can write the table in the form of a 2 × 2 table as shown in Table below



		Attribute		Total
		A	*a*	
Attribute	*B*	(*AB*)	(*aB*)	(*B*)
	b	(*Ab*)	(*ab*)	(*b*)
	Total	(*A*)	(*a*)	*N*

After reducing a contingency table in a two-by-two table through the process of combining some classes, we can work out the association as explained above. But the practice of combining classes is not considered very correct and at times it is inconvenient also, Karl Pearson has suggested a measure known as Coefficient of mean square contingency for studying association in contingency tables. This can be obtained as under:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

where

C = Coefficient of contingency

χ^2 = Chi-square value which is $= \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$

N = number of items.

This is considered a satisfactory measure of studying association in contingency tables.

Conclusion

The statistical decision consists of rejecting or not rejecting the null hypothesis. It is rejected if the computed value of the test statistic falls in the rejection region, and it is not rejected if the computed value of the test statistic falls in the nonrejection region. If is rejected, we conclude that is true.

Questions

1. State the test of significance for large sample.
2. What is standard error of mean?
3. State the test of significance for Small sample.
4. What do you mean by Chi – square test?
5. State the properties of Chi – square distribution.
6. What are the conditions for apply Chi – square test?
7. What are the uses of Chi – square test?
8. What is association of attributes?

Lesson – 12

STANDARD DEVIATION

Introduction

In statistics, the standard deviation (SD, also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

Meaning

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler, though in practice less robust, than the average absolute deviation. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data. There are also other measures of deviation from the norm, including average absolute deviation, which provide different mathematical properties from standard deviation.

Applications of Standard Deviation

Standard deviation is widely used in experimental and industrial settings to test models against real-world data. An example of this in industrial applications is quality control for some product. Standard deviation can be used to calculate a minimum and maximum value within which some aspect of the product should fall some high percentage of the time. In cases where values fall outside the calculated range, it may be necessary to make changes to the production process to ensure quality control.

Standard deviation is also used in weather to determine differences in regional climate. Imagine two cities, one on the coast and one deep inland, that have the same mean temperature of 75°F. While this may prompt the belief that the temperatures of these two cities are virtually the same, the reality could be masked if only the mean is addressed and the standard deviation ignored. Coastal cities tend to have far more stable temperatures due to regulation by large bodies of water, since water has a higher heat capacity than land; essentially, this makes water far less susceptible to changes in temperature, and coastal areas remain warmer in winter, and cooler in summer due to the amount of energy required to change the temperature of water. Hence, while the coastal city may have temperature ranges between 60°F and 85°F over a given period of time to result in a mean of 75°F, an inland city could have temperatures ranging from 30°F to 110°F to result in the same mean.

Another area in which standard deviation is largely used is finance, where it is often used to measure the associated risk in price fluctuations of some asset or portfolio of assets. The use of standard deviation in these cases provides an estimate of the uncertainty of future returns on a given investment. For example, in comparing stock A that has an average return of 7% with a standard deviation of 10% against stock B, that has the same average return but a standard deviation of 50%, the first stock would clearly be the safer option, since standard deviation of stock B is significantly larger, for the exact same return. That is not to say that stock A is definitively a better investment option in this scenario, since standard deviation can skew the mean in either direction. While Stock A has a higher probability of an average return closer to 7%, Stock B can potentially provide a significantly larger return (or loss).

FORMULAE FOR THE STANDARD DEVIATION

Whilst it is not necessary to learn the formula for calculating the standard deviation, there may be times when you wish to include it in a report or dissertation.

The standard deviation of an entire population is known as **σ** (sigma) and is calculated using:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Where **x** represents each value in the population, **μ** is the mean value of the population, **Σ** is the summation (or total), and **N** is the number of values in the population.

The standard deviation of a sample is known as **S** and is calculated using:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Where **x** represents each value in the population, **x** is the mean value of the sample, **Σ** is the summation (or total), and **n-1** is the number of values in the sample minus 1.

Problem :1 The annual salaries of a group of employees are given in the following table:

Salaries (in Rs.000)	45	50	55	60	65	70	75	80
Number of persons	3	5	8	7	9	7	4	7

Calculate the standard deviation of the salaries

Solution

CALCULATION OF STANDARD DEVIATION

Salaries X	No of persons	(X – 60)/5	fd	Fd^2
45	3	-3	-9	27
50	5	-2	-10	20
55	8	-1	-8	8
60	7	0	0	0
65	9	+1	+9	9
70	7	+2	+14	28
75	4	+3	+12	36
80	7	+4	+28	112
	N = 50		∑fd = 36	∑fd$^{2\,=}$ 240

Formula

$\sigma = \surd \sum fd^2/N – (\sum fd/N)^2$ X i

$\sigma = \surd$ 240/50 – $(36/50)^2$ X5

= 10.35

Problem – 2

Calculate mean and standard deviation of following frequency distribution of marks

Marks	No of Students	Marks	No of Students
0 – 10	5	40 – 50	50
10 – 20	12	50 – 60	37
20 – 30	30	60 – 70	21
30 – 40	45		

Solution

CALCULATION OF MEAN AND STADARD DEVIATION

Marks	m.p	f	(m – 35)/10	fd	fd^2
	m		D		
0 – 10	5	5	-3	-15	45
10 – 20	15	12	-2	-24	48
20 – 30	25	30	-1	-30	30
30 – 40	35	45	0	0	0
40 – 50	45	50	+1	+50	50
50 – 60	55	37	+2	+74	148
60 – 70	65	21	+3	+63	189
		N = 200		∑fd = 118	$\sum fd^2 = 510$

$\bar{X} = A + (\sum fd/N)\ I = 35 + (118/200) \times 10 = 35 + 5.9 = 40.9$

$\sigma = \sqrt{}\ \sum fd^2/N – (\sum fd/N)^2 \times i$

$\sigma = \sqrt{}\ 510/200 – (118/200)^2 \times 10$

$= 14.839$

Conclusion

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

Questions

1. What do you mean by standard deviation?
2. What are the Application of standard deviation?

Lesson – 13

CO-EFFICIENT OF VARIATIONS

Introduction

The standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series.

Meaning

The coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable, or less homogeneous. On the other hand, series for which coefficient of variation is less is said to beless variable or more consistent, more uniform, more stable or more homogeneous. Coefficient of variation is denoted by C.V and is obtained as follows

Coefficient of variation or C.V = $= \frac{\sigma}{X} \text{ x } 100$

It may be pointed out that although any measure of dispersion an be sued in conjunction with any average in computing relative dispersion, statisticians, in fact, almost always use the standard deviation as the measure of dispersion and the arithmetic mean as the average. When the relative dispersion is stated in terms of the arithmetic mean and the standard deviation the resulting percentage is known as the coefficient of variation or coefficient of variability

Problem:1

The following table shows that monthly expenditures of 80 students of a university on morning breakfast:

Expenditure (in Rs)	No of students	Expenditure(in Rs)	No of Students
78 – 82	2	53 – 57	13
73 – 77	6	48 – 52	9
68 – 72	7	43 – 47	7
63 – 67	12	38 – 42	4
58 – 62	18	33 – 37	2

Calculate arithmetic mean, standard deviation and coefficient of variation of the above data

Solution:

CCALCULATION OF $\overline{X}$ S.D and C.V

Expenditure	m.p	F	(m – 60)/5	fd	fd^2
(Rs)	m		D		
78 – 82	80	2	+4	+8	32
73 – 77	75	6	+3	+18	65
68 – 72	70	7	+2	+14	28
63 – 67	65	12	+1	+12	12
58 – 62	60	18	0	0	0
53 – 57	55	13	-1	-13	13
48 – 52	60	9	-2	-18	36
43 – 47	45	7	-3	-21	63
38 – 42	40	4	-4	-16	64
33 - 37	35	2	-5	-10	50
		N = 80		∑fd = -26	∑fd^2 = 352

Mean : $\bar{X} = A + \frac{\sum fd}{N} x i = 60 - \frac{26}{80} x\ 5 = 60 - 1.625 = 58.375$

S.D : $\sigma = \sqrt{\ } \sum fd^2/N - (\sum fd/N)^2 \text{ x I} = \sqrt{\ } 352/80 - (-26/80)^2 \text{ X } 5$

$\sqrt{\ } 4.4 - .106 \text{ x } \overline{5 = 2.07}2 \text{ x } 5 = 10.36$

C.V = $\frac{\sigma}{\bar{X}}$X100 = $\frac{10.36}{58.375}$ X 100 = 17.75%

Problem:2

From the prices of shares of X and Y below find out which is more stable in value

X	:	35	54	52	53	56	58	52	50	51	49
Y	:	108	107	105	105	106	107	104	103	104	101

In order to find out which shares are more stable, we have to compare coefficient of variations.

Solutions:

CALCULATION OF COEFFICIENT OF VARIATION

X	(x – x) X	x^2	Y	(y-y) Y	y^2
35	-16	256	108	+3	9
54	+3	9	107	+2	4
52	+1	1	105	0	0
53	+2	4	105	0	0
56	+5	25	106	+1	1
58	+7	49	107	+2	4
52	+1	1	104	-1	1
20	-1	1	103	-2	4
21	0	0	104	-1	1
49	2	4	101	-4	16
∑X = 510	∑x = 0	$\sum X^2 = 350$	∑Y = 1050	∑y = 0	$\sum y^2 = 40$

Coefficient of variation X:

$$C.V = \frac{\sigma}{X} X100$$

$$\bar{X} = \frac{\sum X}{N} = \frac{510}{10} = 51$$

$$\sigma = \sqrt{\sum X^2/N} = \sqrt{350/10} = 5.916$$

$$C.V = \frac{5.916}{51} \times 100 = 11.6$$

Coefficient of Variation Y

$$C.V = \frac{\sigma}{X} X100$$

$$\bar{X} = \frac{\sum Y}{N} = \frac{1050}{10} = 105$$

$$\sigma = \sqrt{\sum Y^2/N} = \sqrt{40/10} = 2$$

$$C.V = \frac{2}{105} \times 100 = 1.905$$

MERITS AND LIMITATIONS

Merits

The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of sampling than most other measures of dispersion.

- It is possible to calculate the combined standard deviation of two or more groups. This is not possible with anyother measure.
- For comparing the variability of two more distributions coefficient of variation is considered to be more appropriate and this is based on mean and standard deviation.
- Standard deviation is most prominently used in further statistical work. For example, in computing skewness, correlation, etc., use is made of standard deviation. It is keynote in sampling and provides a unit of measurement for the normal distribution

Limitations:

As compared to other measures it is difficult to compute. However, this does not reduce the importance of this measure because of high degree of accuracy of results it gives.

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

Conclusion

There was a disagreement between the proposed classification and the classification by Pimentel Gomes, showing the need for the analyzed variables to have their own ranges of classification. The magnitude of the coefficients of variation (CV) obtained in the experiments with sugarcane varies according to the nature of the variables. The percentage of sucrose (PSU) presented the lower limits in ranges of CV, while the variable tons of stalks per hectare (TSH) and tons of pol per hectare (TPH) presented the highest limits in ranges of coefficient of variation. The CV upper limits for considering a sugarcane experiment as of good to high precision is 10, 15 and 19% for PSU, TSH and TPH, respectively.

Question

1. What is co – efficient of variation ?
2. Explain the merits and demerits of co-efficient of variation.

UNIT - IV

Lesson – 14

STATISTICAL APPLICATIONS

INTRODUCTION

A manager in a business organization – whether in the top level, or the middle level, or the bottom level - has to perform an important role of decision making. For solving any organizational problem – which most of the times happens to be complex in nature -, he has to identify a set of alternatives, evaluate them and choose the best alternative. The experience, expertise, rationality and wisdom gained by the manager over a period of time will definitely stand in good stead in the evaluation of the alternatives available at his disposal. He has to consider several factors, sometimes singly and sometimes jointly, during the process of decision making. He has to deal with the data of not only his organization but also of other competing organizations.

Meaning

It would be a challenging situation for a manager when he has to face so many variables operating simultaneously, something internal and something external. Among them, he has to identify the important variables or the dominating factors and he should be able to distinguish one factor from the other. He should be able to find which factors have similar characteristics and which factors stand apart. He should be able to know which factors have an inter play with each other and which factors remain independent. It would be advantageous to him to know whether there is any clear pattern followed by the variables under consideration. At times he may be required to have a good idea of the values that the variables would assume in future occasions. The task of a manager becomes all the more difficult in view of the risks and uncertainties surrounding the future events. It is imperative on the part of a manager to understand the impact of various policies and programmes on the development of the organization as well as the environment. Also he should be able to understand the impact of several of the environmental factors on his organization. Sometimes a manager has to take a single stage decision and at times he is called for to take a multistage decision on the basis of various factors operating in a situation.

Statistical analysis is a tool for a manager in the process of decision making by means of the data on hand. All managerial activities involve an analysis of data. Statistical approach would enable a manager to have a scientific guess of the future events also. Statistical methods are systematic and built by several experts on firmly established theories and consequently they would enable a manager to overcome the uncertainties associated with future occasions.

However, statistical tools have their shortcomings too. The limitations do not reflect on the subject. Rather they shall be traced to the methods of data collection and recording of data.

Even with highly sophisticated statistical methods, one may not arrive at valid conclusions if the data collected are devoid of representative character.

In any practical problem, one has to see whether the assumptions are reasonable or not, whether the data represents a wide spectrum, whether the data is adequate, whether all the conditions for the statistical tests have been fulfilled, etc. If one takes care of these aspects, it would be possible to arrive at better alternatives and more reliable solutions, thereby avoiding future shocks. While it is true that a statistical analysis, by itself, cannot solve all the problems faced by an organization, it will definitely enable a manager to comprehend the ground realities of the situation. It will for sure provide a foresight in the identification of the crucial variables and the key areas so that he can locate a set of possible solutions within his ambit. A manager has to have a proper blend of the statistical theories and practical wisdom and he shall always strive for a holistic approach to solve any organizational problem. A manager has to provide some safe-guarding measures against the limitations of the statistical tools. In the process he will be able to draw valid inferences thereby providing a clue as to the direction in which the organization shall move in future. He will be ably guided by the statistical results in the formulation of appropriate strategies for the organization. Further, he can prepare the organization to face the possible problems of business fluctuations in future and minimize the risks with the help of the early warning signals indicated by the relevant statistical tools.

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

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

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

The interpretation of the results from statistical analysis occupies an important place. Statistics is concerned with the aggregates and not just the individual data items or isolated measurements of certain variables. Therefore the conclusions from a statistical study will be valid for a majority of the objects and normal situations only. There are always extreme cases in any problem and they have to be dealt with separately. Statistical tools will enable a manager to identify such outliers (abnormal cases or extreme variables) in a problem. A manager has to evaluate the statistical inferences, interpret them in the proper context and apply them in appropriate situations.

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

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

SIMPLE CORRELATION

Correlation

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

Simple correlation

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

TYPES OF CORRELATION

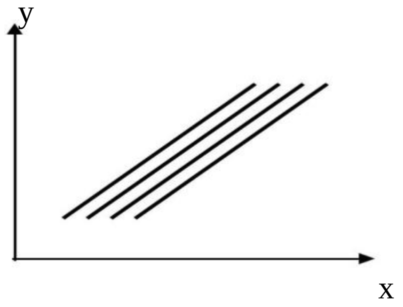
Positive correlation

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

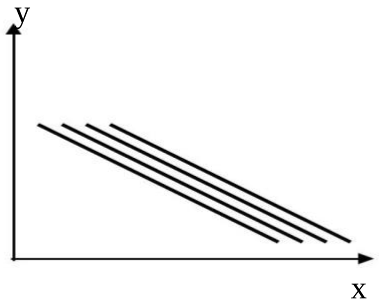
Negative correlation

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

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



Positive Correlation



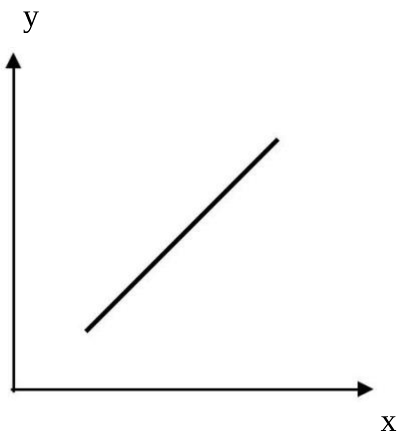
Negative Correlation

Perfect Positive Correlation

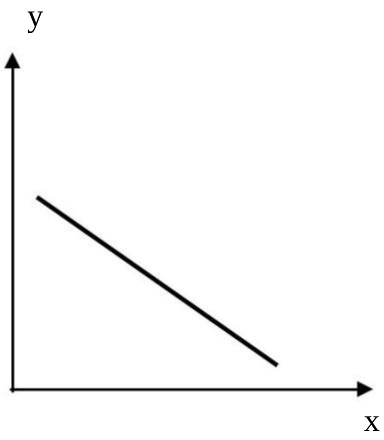
If changes in two variables are in the same direction and the changes are in equal proportion, we say that there is a perfect positive correlation between them.

Perfect Negative Correlation

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



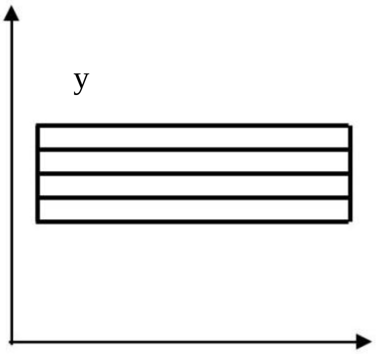
Perfect Positive Correlation



Perfect Negative Correlation

Zero Correlation

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



x

Zero Correlation

Linear Correlation

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

Coefficient of Correlation

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

Karl Pearson’s Coefficient of Simple Correlation:

Let N denote the number of pairs of observations of two variables X and Y. The correlation coefficient r between X and Y is defined by

$$r = \frac{\dfrac{\sum XY - (\sum X)(\sum Y)}{N}}{\sqrt{N\sum X2-(\sum X)^2}\;\sqrt{N\sum Y2-(\sum Y)^2}}$$

This formula is suitable for solving problems with hand calculators. To apply this formula, we have to calculate ∑ X,∑Y, ∑XY, ∑X2, ∑Y2.

Properties of Correlation Coefficient

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

The value of r ranges from – 1.0 to 0.0 or from 0.0 to 1.0

A value of r = 1.0 indicates that there exists perfect positive correlation between the two variables.

A value of r = - 1.0 indicates that there exists perfect negative correlation between the two variables.

A value r = 0.0 indicates zero correlation i.e., it shows that there is no correlation at all between the two variables.

A positive value of r shows a positive correlation between the two variables.

A negative value of r shows a negative correlation between the two variables.

A value of r = 0.9 and above indicates a very high degree of positive correlation between the two variables.

A value of - 0.9 ≥ r > - 1.0 shows a very high degree of negative correlation between the two variables.

For a reasonably high degree of positive correlation, we require r to be from 0.75 to 1.0.

A value of r from 0.6 to 0.75 may be taken as a moderate degree of positive correlation.

Problem 1

The following are data on Advertising Expenditure (in 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 the result.

Solution:

We have N = 6. Calculate ∑ X, ∑Y, ∑XY, ∑Y2, ∑Y2 as follows:

X	Y	XY	X2	Y2
18	17	306	324	289
19	17	323	361	289
20	18	360	400	324
21	19	399	441	361
22	19	418	484	361
23	19	437	529	361
Total :123	109	2243	2539	1985

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

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{}$$

$N \sum X2 - (\sum X)^2$ $\quad$ $N \sum Y2 - (\sum Y)^2$

$$6 \times 2243 - 123 \times 109$$

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

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

=51/{√105 √29} = 51/ (10.247 x 5.365)

=51/ 54.975

=0.9277

Interpretation

The value of r is 0.92. It shows that there is a high, positive correlation between the two variables ‘Advertising Expenditure’ and ‘Sales’. This provides a basis to consider some functional relationship between them.

Problem 2

Consider the following data on two variables X and Y.

X	: 12	14	18	23	24	27
Y	: 18	13	12	30	25	10

Determine the correlation coefficient between the two variables and interpret the result.

Solution:

we have N = 6. Calculate ∑ X, ∑Y, ∑XY, ∑X2, ∑Y2 as follows:

X	Y	XY	X2	Y2

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 and interpret the result.

Solution:

We have N = 8. Take X = Supply and Y = Price.

Calculate ∑ X, ∑Y, ∑XY, ∑X2, ∑Y2 as follows:

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 and interpret the result.

Solution:

We have N = 10. Take X = Income and Y = Savings.

Calculate ∑ X, ∑Y, ∑XY, ∑X2, ∑Y2 as follows:

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 = 0.9371135

Interpretation

The value of r is 0.93. The positive sign in r shows that the two variables move in the same direction. The value of r is very high. Therefore we conclude that there is high positive correlation between the two variables ‘Income’ and ‘Savings’. As a result, we can construct a functional relationship between them.

RANK CORRELATION

Spearman’s rank correlation coefficient

If ranks can be assigned to pairs of observations for two variables X and Y, then the correlation between the ranks is called the **rank correlationcoefficient**. It is usually denoted by the **symbol** ρ(rho). It is given by theformula

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

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:

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	10	7	2	6	5	4	1	9	8	3

Find out whether there is any correlation between the written and oral communication skills of the short listed candidates.

Solution:

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

RANK OF X: R_1	RANK OF Y: R_2	D=R_1- R_2	^{D}2
8	10	- 2	4
7	7	0	0
2	2	0	0
10	6	4	16
3	5	- 2	4
5	4	1	1
1	1	0	0
9	9	0	0
6	8	- 2	4
4	3	1	1

Total: 30

We have N = 10. The rank correlation coefficient is

= 1 - {6 ∑ D2 / (N3 – N)}

1 – {6 x 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	D2
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

= 1 - {6 ∑ D2 / (N3 – N)}

1 – { 6 x 82 / (1000 – 10) }

1 – (492 / 990)

1 – 0.497

= 0.503

Inference:

The rank correlation coefficient is not high.

Problem 7 (conversion of scores into ranks)

Calculate the rank correlation to determine the relationship between equity shares and preference shares given by the following data on their price.

Equity	90.0	92.4	98.5	98.3	95.4	91.3	98.0	92.0

	76.0	74.2	75.0	77.4	78.3	78.8	73.2	76.5

Solution:

From the given data on share price, we have to find out the ranks for equity shares and preference shares.

Step 1.

First, consider the equity shares and arrange them in descending order of their price as 1,2,…,8. We have the following ranks:

Equity share	98.5	98.3	98.0	95.4	92.4	92.0	91.3	90.0
Rank	1	2	3	4	5	6	7	8

Step 2.

Next, take the preference shares and arrange them in descending order of their price as 1,2,…,8. We obtain the following ranks:

Preference share	78.8	78.3	77.4	76.5	76.0	75.0	74.2	73.2
Rank	1	2	3	4	5	6	7	8

Step 3.

Calculation of D2:

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**	**Rank of Y:**	**D=R_1- R_2**	**D2**

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

= 1 - { 6 ∑ D2 / (N3 – N)}
1 – { 6 x 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	Manager I	Manager II	Manager III	$(R_1-R_2)^2$	$(R_1-R_3)^2$	$(R_2-R_3)^2$
1	8	7	4	1	16	9
2	7	8	5	1	4	9
3	6	4	1	4	25	9
4	1	6	8	25	49	4
5	5	5	9	0	16	16
6	9	10	10	1	1	0

7	10	9	6	1	16	9
8	2	3	7	1	25	16
9	3	2	3	1	0	1
10	4	1	2	9	4	1
			Total	44	156	74

We have N = 10. The rank correlation coefficient between mangers I and II is

= 1 - { 6 ∑ D2 / (N3 – N)}

1 – { 6 x 44 / (1000 – 10) }

1 – (264 / 990)

1 – 0.27

= 0.73

The rank correlation coefficient between mangers I and III is1 – { 6 x 156 / (1000 – 10) }

1 – (936 / 990)

1 – 0.95

= 0.05

The rank correlation coefficient between mangers II and III is

1 – { 6 x 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 certaincandidates 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 fractions also.

If there are three items with equal values, their ranks will be three consecutive integers, say s, s + 1 and s + 2. Their average is { s + (s+1)
(s+2) } / 3 = (3s + 3) / 3 = s + 1. Assign this rank to all 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 D2 in ρ. We have to add as many terms like (m3 – m) / 12 as there are ties.

Let us calculate the correction terms for certain values of m. These are provided in the following table.

m	m3	m3-m	Correction term $= \frac{m^3 - m}{12}$
2	8	6	0.5
3	27	24	2

4	64	60	5
5	125	120	10

Illustrative examples:

If there is a tie involving 2 items, then the correction term is 0.5

If there are 2 ties involving 2 items each, then the correction term is 0.5 + 0.5 = 1

If there are 3 ties with 2 items each, then the correction term is 0.5 + 0.5 + 0.5 = 1.5

If there is a tie involving 3 items, then the correction term is 2

If there are 2 ties involving 3 items each, then the correction term is 2 + 2 = 4

If there is a tie with 2 items and another tie with 3 items, then the correction term is 0.5 + 2 = 2.5

If there are 2 ties with 2 items each and another tie with 3 items, then the correction term is 0.5 + 0.5 + 2 = 3

Problem 9 :***Resolving ties in ranks***

The following are the details of ratings scored by two popular insurance schemes. Determine the rank correlation coefficient between them.

Scheme I	80	80	83	84	87	87	89	90
Scheme II	55	56	57	57	57	58	59	60

Solution:

From the given values, we have to determine the ranks.

Step 1.

Arrange the scores for Insurance Scheme I in descending order and rank them as 1,2,3,…,8.

Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3	4	5	6	7	8

The score 87 appears twice. The corresponding ranks are 3, 4. Their average is (3 + 4) / 2 = 3.5. Assign this rank to the two equal scores in Scheme I.

The score 80 appears twice. The corresponding ranks are 7, 8. Their average is (7 + 8) / 2 = 7.5. Assign this rank to the two equal scores in Scheme I.

The revised ranks for Insurance Scheme I are as follows:

Scheme I Score	90	89	87	87	84	83	80	80
Rank	1	2	3.5	3.5	5	6	7.5	7.5

Step 2.

Arrange the scores for Insurance Scheme II in descending order and rank them as 1,2,3,…,8.

Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	4	5	6	7	8

The score 57 appears thrice. The corresponding ranks are 4, 5, 6.

Their average is (4 + 5 + 6) / 3 = 15 / 3 = 5. Assign this rank to the three equal scores in Scheme II.

The revised ranks for Insurance Scheme II are as follows:

Scheme II Score	60	59	58	57	57	57	56	55
Rank	1	2	3	5	5	5	7	8

Step 3.

Calculation of D2: Assign the revised ranks to the given pairs ofvalues and calculate D2 as follows:

Scheme I Score	Scheme II Score	Scheme I Rank: R_1	Scheme II Rank: R_2	D=R_1- R_2	D2

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

= 1 - [{ 6 ∑ D2 + (1/2) + (1/2) +2 }/ (N3 – N)}]

1 – { 6 (4.+0.5+0.5+2) / (512 – 8) } = 1 – (6 x 7 / 504) = 1 - (42/504)

1 - 0.083 = 0.917

Inference:

It is inferred that the two insurance schemes are highly, positively correlated.

REGRESSION

In the pairs of observations, if there is a cause and effect relationship between the variables X and Y, then the average relationship between these two variables is called regression, 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 **regressionline or line of the best fit**. The regression of X on Y is different from theregression 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: $Y-Y=\bar{b}_{yx}(X-X)$

Regression of X on Y: $X-X=\bar{b}_{xy}(Y-Y)$

Where $\bar{X}, \bar{Y}$ – are the means of X, Y respectively.

Result:

Let σ_x, σ_y denote the standard deviations of x, y respectively. We have the following result.

$$b_{yx} = r\frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{xy} = r\frac{\sigma_X}{\sigma_Y}$$

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

Result:

The coefficient of correlation r between X and Y is the square root of the product of the b values in the two regression equations. We can find r by this way also.

Conclusion

This is not to say that all software problems are going to be solved by statistical means, just as not all automobile manufacturing problems can be solved by statistical means. On the contrary, the software industry has been technology driven, and the bulk of future gains in productivity will come from new, creative ideas. For example, much of the gain in productivity

Questions

1. What is statistical applications?
2. What is correlation?
3. Explain the types of correlation.
4. State the formulae of karlpearson’s correlation.
5. State the properties of correlation.
6. What is rank correlation?
7. What do you mean by regression?

Lesson - 15

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

The parent population is normal.

The error component is distributed normally with zero mean and constant variance.

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.

Classification of models

Linear models for the sample data may broadly be classified into three
types as follows:

Random effect model

Fixed effect model

Mixed effect model

In any variance components model, the error component has always random effects, since it occurs purely in a random manner. All other components may be either mixed or random.

Random effect model

A model in which each of the factors has random effect (including error effect) is called a random effect model or simply a random model.

Fixed effect model

A model in which each of the factors has fixed effects, buy only the error effect is random is called a fixed effect model or simply a fixed model.

Mixed effect model

A model in which some of the factors have fixed effects and some others have random effects is called a mixed effect model or simply a mixed model.

In what follows, we shall restrict ourselves to a fixed effect model. In a fixed effect model, the main objective is to estimate the effects and find the measure of variability among each of the factors and finally tofind the variability among the error effects.

The ANOVA technique is mainly based on the linear model which depends on the types of data used in the linear model. There are several types of data in ANOVA, depending on the number of sources of variation namely,

One-way classified data,
Two-way classified data, …
m-way classified data.

One-way classified data

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

ANOVA for One-way classified data

Let y denote the jth observation corresponding to the ith level of factor A and Y_{ij} the corresponding random variate.

Define the linear model for the sample data obtained from the experiment of the equation

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

Where µ represents the general mean effect which is fixed and which represents the general condition of the experimental units, a_i denotes the fixed effect due to ith level of the factor A (i=1,2,…,k) and hence the variation due to a_i (i=1,2,…,k) is said to be control.

The last component of the model e_{ij} is the random variable. It is called the error component and it makes the Y_{ij} a random variate. The variation in e_{ij} is due to all the uncontrolled factors and e_{ij} is independently, identically and normally distributed with mean zero and constant variance σ2 .

For the realization of the random variateY_{ij}, consider y_{ij} definedby

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

The expected value of the general observation y_{ij} in the experimental units is given by

$$E(y_{ij}) = \mu_i \text{ for all } i = 1, 2, ..., k$$

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

$$\mu_i \neq \mu \text{ for all } i = 1, 2, ..., k$$

$$i.e., \mu_i - \mu \neq 0 \text{ for all } i = 1, 2, ..., k$$

Suppose $\mu_i - \mu \neq a_i$.

Then we have $\mu_i \neq \mu + a_i$ *for all* $i = 1, 2, ..., k$

On substitution for μ_i in the above equation, the linear model reduces to

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

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 doneby minimizing the residual sum of squares defined by

$$= \sum_{ij} e^2_{ij}$$

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

Using (1). The normal equations can be obtained by partiallydifferentiating E with respect to μ and a_i *for all* $i = 1, 2, ..., k$ and equatingthe results to zero. We obtain

$$G = N\mu + \sum_i n_i a_i \qquad (2)$$

and $$T_i = n_i \mu + n_i a_i, \; i = 1,2,\dots,k \qquad (3)$$

Where N = nk. We see that the number of variables (k+1) is more than the number of independent equations (k). So, by the theorem on a system of linear equations, it follows that unique solution for this system is not possible.

However, by making the assumption that $\sum n_i a_i$ = 0 , we can get a

i

unique solution μ for and a_i (i = 1,2,…,k). Using this condition in equation (2), we get

$= N\mu$

$i.e.\mu{=}\frac{G}{N}$

μ G

Therefore the estimate of μ is given by $\mu = \overline{N}$ (4)

Again from equation (2), we have

T_i

$\overline{}$

$= \mu + a$

i

n_i

^{T}i

$\overline{}$

Hence, $a =$ $- \mu$

i

n_i

Therefore, the estimate of a_i is given by

$\underline{T}$

μ= $^{i}{-}\mu$μ

a_i $_{n}$

i

i.e.,

μ ^{T}i G

$a_i{=}$ $\overline{n}$ $-\overline{N}$ (5)

i

Substituting the least square estimates of μ and $μ_i$ in the residual sum of squares, we get

μ $

$E = (y_{ij} - \mu - ai)^2$

ij

After carrying out some calculations and using the normal equations (2) and (3) we obtain

$$E = \sum_{ij} y_{ij}^2 - \frac{G^2}{N} - \sum_{i} \frac{T_i^2}{n_i} - \frac{G^2}{N} \qquad (6)$$

The first term in the RHS of equation (6) is called the **corrected total sum of squares** while $\sum_{ij} y_{ij}^2$ is called the **uncorrected total sum ofsquares**

for measuring the variation due to treatment (controlled factor), we consider the null hypothesis that all the treatment effects are equal.

i.e.,

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

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

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

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

Under H_o , the linear model reduces to

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

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

$$E_1 = \sum_{ij} y_{ij}^2 - \frac{G^2}{N} \qquad (7)$$

Actually, E_1 contains the variation due to both treatment and error. Therefore a measure of variation due to treatment can be obtainedby " $E1 - E$ ". Using (6) and (7), we get

$$E_1 - E = \sum_{i}^{k} \frac{T_i^2}{} - \frac{G^2}{} \qquad (8)$$

$i=$

$1 \quad n_i \qquad N$

The expression in (8) is usually called the **corrected treatment sum**

$k \;\; _T2$

i

of squares while the term$\sum$ $\frac{}{}$ is called **uncorrected treatment sum**

$i=1ni$

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 asthe denoted by **α**. The testing of the null hypothesis H_o may be carried out by F test.

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 thetheoretical value or the expected value. The calculated value is referred to as the observed value.

Inference

If the observed value of F is less than the expected value of F (i.e., $F_o < F_e$) for the given level of significance **α** , then the null hypothesis H_o is accepted. In this case, we conclude that there is no significant difference between the treatment effects.

On the other hand, if the observed value of F is greater than the expected value of F (i.e.,) for the given level of significance **α** , then the null hypothesis H_o is rejected. In this case, we conclude that all the treatment effects are not equal.

Note:

If the calculated value of F and the table value of f are equal, we can try some other value of **α** .

Problem 1

The following are the details of sales effected by three sales persons in three door-to-door campaigns.

Sales person	Sales in door – to – door campaign			
A	8	9	5	10
B	7	6	6	9
C	6	6	7	5

Construct an ANOVA table and find out whether there is any significant difference in the performance of the sales persons.

Solution:

Method I (Direct method) :

$$A = 8+9+5+10 = 32$$

$$B = 7+6+6+9 = 28$$

$$C = 6+6+7+5 = 24$$

Sample mean for A : $\overline{A} = \frac{32}{4} = 8$

Sample mean for B : $\overline{B} = \frac{28}{4} = 7$

Sample mean for C : $\overline{C} = \frac{24}{4} = 6$

Total number of sample items = No. of items for A + No. of items for B + No. of items for C = 4 + 4 + 4 = 12

Mean of all the samples $\overline{X} = \frac{32 + 28 + 24}{12} = \frac{84}{12} = 7$

Sum of squares of deviations for A:

A	$A-\bar{A}=A-8$	$(A-\bar{A})^2$
8	0	0
9	1	1
5	-3	9
10	2	4
		14

Sum of squares of deviations for B:

B	$B-\bar{B}=B-7$	$(B-\bar{B})^2$
7	0	0
6	-1	1
6	-1	1
9	2	4
		6

Sum of squares of deviations for C:

C	$C-\bar{C}=C-6$	$(C-\bar{C})^2$

6	0	0
6	0	0
7	-1	1
5	-1	1
		2

Sum of squares of deviations within

Varieties = $\sum(A-A)^2+\sum(B-B)^2+\sum(\bar{C}-C)^2$ −

= 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:

$$F = \frac{\textit{Greater Variance}}{\textit{Smaller Variance}} = \frac{4.00}{2.44} = 1.6393$$

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):

∑ A = 32, ∑ B = 28, ∑ 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

$$\text{Correction Factor} = \frac{T^2}{N} = \frac{84^2}{12} = 588$$

Calculate the sum of squares of the observed values as follows:

Sales Person	X	X2
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 = $\sum X^2$– correctionfactor

= 618 – 588 = 30.

Sum of squares of deviations for variance between samples

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

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

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

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 II are one and the same. For the further steps of calculation of F value and drawing inference, refer to method I.

Problem 2

The following are the details of plinth areas of ownership apartment flats offered by 3 housing companies A,B,C. Use analysis of variance to determine whether there is any significant difference in the plinth areas of the apartment flats.

Housing Company	**Plinth area of apartment flats**			
A	1500	1430	1550	1450
B	1450	1550	1600	1480

C	1550	1420	1450	1430

Use analysis of variance to determine whether there is any significant difference in the plinth areas of the apartment’s flats.

Note:

As the given figures are large, working with them will be difficult.

Therefore, we use the following facts:

i). Variance ratio is independent of the change of origin.

ii.) Variance ratio is independent of the change of scale.

In the problem under consideration, the numbers vary from 1420 to 1600. So we follow a method called the **coding method**. First, let us subtract 1400 from each item. We get the following transformed data:

Company	Transformed measurement			
A	100	30	150	50
B	50	150	100	80
C	150	20	50	30

Next, divide each entry by 10.

Thetransformed 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

$$A=10+3+15+5=33$$

$$B=5+15+10+8=38$$

$$C=15+2+5+3=25$$

$$T=\sum A+\sum B+\sum C$$

$$33+38+25$$

$$=96$$

N = Total number of items in all the samples = 4 + 4 + 4 = 12

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

Calculate the sum of squares of the observed values as follows:

Company	X	X^2
A	10	100
A	3	9
A	15	225
A	5	25
B	5	25
B	15	225
B	10	100
B	8	64
C	15	225

C	2	4
C	5	25
C	3	9
		1036

Sum of squares of deviations for total variance = $\sum X^2$ - correctionfactor

= 1036 – 768 = 268

Sum of squares of deviations for variance between samples

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

$\frac{33^2}{4}+\frac{38^2}{4}+\frac{25^2}{4}-768$

$\frac{1089}{4}+\frac{1444}{4}+\frac{625}{4}-768$

$272.25+361+156.25-768$

$789.5-768$

$= 21.5$

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:

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

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 = 19.38

Inference:

Since the calculated value of F is less than the table value of F, the null hypothesis is accepted and it is concluded that there is no significant difference in the plinth areas of ownership apartment flats offered by the three companies, at 5% level of significance.

Problem 3

A finance manager has collected the following information on the performance of three financial schemes.

Source of variation	Degrees of freedom	Sum of squares of deviations
Treatments	5	15
Residual	2	25
Total (corrected)	7	40

Interpret the information obtained by him.

Note: ‘Treatments’ means ‘Between varieties’.

‘Residual’ means ‘Within varieties’ or ‘Error’.

Solution:

Number of schemes = 3 (since 3 – 1 = 2)

Total number of sample items = 8

Let us calculate the variance.

Variance between varieties = (since 8 – 1 = 7)

Variance between varieties = $\frac{15}{2} = 7.5$

$\frac{25}{5} = 5$

$$F = \frac{\textit{Greater Variance}}{\textit{Smaller Variance}} = \frac{7.5}{5} = 1.5$$

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 F we accept the null-hypothesis and conclude that there is no significant difference in the performance of the three financial schemes.

Conclusion

The specific test considered here is called analysis of variance (ANOVA) and is a test of hypothesis that is appropriate to compare means of a continuous variable in two or more independent comparison groups. For example, in some clinical trials there are more than two comparison groups.

Questions

1. Define ANOVA

2.What are the assumption of ANOVA?

3.State the classification of ANOVA.

4.State the variance ratio.

Lesson – 16

PARTIAL AND MULTIPLE CORRELATION

Introduction

Simple correlation does not prove to be an all-encompassing technique especially under the above circumstances. In order to get a correct picture of the relationship between two variables, we should first eliminate the influence of other variables.

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 betweenand 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

$$r12.3= \frac{r\,12-r\,13\ \ r23}{\overline{(1-r^2\,13)\,(1-r^2\,23)}}$$

Similarly,

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

and

$$r_{32.1} = \frac{r_{23} - r_{21}\, r_{13}}{\sqrt{(1 - r^2_{21})\,(1 - r^2_{13})}}$$

Problem 1

Given that $r_{12} = 0.6$, $r_{13} = 0.58$, $r_{23} = 0.70$ determine the partial correlation coefficient $r_{12.3}$

Solution:

We have

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

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

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

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

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

$= 0.3335$

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 influenceof several independent variables on the dependent variable. Multiple correlations arise in this context.

Suppose Y is a dependent variable, which is influenced by n other variables X_1, X_2, …,X_n. The multiple correlation is a measure of the relationship between Y and X_1, X_2,…,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,…, 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.

Linear Correlations:

The coefficient of multiple linear correlations R is a non-negative quantity. It varies between 0 and 1.

$^{R}1.23 \quad =$

$^{R}1.32$

$^{R}2.13 \quad =$

$^{R}2.31$

$R_{3.12} = R_{3.21}$, etc.

$R_{1.23} \geq |r_{12}|$,

$R_{1.32} \geq |r_{13}|$, 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

$$= \frac{\sqrt{r^2\,12 + r^2\,13 - 2\,r12\,r13\,r23}}{\sqrt{-\,r^2\,23}}$$

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

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

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

$$\sqrt{0.1585}$$

$$\sqrt{0.36}$$

$$\sqrt{0.4403}$$

$$0.6636$$

Problem 4

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

Solution:

We have $\boldsymbol{R}_{\mathbf{2.13}} = \sqrt{\dfrac{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}{1 - r^2_{13}}}$

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

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

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

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

$$\sqrt{0.7314}$$

$=0.8552$

Conclusion

Once the parameters responsible for performance are identified, an effective training schedule can be developed to improve the performance. The correlation coefficient gives a fair estimate of the extent of relationship between the two variables. Although the correlation coefficient may not give a clear picture of the real relationship between the two variables, yet it provides inputs for computing partial and multiple correlations. This chapter discusses the procedure for computing correlation matrix and partial correlations, and their application in research. If elements in a matrix are correlation coefficients, it is known as correlation matrix. Product moment correlation coefficient is the measure of relationship between any two variables. The chapter explains how to formulate research problems where correlation matrix and partial correlation can be used to draw conclusions.

QUESTIONS

1. Explain partial correlation.
2. Explain multiple correlations.
3. State the properties of the coefficient of multiple linear correlations.

Lesson - 17

FACTOR ANALYSIS AND CONJOINT ANALYSIS

Introduction

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales.

It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

FACTOR ANALYSIS

In a real life situation, several variables are operating. Some variables may be highly correlated among themselves. For example, if manager of a restaurant has to analyse six attributes of a new product. He undertakesa sample survey and finds out the responses of potential consumers. He obtains the following attribute correlation matrix.

Attribute

	1	2	3	4	5	6
1	1.00	0.05	0.10	0.95	0.20	0.02
2	0.05	1.00	0.15	0.10	0.60	0.85
3	0.10	0.15	1.00	0.50	0.55	0.10
4	0.95	0.10	0.50	1.00	0.12	0.08
5	0.20	0.60	0.55	0.12	1.00	0.80
6	0.02	0.85	0.10	0.08	0.80	1.00

Attribute Correlation Matrix

We try to group the attributes by their correlations. The high correlation values are observed for the following attributes:

Attributes 1, 4 with a very high correlation coefficient of 0.95.

Attributes 2, 4 with a high correlation coefficient of 0.85.

Attributes 3, 4 with a high correlation coefficient of 0.85.

As a result, it is seen that not all the attributes are independent. The attributes 1 and 4 have mutual influence on each other while the attributes 2, 5 and 6 have mutual influence among themselves. As far as attribute 3 is concerned, it has little correlation with the attributes 1, 2 and 6. Even with the other attributes 4 and 5, its correlation is not high. However, we can say that attribute 3 is somewhat closer to the variables 4 and 5 rather than the attributes 1, 2 and 6. Thus, from the given list of 6 attributes, it is possible to find out 2 or 3 common factors as follows: I.

The common features of the attributes 1,3,4 will give a factor

The common features of the attributes 2, 5, 6 will give a factor

Or

II.The common features of the attributes 1,4 will give a factor

The common features of the attributes 2,5,6 will give a factorthe attribute 3 can be considered to be an independent factor

The factor analysis is a multivariate method. It is a statistical technique to identify the underlying factors among a large number of interdependent variables. It seeks to extract common factor variances from a given set of observations. It splits a number of attributes or variables into a smaller group of uncorrelated factors. It determines which variables belong together. This method is suitable for the cases with a number of variables having a high degree of correlation.

In the above example, we would like to filter down the attributes 1, 4 into a single attribute. Also we would like to do the same for the attributes 2, 5, 6. If a set of attributes (variables) A1, A2, …,Ak filter down to an attribute A_i ($1 \leq i \leq k$), we say that these attributes are loaded on the factor A_i or saturated with the factor A_i. Sometimes, more than one factor also may be identified.

Basic concepts in factor analysis

The following are the key concepts on which factor analysis isbased.

Factor:

A factor plays a fundamental role among a set of attributes or variables. These variables can be filtered down to the factor. A factor represents the combined effect of a set of attributes. Either there may be one such factor or several such factors in a real life problem based on the complexity of the situation and the number of variables operating.

Factor loading:

A factor loading is a value that explains how closely the variables are related to the factor. It is the correlation between the factor and the variable. While interpreting a factor, the absolute value of the factor is taken into account.

Communality:

It is a measure of how much each variable is accounted for by the underlying factors together. It is the sum of the squares of the loadings of the variable on the common factors. If A,B,C,… are the factors, then the communality of a variable is computed using the relation

h2 = (The factor loading of the variable with respect to factor A)2 +

(The factor loading of the variable with respect to factor B)2 +

(The factor loading of the variable with respect to factor C)2 + …..

Eigen value:

The sum of the squared values of factor loadings pertaining to a factor is called an eigen value. It is a measure of the relative importance of each factor under consideration.

Total Sum Of Squares (TSS)

It is the sum of the eigen values of all the factors.

Application Of Factor Analysis:

1. Model Building For New Product Development:

As pointed out earlier, a real life situation is highly complex and it consists of several variables. A model for the real life situation can be built by incorporating as many features of the situation as possible. But then, with a multitude of features, it is very difficult to build such a highly idealistic model. A practical way is to identify the important variables and incorporate them in the model. Factor analysis seeks to identify those variables which are highly correlated among themselves and find a common factor which can be taken as a representative of those variables. Based on the factor loading, some of variables can be merged together to give a common factor and then a model can be built by incorporating such factors. Identification of the most common features of a product preferred by the consumers will be helpful in the development of new products.

2. Model Building For Consumers:

Another application of factor analysis is to carry out a similar exercise for the respondents instead of the variables themselves. Using the factor loading, the respondents in a research survey can be sorted out into various groups in such a way that the respondents in a group have more or less homogeneous opinions on the topics of the survey. Thus a model can be constructed on the groups of consumers. The results emanating from such an exercise will guide the management in evolving appropriate strategies towards market segmentation.

CONJOINT ANALYSIS

Introduction

Everything in the world is undergoing a change. There is a proverb saying that “the old order changes, yielding place to new”. Due to rapid advancement in science and technology, there is fast communication across the world. Consequently, the whole world has shrunk into something like a village and thus now-a-days one speaks of the “global village”. Under the present set-up, one can purchase any product of his choice from whatever part of the world it may be available. Because of this reason, what was a seller’s market a few years back has transformed into a buyer’s market now.

In a seller's market of yesterday, the manufacturer or the seller could pass on a product according to his own perceptions and prescriptions. In the buyer's market of today, a buyer decides what he should purchase, what should be the quality of the product, how much to purchase, where to purchase, when to purchase, at what cost to purchase, from whom to purchase, etc. A manager is perplexed at the way a consumer takes a decision on the purchase of a product. In this background, conjoint analysis is an effective tool to understand a buyer's preferences for a good or service.

Meaning of Conjoint Analysis

A product or service has several attributes. By an attribute, we mean a characteristic, a property, a feature, a quality, a specification oran aspect. A buyer's decision to purchase a good or service is based on 193not just one attribute but a combination of several attributes. i.e., he is concerned with a join of attributes.

Therefore, finding out the consumer's preferences for individual attributes of a product or service may not yield accurate results for a marketing research problem. In view of this fact, conjoint analysis seeks to find out the consumer's preferences for a 'join of attributes', i.e., a combination of several attributes.

Let us consider an example. Suppose a consumer desires to purchase a wrist watch. He would take into consideration several attributes of a wrist watch, namely the configuration details such as mechanism, size, dial, appearance, colour and other particulars such as strap, price, durability, warranty, after-sales service, etc. If a consumer is asked what the important aspect among the above list is, he would reply that all attributes are important for him and so a manager cannot arrive at a decision on the design of a wrist watch. Conjoint analysis assumes that the buyer will base his decision not on just the individual attributes of the product but rather he would consider various combinations of the attributes, such as

'mechanism, colour, price, after-sales service',

or 'dial, colour, durability, warranty',

or 'dial, appearance, price, durability', etc.

This analysis would enable a manager in his decision making process in the identification of some of the preferred combinations of the features of the product.

The rank correlation method seeks to assess the consumer's preferences for individual attributes. In contrast, the conjoint analysis seeks to assess the consumer's preferences for combinations (or groups) of attributes of a product or a service. This method is also called an '**unfolding technique**' because preferences on groups of attributes unfold from the rankings expressed by the consumers. Another name for this method is '**multi-attribute compositional model**' because it deals with combinations of attributes.

STEPS IN THE DEVELOPMENT OF CONJOINT ANALYSIS

The development of conjoint analysis comprises of the followingsteps:

Collect a list of the attributes (features) of a product or a service.

For each attribute, fix a certain number of points or marks. The more the number of points for an attribute, the more serious the consumers' concern on that attribute.

Select a list of combinations of various attributes.

Decide a mode of presentation of the attributes to the respondents of the study i.e., whether it should be in written form, or oral form, or a pictorial representation etc.

Inform the combinations of the attributes to the prospective customers.Request the respondents to rank the combinations, or to rate them on a suitable scale, or to choose between two different combinations at a time.Decide a procedure to aggregate the responses from the consumers.

Any one of the following procedures may be adopted:

Go by the individual responses of the consumers.

Put all the responses together and construct a single utilityFunction.

Split the responses into a certain number of segments such that within each segment, the preferences would be similar.

Choose the appropriate technique to analyze the data collected from the respondents.

Identify the most preferred combination of attributes.

Incorporate the result in designing a new product, construction of an advertisement copy, etc.

Applications Of Conjoint Analysis

An idea of consumer's preferences for combinations of attributes will be useful in designing new products or modification of an existing product.

A forecast of the profits to be earned by a product or a service.

A forecast of the market share for the company's product.

A forecast of the shift in brand loyalty of the consumers.

A forecast of differences in responses of various segments of the product.

Formulation of marketing strategies for the promotion of the product.

Evaluation of the impact of alternative advertising strategies.

A forecast of the consumers' reaction to pricing policies.

A forecast of the consumers' reaction on the channels of distribution.

Evolving an appropriate marketing mix.

Even though the technique of conjoint analysis was developed for the formulation of corporate strategy, this method can be used to have a comprehensive knowledge of a wide range of areas such as family decision making process, pharmaceuticals, tourism development, public transport system, etc.

Advantages of Conjoint Analysis

- The analysis can be carried out on physical variables.
- Preferences by different individuals can be measured and pooled together to arrive at a decision.

Disadvantages of Conjoint Analysis

- When more and more attributes of a product are included in the study, the number of combinations of attributes also increases, rendering the study highly difficult. Consequently, only a few selected attributes can be included in the study.
- Gathering of information from the respondents will be a tough job.
- Whenever novel combinations of attributes are included, the respondents will have difficulty in capturing such combinations.
- The psychological measurements of the respondents may not be accurate. In spite of the above stated disadvantages, conjoint analysis offers more scope to the researchers in identifying the consumers’ preferences for groups of attributes.

Illustrative Problem 1 : Application Of Rating Scale Technique

A wrist watch manufacturer desires to find out the combinations of attributes that a consumer would be interested in. After considering several attributes, the manufacturer identifies the following combinations of attributes for carrying out marketing research.

Combination – I	Mechanism, colour, price, after-scales service
Combination – II	Dial, colour, durability, warranty
Combination – III	Dial, appearance, price, durability
Combination – IV	Mechanism, dial, price, warranty

12 respondents are asked to rate the 4 combinations on the following3-point rating scale.

Scale – 1	:	Less important
Scale – 2	:	Somewhat important
Scale – 3	:	Very important

Their responses are given in the following table:

Rating of Combination

Respondent No.	**Combination I**	**Combination II**	**Combination III**	**Combination IV**
1	Less important	Some what important	Very important	Some what important
2	Some what important	Very important	Less important	Some what important
3	Some what important	Less important	Some what important	Very important
4	Less important	Less important	Very important	Some what important
5	Some what important	Very important	Very important	Less important
6	Some what	Very	Some what	Less

	important	important	important	important
7	Some what important	Less important	Very important	Less important
8	Very important	Some what important	Less important	Some what important
9	Very important	Less important	Some what important	Some what important
10	Some what important	Very important	Less important	Some what important
11	Very important	Some what important	Very important	Some what important

12	Very important	Less important	Very important	Some what important

Determine the most important and the least important combinations of the attributes.

Solution:

Let us assign scores to the scales as follows:

Sl. No.	**Scale**	**Score**
1	Less important	1
2	Some what important	3
3	Very important	5

The scores for the four combinations are calculated as follows:

Combination	**Response**	**Score for Response**	**No. Of Respondents**	**Total score**
I	Less important Some what important	1 3 5	2 6 4	1X2=2 3X6=18 5X4=20
			12	40
II	Less important Some what important	1 3 5	5 3 4	1X5=5 3X3=9 5X4=20
			12	34

III	Less important Some what important	1 3 5	3 3 6	1 X 3 = 3 3 X 3 = 9 5 X 6 = 30
			12	42

Let us tabulate the scores earned by the four combinations as follows:

Combination	Total scores
I	40
II	34
III	42
IV	32

Inference:

It is concluded that the consumers consider combination III as the most important and combination IV as the least important.

Note:

For illustrating the concepts involved, we have taken up 12 respondents in the above problem. In actual research work, we should take a large number of respondents, say 200 or 100. In any case, the number of respondents shall not be less than 30.

Illustrative Problem 2: Application Of Ranking Method

A marketing manager selects four combinations of features of a product for study. The following are the ranks awarded by 10 respondents. Rank one means the most important and rank 4 means the least important.

Respondent No.	Rank Awarded			
	Combination I	Combination II	Combination III	Combination IV
1	2	1	3	4
2	1	4	2	3
3	1	2	3	4
4	3	2	4	1
5	4	1	2	3
6	1	2	3	4
7	4	3	2	1
8	3	1	2	4
9	3	1	4	2
10	4	1	2	3

Determine the most important and the least important combinations of the features of the product.

Solution:

Let us assign scores to the ranks as follows:

Rank	Score
1	10
2	8
3	6
4	4

The scores for the 4 combinations are calculated as follows:

Combination	Rank	Score for rank	No. of	Total score
I	1 2 3 4	10 8 6 4	3 1 3 3	10 X 3 = 30 8 X 1= 8 6 X 3 = 18 4 X 3 = 12
			10	68
II	1 2 3 4	10 8 6 4	5 3 1 1	10 X 5 = 50 8 X 3 = 24 6 X 1 = 6 4 X 1 = 4

			10	84
	1 2 3 4	10 8 6 4	Nil 5 3 2	-- 8 X 5 = 40 6 X 3 = 18 4 X 2 = 8
			10	66
	1 2 3 4	10 8 6 4	2 1 3 4	10 X 2 = 20 8 X 1 = 8 6 X 3 = 18 4 X 4 = 16
			10	62

The final scores for the 4 combinations are as follows:

Combination	Score
I	68
II	84
III	66
IV	62

Inference:

It is seen that combination II is the most preferred one by the consumers and combination IV is the least preferred one.

Illustrative Problem 3:

Application Of Mini-Max Scaling Method

An insurance manager chooses 5 combinations of attributes of a social security plan for analysis. He requests 10 respondents to indicate their perceptions on the importance of the combinations by awardingthe minimum score and the maximum score for each combination in the range of 0 to 100. The details of the responses are given below. Help the manager in the identification of the most important and the least important combinations of the attributes of the social security plan.

Respondent No.	Combination I		Combination II		Combination III		Combination IV		Combination V	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1	30	60	45	85	50	70	40	75	50	80
2	35	65	50	80	50	80	35	75	40	75
3	40	70	35	80	60	80	40	70	50	80
4	40	80	40	80	60	85	50	75	60	80
5	30	75	50	80	60	75	60	75	60	85
6	35	70	35	85	50	80	40	80	40	80
7	40	80	40	75	45	75	50	70	40	80
8	30	80	40	75	50	80	50	70	60	80
9	45	75	45	75	50	80	50	80	50	80
10	55	75	40	85	35	75	45	80	40	80

Solution:

For each combination, consider the minimum score and the maximum score separately and calculate the average in each case.

	Com-bination I		Com-bination II		Com-bination III		Com-bination IV		Com-bination V	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Total	380	730	420	800	510	780	460	750	490	800
Average	38	73	42	80	51	78	46	75	49	80

Consider the mean values obtained for the minimum and maximum of each combination and calculate the range for each combination as

Range = Maximum Value – Minimum Value

The measure of importance for each combination is calculated as follows:

Measure of importance for a combination of attributes

$$= \frac{\text{Range for that combination}}{\text{Sum of the ranges for all the combinations}} \times 100$$

Tabulate the results as follows:

Combination	Max. Value	Min. Value	Range	Measure of Importance
I	73	38	35	21.875
II	80	42	38	23.750
III	78	51	27	16.875
IV	75	46	29	18.125
V	80	49	31	19.375
Sum of the ranges			160	100.000

Inference:

It is concluded that combination II is the most important one and combination III is the least important one.

APPROACHES FOR CONJOINT ANALYSIS

The following two approaches are available for conjoint analysis:

Multi-factor evaluation approach

Two-factor evaluation approach

MULTI-FACTOR EVALUATION APPROACH IN CONJOINT ANALYSIS

Suppose a researcher has to analyze n factors. It is possible that each factor can assume a value in different levels.

Product Profile

A product profile is a description of all the factors under consideration, with any one level for each factor.

Suppose, for example, there are 3 factors with the levels given below.

Factor 1	:	3 levels
Factor 2	:	2 levels
Factor 3	:	4 levels

Then we have 3 × 2 × 4 = 24 product profiles. For each respondent in the research survey, we have to provide 24 data sheets such that each data sheet contains a distinct profile. In each profile, the respondent is requested to indicate his preference for that profile in a rating scale of 0 to 10. A rating of 10 indicates that the respondent's preference for that profile is the highest and a rating of 0 means that he is not all interested in the product with that profile.

Example:

Consider the product 'Refrigerator' with the following factors andlevels:

Factor 1	:	Capacity of 180 liters; 200 liters; 230 liters
Factor 2	:	Number of doors: either 1 or 2
Factor 3	:	Price :rs. 9000; rs. 10,000; rs. 12,000

Sample profile of the product

Profile Number	:	
Capacity	:	200 liters
Number of Doors	:	1

Price : Rs. 10,000

Rating of Respondent:

(in the scale of 0 to 10)

Steps In Multi-Factor Evaluation Approach:

- ❖ Identify the factors or features of a product to be analyzed. If they are too many, select the important ones by discussion with experts.
- ❖ Find out the levels for each factor selected in step 1.
- ❖ Design all possible product profiles. If there are n factors with levels L_1, L_2,…L_n respectively, then the total number of profiles = $L_1L_2…L_n$.
- ❖ Select the scaling technique to be adopted for multi-factor evaluation approach (rating scale or ranking method).
- ❖ Select the list of respondents using the standard sampling technique.
- ❖ Request each respondent to give his rating scale for all the profiles of the product. Another way of collecting the responses is to request each respondent to award ranks to all the profiles: i.e., rank 1 for the best profile, rank 2 for the next best profile etc.
- ❖ For each factor profile, collect all the responses from all the participating respondents in the survey work.
- ❖ With the rating scale awarded by the respondents, find out the score secured by each profile.
- ❖ Tabulate the results in step 8. Select the profile with the highest score. This is the most preferred profile. Implement the most preferred profile in the design of a new product.

Two-Factor Evaluation Approach In Conjoint Analysis

When several factors with different levels for each factor have to be analyzed, the respondents will have difficulty in evaluating all the profiles in the multi-factor evaluation approach. Because of this reason, two-factor evaluation approach is widely used in conjoint analysis.

Suppose there are several factors to be analyzed with different levels of values for each factor, then we consider any two factors at a time with their levels of values. For each such case, we have a data sheet called a **two-factor table**. If there are n factors, then the number of such data sheets is .

$$\binom{n}{2}=\frac{n(n-1)}{2}$$

Let us consider the example of ‘Refrigerator’ described in the multi-factor approach. For the two factors (i) capacity and (ii) price, we have the following **data sheet.**

Data Sheet (Two Factor Table) No:

Steps in two-factor evaluation approach:

- ❖ Identify the factors or features of a product to be analyzed.
- ❖ Find out the levels for each factor selected in step 1.
- ❖ Consider all possible pairs of factors. If there are n factors, then the number of pairs is $\binom{n}{2}=\frac{n(n-1)}{2}$. For each pair of factors, prepare two-factor table, indicating all the levels for the two factors. If L_1 and L_2 are the respective levels for the two factors, then the number of cells in the corresponding table is L_1L2.

Select the list of respondents using the standard sampling technique.

Request each respondent to award ranks for the cells in each two-factor table. I.e., rank 1 for the best cell, rank 2 for the next best cell, etc.

For each two-factor table, collect all the responses from all the participating respondents in the survey work.

With the ranks awarded by the respondents, find out the score secured by each cell in each two-factor table.

Tabulate the results in step 7. Select the cell with the highest score. Identify the two factors and their corresponding levels.

Implement the most preferred combination of the factors and their levels in the design of a new product.

Application:

The two factor approach is useful when a manager goes for market segmentation to promote his product. The approach will enable the top level management to evolve a policy decision as to which segment of the market has to be concentrated more in order to maximize the profit from the product under consideration.

QUESTIONS

1. Explain the purpose of ‘factor analysis’.
2. What is the objective of ‘conjoint analysis’? Explain.
3. State the steps in the development of conjoint analysis.
4. State the applications of conjoint analysis.
5. What are the application used in the factor analysis?
6. What is conjoint?
7. What are Advantages and disadvantages of conjoint analysis?
8. Explain the approaches of conjoint analysis.
9. What is multifactor evaluation approach?
10. State the two factor evaluation approach.
11. Explain the steps in two factor analysis.

UNIT - V

Lesson – 19

RESEARCH REPORTS

Introduction

A report is a written document on a particular topic, which conveys information and ideas and may also make recommendations. Reports often form the basis of crucial decision making.

WHAT IS A REPORT?

Inaccurate, incomplete and poorly written reports fail to achieve their purpose and reflect on the decision, which will ultimately be made. This will also be the case if the report is excessively long, jargonistic and/ or structure less. A good report can be written by keeping the following features in mind:

All points in the report should be clear to the intended reader.

- The report should be concise with information kept to a necessary minimum and arranged logically under various headings and sub-headings.
- All information should be correct and supported by evidence.
- All relevant material should be included in a complete report.

Purpose Of Research Report

- Why am i writing this report? Do i want to inform/ explain/ persuade, or indeed all of these.
- Who is going to read this report? Managers/ academicians/ researchers! What do they already know? What do they need to know? Do any of them have certain attitudes or prejudices?
- What resources do we have? Do i have access to a computer? Do i have enough time? Can any of my colleagues help?
- Think about the content of your report – what am i going to put in it? What are my main themes? How much should be the text, and how much should be the illustrations?

Framework of a Report

The various frameworks can be used depending on the content of the report, but generally the same rules apply. Introduction, method, results and discussion with references or bibliography at the end, and an abstract at the beginning could form the framework.

STRUCTURE OF A REPORT

Structure your writing around the IMR&D framework and you will ensure a beginning, middle and end to your report.

I	Introduction	Why did i do this research?	(beginning)
M	Method	What did i do and how did i go about doing it?	(middle)
R	Results	What did i find?	(middle)
AND			
D	Discussion	What does it all mean?	(end)

What do I put in the beginning part?

TITLE PAGE	Title of project, Sub–title (where appropriate), Date, Author, Organization, Logo
BACKGROUND	History(if any) behind project
ACKNOWLEDGEMENT	Author thanks people and organization who helped during the project
SUMMARY(sometimes called abstract of the synopsis)	A condensed version of a report – outlines salient points, emphasizes main conclusions and (where appropriate) the main recommendations. N.B this is often difficult to write and it is suggested that you write it last.
	An at- a – glance list that tells the reader what is in the report and what page number(s) to find it on.

	As above, specifically for tables.
	As above, specifically for appendices.
	Author sets the scene and states his/ her intentions.
	AIMS – general aims of the audit/ project, broad statement of intent. OBJECTIVES – specific things expected to do/ deliver(e.g. expected outcomes)

What do I Put In the middle Part?

METHOD	Work steps; what was done – how, by whom, when?
RESULT/FINDINGS	Honest presentation of the findings, whether these were as expected or not. Give the facts, including any inconsistencies or difficulties encountered

What do I put in the end part?

Conclusion	Explanation of the results.(you might like to keep the SWOT analysis in mind and think about your project's strengths, weakness, opportunities and threats, as you write)
Recommendation	The author links the results/ findings with the points made in the introduction and strives to reach

	clear, simply stated and unbiased conclusions. Make sure they are fully supported by evidence and arguments of the main body of your audit/project.
	The author states what specific actions should be taken, by whom and why. They must always be linked to the future and should always be realistic. Don’t make them unless asked to.
REFERENCES	A section of a report, which provides full details of publications mentioned in the text, or from which extracts have been quoted.

APPENDIX	The purpose of an appendix is to supplement the information contained in the main body of the report.

PRACTICAL REPORTS VS. ACADEMIC REPORTS

Practical Reports:

In the practical world of business or government, a report conveys an information and (sometimes) recommendations from a researcher who has investigated a topic in detail. A report like this will usually be requested by people who need the information for a specific purpose and their request may be written in terms of reference or the brief. Whatever the report, it is important to look at the instruction for what is wanted. A report like this differs from an essay in thatit is designed to provide information which will be acted on, rather than to be read by people interested in the ideas for their own sake.Because of this, it has a different structure and layout.

Academic Reports:

A report written for an academic course can be thought of as a simulation. We can imagine that someone wants the report for a practical purpose, although we are really writing the report as an academic exercise for assessment. Theoretical ideas will be more to the front in an academic report than in a practical one. Sometimes a report seems to serve academic and practical purposes. Students on placement with organizations often have to produce a report for the organization and for assessment on the course. Although the background work for both will be related, in practice, the report the student produces for academic assessment will be different from the report produced for the organization, because the needs of each are different.

RESEARCH REPORT: PRELIMINARIES

It is not sensible to leave all your writing until the end. There is always the possibility that it will take much longer than you anticipate and you will not have enough time. There could also be pressure upon available word processors as other students try to complete their own reports. It is wise to begin writing up some aspects of your research as you go along. Remember that you do not have to write your report in the order it will be read. Often it is easiest to start with the method section. Leave the introduction and the abstract to last. The use of a word processor makes it very straightforward to modify and rearrange what you have written as your research progresses and your ideas change. The very process of writing will help your ideas to develop. Last but by no means least, ask someone to proofread your work.

STRUCTURE OF A RESEARCH REPORT

A research report has a different structure and layout in comparison to a project report. A research report is for reference and is often quite a long document. It has to be clearly structured for the readers to quickly find the information wanted. It needs to be planned carefully to make sure that the information given in the report is put under correct headings.

PARTS OF RESEARCH REPORT

Cover sheet: This should contain some or all of the following:

Full title of the report
Name of the researcher
Name of the unit of which the project is a part
Name of the institution Date/year.

Title page: Full title of the report.

Your name

Acknowledgement: a thanks giving to the people who helped you.

Contents

List of the Tables

Headings and sub-headings used in the report should be given with heir page numbers. Each chapter should begin on a new page. Use a consistent system in dividing the report into parts. The simplest may be to use chapters for each major part and subdivide these into sections and sub-sections. 1, 2, 3 etc.Can be used as the numbers for each chapter. The sections of chapter 3 (for example) would be 3.1, 3.2, 3.3, and so on. For further sub-division of a sub-section you may use 3.2.1, 3.2.2, and so on.

Abstract or **Summary** or **Executive Summary** or **Introduction:**

This presents an overview of the whole report. It should let the reader see in advance, what is in the report. This includes what you set out to do, how review of literature is focused and narrowed in your research, the relation of the methodology you chose to your objectives, a summary of your findings and analysis of the findings

BODY

Aims And Purpose or **Aims And Objectives:**

Why did you do this work? What was the problem you were investigating? If you are not including review of literature, mention the specific research/es which is/are relevant to your work.

Review of Literature

This should help to put your research into a background context and to explain its importance. Include only the books and articles which relate directly to your topic. You need to be analytical and critical, and not just describe the works that you have read.

Methodology

Methodology deals with the methods and principles used in an activity, in this case research. In the methodology chapter, explain the method/syou used for the research and why you thought they were the appropriate ones. You may, for example, be depending mostly upon secondary data or you might have collected your own data. You should explain the method of data collection, materials used, subjects interviewed, or places you visited. Give a detailed account of how and when you carried out your research and explain why you used the particular method/s, rather than other methods. Included in this chapter should be an examination of ethical issues, if any.

Results or Findings

What did you find out? Give a clear presentation of your results. Show the essential data and calculations here. You may use tables, graphs and figures.

Analysis and Discussion

Interpret your results. What do you make out of them? How do they compare with those of others who have done research in this area? The accuracy of your measurements/results should be discussed and deficiencies, if any, in the research design should be mentioned.

Conclusions

What do you conclude? Summarize briefly the main conclusions which you discussed under “Results.” were you able to answer some or all of the questions which you raised in your aims and objectives? Do not be tempted to draw conclusions which are not backed up by your evidence. Note the deviation/s from expected results and any failure to achieve all that you had hoped.

Recommendations

Make your recommendations, if required. The suggestions for action and further research should be given.

Appendix

You may not need an appendix, or you may need several. If you have used questionnaires, it is usual to include a blank copy in the appendix. You could include data or calculations, not given in the body, that arenecessary, or useful, to get the full benefit from your report. There may be maps, drawings, photographs or plans that you want to include. If you have used special equipment, you may include information about it.

The plural of an **appendix** is appendicesare needed, design them**Referencesappendices**. If an appendix orthoughtfully in a way that yourreaders find it/them convenient to use.List all the sources which you referred in the body of the report. You may use the pattern prescribed by American Psychological Association, or any other standard pattern recognized internationally.

REVIEW OF LITERATURE

In the case of small projects, this may not be in the form of a critical review of the literature, but this is often asked for and is a standard part of larger projects. Sometimes students are asked to write Review of Literature on a topic as a piece of work in its own right. In its simplest form, the review of literature is a list of relevant books and other sources, each followed by a description and comment on its relevance.

The literature review should demonstrate that you have read and analysed the literature relevant to your topic. From your readings, you may get ideas about methods of data collection and analysis.

If the review is part of a project, you will be required to relate your readings to the issues in the project, and while describing the readings, you should apply them to your topic. A review should include only relevant studies. The review should provide the reader with a picture of the state of knowledge in the subject.

Your literature search should establish what previous researches have been carried out in the subject area. Broadly speaking, there are three kinds of sources that you should consult:

Introductory material;
Journal articles and
Books.

To get an idea about the background of your topic, you may consult one or more textbooks at the appropriate time. It is a good practice to review in cumulative stages - that is, do not think you can do it all at one go. Keep a careful record of what you have searched, how you have gone about it, and the exact citations and page numbers of your readings. Write notes as you go along. Record suitable notes on everything you read and note the methods of investigations. Make sure that you keep a full reference, complete with page numbers. You will have to find your own balance between taking notes that are too long and detailed, and ones too brief to be of any use. It is best to write your notes in complete sentences and paragraphs, because research has shown that you are more likely to understand your notes later if they are written in a way that other people would understand. Keep your notes from different sources and/or about different points on separate index cards or on separate sheets of paper. You will do mainly basic reading while you are trying to decide on your topic. You may scan and make notes on the abstracts or summaries of work in the area. Then do a more thorough job of reading later on, when you are more confident of what you are doing. If your project spans several months, it would be advisable towards the end to check whether there are any new and recent references.

REFERENCES

There are many methods of referencing your work; some of the most common ones are the numbered style, american psychological association style and the harvard method, with many other variations. Just use the one you are most familiar and comfortable with. Details of all the works referred by you should be given in the reference section.

THE PRESENTATION OF REPORT

Well-produced, appropriate illustrations enhance the presentability of a report. With today’s computer packages, almost anything is possible. However, histograms, bar charts and pie charts are still the three ‘staples’. Readers like illustrated information, because it is easier to absorb and it’s more memorable. Illustrations are useful only when they are easier to understand than words or figures and they must be relevant to the text. Use the *algorithm* included to help you decide whether or not to use an illustration. They should never be included for their own sake, and don’t overdo it; too many illustrations distract the attention of readers.

Conclusion

The conclusion should provide the reader with a sense of finality. The reader should feel that writer has made the point they wish to make, and supported their thesis by effectively arguing their case. However, the conclusion should pique the reader’s curiosity, and instill in the reader a desire to learn more about ideas, issues, and questions that are raised by the report.

Questions

1. What do you mean by report?
2. What is the purpose of research report?
3. State the structure of research report.
4. Distinguish between the practical Vs academic report.
5. Explain the parts of research report.

Lesson – 20

TYPES OF REPORTS

Introduction

Reports vary in length and type. Students' study reports are often called term papers, project reports, theses, dissertations depending on the nature of the report. Reports of researchers are in the form of monographs, research papers, research thesis, etc. In business organizations a wide variety of reports are under use: project reports, annual reports of financial statements, report of consulting groups, project proposals etc. News items in daily papers are also one form of report writing. In this lesson, let us identify different forms of reports and their major components.

Types of reports

Reports may be categorized broadly as Technical Reports and General Reports based on the nature of methods, terms of reference and the extent of in-depth enquiry made etc. On the basis of usage pattern, the reports may also be classified as information oriented reports, decision oriented reports and research based reports. Further, reports may also differ based on the communication situation. For example, the reports may be in the form of Memo, which is appropriate for informal situations or for short periods. On the other hand, the projects that extend over a period of time, often call for project reports. Thus, there is no standard format of reports. The most important thing that helps in classifying the reports is the outline of its purpose and answers for the following questions:

- What did you do?
- Why did you choose the particular research method that you used?
- What did you learn and what are the implications of what you learned?
- If you are writing a recommendation report, what action are you recommending in response to what you learned?

Two types of report formats are described below:

A Technical Report

A technical report mainly focuses on methods employed, assumptions made while conducting a study, detailed presentation of findings and drawing inferences and comparisons with earlier findings based on the type of data drawn from the empirical work.

An outline of a Technical Report mostly consists of the following:

Title and nature of the study:

Brief title and the nature of work sometimes followed by subtitle indicate more appropriately either the method or tools used. Description of objectives of the study, research design, operational terms, working hypothesis, type of analysis and data required should be present.

Abstract of Findings:

A brief review of the main findings just can be made either in a paragraph or in one/two pages.

Review of current status:

A quick review of past observations and contradictions reported, applications observed and reported are reviewed based on the in-house resources or based on published observations.

Sampling and Methods employed

Specific methods used in the study and their limitations. In the case of experimental methods, the nature of subjects and control conditions are to be specified. In the case of sample studies, details of the sample design i.e., sample size, sample selection etc are given.

Data sources and experiment conducted

Sources of data, their characteristics and limitations should be specified. In the case of primary survey, the manner in which data has been collected should be described.

Analysis of data and tools used.

The analysis of data and presentation of findings of the study with supporting data in the form of tables and charts are to be narrated. This constitutes the major component of the research report.

Summary of findings

A detailed summary of findings of the study and major observations should be stated. Decision inputs if any, policy implications from theobservations should be specified.

References

A brief list of studies conducted on similar lines, either preceding the present study or conducted under different experimental conditions is listed.

Technical appendices

These appendices include the design of experiments or questionnaires used in conducting the study, mathematical derivations, elaboration on particular techniques of analysis etc.

General Reports

General reports often relate popular policy issues mostly related to social issues. These reports are generally simple, less technical, good use of tables and charts. Most often they reflect the journalistic style. Example for this type of report is the “Best B-Schools Survey in Business Magazines”. The outline of these reports is as follows:

Major Findings and their Implications

Recommendations for Action

Objectives of the Study

Method Employed for Collecting Data

Results

Writing Styles

There are atleast 3 distinct report writing styles that can be applied by students of Business Studies. They are called:

Conservative

Key points

Holistic

1. Conservative Style

Essentially, the conservative approach takes the best structural elements from essay writing and integrates these with appropriate report writing tools. Thus, headings are used to deliberate upon different sections of the answer. In addition, the space is well utilized by ensuring that each paragraph is distinct (perhaps separated from other paragraphs by leaving two blank lines in between).

2. Key Point Style

This style utilizes all of the report writing tools and is thus more overtly 'report-looking'. Use of headings, underlining, margins, diagrams and tables are common. Occasionally reporting might even use indentation and dot points. The important thing to remember is that the tools should be applied in a way that adds to the report. The question must be addressed and the tools applied should assist in doing that. An advantage of this style is the enormous amount of information that can be delivered relatively quickly.

3. Holistic Style

The most complex and unusual of the styles, holistic report writing aims to answer the question from a thematic and integrative perspective. This style of report writing requires the researcher to have a strong understanding of the course and is able to see which outcomes are being targeted by the question.

Essentials of A Good Report:

Good research report should satisfy some of the following basic characteristics:

STYLE

Reports should be easy to read and understand. The style of the writer should ensure that sentences are succinct and the language used is simple, to the point and avoiding excessive jargon.

LAYOUT

A good layout enables the reader to follow the report's intentions, and aids the communication process. Sections and paragraphs should be given headings and sub¬-headings. You may also consider a system of numbering or lettering to identify the relative importance of paragraphs and sub-paragraphs. Bullet points are an option for highlighting important points in your report.

ACCURACY

Make sure everything you write is factually accurate. If you would mislead or misinform, you will be doing a disservice not only to yourself but also to the readers, and your credibility will be destroyed. Remember to refer to any information you have used to support your work.

CLARITY

Take a break from writing. When you would come back to it, you'll have the degree of objectivity that you need. Use simple language to express your point of view.

READABILITY

Experts agree that the factors, which affect readability the most, are:

Attractive appearance
Non-technical subject matter
Clear and direct style
Short sentences
Short and familiar words

REVISION

When first draft of the report is completed, it should be put to one side atleast for 24 hours. The report should then be read as if with eyes of the intended reader. It should be checked for spelling and grammatical errors. Remember the spell and grammar check on your computer. Use it!

REINFORCEMENT

Reinforcement usually gets the message across. This old adage is well known and is used to good effect in all sorts of circumstances e.g., presentations - not just report writing.

TELL THEM WHAT YOU ARE GOING TO SAY: in the introduction and summary you set the scene for what follows in your report.

THEN SAY IT : you spell things out in results/findings
THEN TELL THEM WHAT YOU SAID: you remind your readers through the discussion what it was all about.

FEEDBACK MEETING

It is useful to circulate copies of your report prior to the feedback meeting. Meaningful discussion can then take place during the feedback meeting with recommendations for change more likely to be agreed upon which can then be included in your conclusion. The following questions should be asked at this stage to check whether the Report served the purpose:

Does the report have impact?
Do the summary /abstract do justice to the report?
Does the introduction encourage the reader to read more?
Is the content consistent with the purpose of the report?
Have the objectives been met?
Is the structure logical and clear?
Have the conclusions been clearly stated?
Are the recommendations based on the conclusions and expressed clearly and logically?

Different Parts of a Report

- ✓ Generally different parts of a report include:
- ✓ Cover Page / Title Page
- ✓ Introductory Pages (Foreword, Preface, Acknowledgement, Table of ontents, List of Tables, List of Illustrations or Figures, Key Words / Abbreviations Used Etc.)
- ✓ Contents of the Report (Which Generally Includes a Macro Setting, Research Problem, Methodology Used, Objectives of the Study, Review of Studies, Tools Used for Data Collection and Analysis, Empirical Results in One/Two Sections, Summary of Observations etc.)
- ✓ References (Including Appendices, Glossary of Terms Used, Source Data, Derivations of Formulas for Models Used in the Analysis etc.)

Title Page:

The Cover Page or Title Page of a Research Report should contain the following information:

Title of the Project / Subject

Who has conducted the study

For what purpose

Organization

Period of submission

A Model:

An example of a Summer Project Report conducted by an MBA student generally follows the following Title Page:

A STUDY ON THE USE OF COMPUTER TECHNOLOGY IN BANKING OPERATIONS IN XXX BANK LTD., PONDICHERRY

A SUMMER PROJECT REPORT

PREPARED BY

Ms. MADAVI LATHA

Submitted at

SCHOOL OF MANAGEMENT

PONDICHERRY UNIVERSITY

PONDICHERRY – 605 014

2017

Introductory Pages:

Introductory pages generally do not constitute the Write up of the Research work done. These introductory pages basically form the Index of the work done. These pages are usually numbered in Roman numerical(eg, I, ii, iii etc). The introductory pages include the following components

- Foreword
- Preface
- Acknowledgements
- Table of Contents
- List of Tables
- List of Figures / Charts

Foreword is usually one page write up or a citation about thework by any eminent / popular personality or a specialist in the given field of study. Generally, the write up includes a brief background on the contemporary issues and suitability of the present subject and its timeliness, major highlights of the present work, brief background of the author etc. The writer of the foreword generally gives the foreword on his letter head**Preface** is again one/two pages write up by the author of the bookreport stating circumstances under which the present work is taken up, importance of the work, major dimensions examined and intended audience for the given work. The author gives his signature and address at the bottom of the page along with date and year of the work

Acknowledgements is a short section, mostly a paragraph. Itmostly consists of sentences giving thanks to all those associated and encouraged to carry out the present work. Generally, author takes time to acknowledge the liberal funding by any funding agency to carry out the work, and agencies which had given permission to use their resources. At the end, the author thanks everybody and gives his signature.

Table of Contents refers to the index of all pages of the said ResearchReport. These contents provide the information about the chapters, sub-sections, annexure for each chapter, if any, etc. Further, the page numbers of the content of the report greatly helps any one to refer to those pages for necessary details. Most authors use different forms while listing the sub contents. These include alphabet classification and decimal classification.

Examples for both of them are given below:

Example of content sheet (alphabet classification)

An example of Content Sheet with decimal classification

CONTENTS

List of Tables and Charts:

Details of Charts and Tables given in the Research Report are numbered and presented on separate pages and the lists of such tables and charts are given on a separate page. Tables are generally numbered either in Arabic numerals or in decimal form. In the case of decimal form, it is possible to indicate the chapter to which the said table belongs. For example, Table 2.1 refers to Table 1 in Chapter 2.

Executive Summary:

Most Business Reports or Project works conducted on a specific issue carry one or two pages of Executive Summary. This summary precedes the chapters of the Regular Research Report. This summary generally contains a brief description of problem under enquiry, methods used and the findings. A line about the possible alternatives for decision making would be the last line of the Executive Summary.

BODY OF THE REPORT:

The body of the Report is the most important part of the report. This body of report may be segmented into a handful of Units or Chaptersarranged in a sequential order. Research Report often present the Methodology, Objectives of the study, data tools, etc in the first or second chapters along with a brief background of the study, review of relevant studies. The major findings of the study are incorporated into two or three chapters based on the major or minor hypothesis tested or based on the sequence of objectives of the study. Further, the chapter plan may also be based likely on different dimensions of the problem under enquiry.

Each Chapter may be divided into sections. While the first section may narrate the descriptive characteristics of the problem under enquiry, the second and subsequent sections may focus on empirical results based on deeper insights of the problem of study. Each chapter based on research studies mostly contain major headings, sub headings, quotations drawn from observations made by earlier writers, footnotes and exhibits.

Use of References:

There are two types of reference formatting. The first is the ‘in-text’ reference format, where previous researchers and authors are cited during the building of arguments in the introduction and discussion sections. The second type of format is that adopted for the Reference section for writing footnotes or Bibliography.

Citations in the text

The names and dates of researchers go in the text as they are mentioned e.g., “This idea has been explored in the work of Smith (1992).” it is generally unacceptable to refer to authors and previous researchers etc.

Examples of Citing References (Single Author)

Duranti (1995) has argued or it has been argued that (Duranti, 1995)

In the case of more authors,

Moore, Maguire, and Smyth (1992) proposed or it has been proposed that(Moore, Macquire, & Smyth, 1992)

The reference section:

The report ends with reference section, which comes immediately after the Recommendations and begins on a new page. It is titled as ‘References’ in upper and lower case letters centered across the page.

Published Journal Articles

Beckerian, D.A. (1993). In search of the typical eyewitness.*American psychologist*, 48, 574-576.

Gubbay, S.S., Ellis, W., Walton, J.N., and Court, S.D.M. (1965).Clumsy Children: a study of apraxic and agnosic defects in 21 children.*Brain,* 88, 295-312.

Authored Books

Cone, J.D., and Foster, S.L. (1993).*Dissertations and theses from startto finish: psychology and related fields.* Washington, Dc: AmericanPsychological Association.

Cone, J.D., and Foster, S.L. (1993).*Dissertations and theses from start tofinish: psychology and related fields* (2ndEd.). Washington, Dc: AmericanPsychological Association.

APPENDICES:

The purpose of the appendices is to supplement the main body of your text and provide additional information that may be of interest to the reader.

There is no major heading for the appendices. You simply need to include each one, starting on a new page, numbered, using capital letters, and headed with a centered brief descriptive title. For example:

Appendix A: List of stimulus words presented to the participants

Dos and Don’ts of Report Writing

- Choose a font size that is not too small or too large; 11 or 12 is a good font size to use.
- Acknowledgment need not be a separate page, except in the final report. In fact, you could just drop it altogether for the first- and second-stage reports. Your guide already knows how much you appreciate his/her support. Express your gratitude by working harder instead of writing a flowery acknowledgment.
- Make sure your paragraphs have some indentation and that it is not too large. Refer to some text books or journal papers if you are not sure.
- If figures, equations, or trends are taken from some reference, the reference must be cited right there, even if you have cited it earlier.
- The correct way of referring to a figure is Fig. 4 or Fig. 1.2 (note that there is a space after Fig.). The same applies to Section, Equation, etc. (e.g., sec. 2, eq. 3.1).
- Cite a reference as, for example, “The threshold voltage is a strong function of the implant dose [1].” note that there must be a space before the bracket.
- Follow some standard format while writing references. For example, you could look up any IEEE transactions issue and check out the format for journal papers, books, conference papers, etc.
- Do not type references (for that matter, any titles or captions) entirely in capital letters. The only capital letters required are (i) the first letter of a name, (ii) acronyms, (iii) the first letter of the title of an article (iv) the first letter of a sentence.
- The order of references is very important. In the list of your references, the first reference must be the one which is cited before any other reference, and so on. Also, every reference in the list must be cited at least once (this also applies to figures). In handling references and figure numbers, Latex turns out to be far better than Word.

- Many commercial packages allow “screen dump” of figures. While this is useful in preparing reports, it is often very wasteful (in terms oftoner or ink) since the background is black. Please see if you can invert the image or use a plotting program with the raw data such that the background is white.

The following tips may be useful: (a) for windows, open the file in Paint and select Image/Invert Colors. (b) For Linux, open the file in Image Magick (this can be done by typing display) and then selecting Enhance/Negate.

- As far as possible, place each figure close to the part of the text where it is referred to.
- A list of figures is not required except for the final project report. It generally does not do more than wasting paper.
- The figures, when viewed together with the caption, must be, as far as possible, self-explanatory. There are times when one must say, “see text for details”. However, this is an exception and not a rule.
- The purpose of a figure caption is simply to state what is being presented in the figure. It is not the right place for making comments or comparisons; that should appear only in the text.
- If you are showing comparison of two (or more) quantities, use the same notation through out the report. For example, suppose you want to compare measured data with analytical model in four different figures, in each figure, make sure that the measured data is represented by the same line type or symbol. The same should be followed for the analytical model. This makes it easier for the reader to focus on the important aspects of the report rather than getting lost in lines and symbols.
- If you must resize a plot or a figure, make sure that you do it simultaneously in both x and y directions. Otherwise, circles in the original figure will appear as ellipses, letters will appear too fat or too narrow, and other similar calamities will occur.
- In the beginning of any chapter, you need to add a brief introduction and then start sections. The same is true about sections and subsections. If you have sections that are too small, it only means that there is not enough material to make a separate section. In that case, do not make a separate section. Include the same material in the main section or elsewhere.

Remember, a short report is perfectly acceptable if you have put in the effort and covered all important aspects of your work. Adding unnecessary sections and subsections will create the impression that you are only covering up the lack of effort.

❖ Do not make one-line paragraphs.

- Always add a space after a full stop, comma, colon, etc. Also, leave a space before opening a bracket. If the sentence ends with a closing bracket, add the full stop (or comma or semicolon, etc) after the bracket.
- Do not add a space before a full stop, comma, colon, etc.
- Using a hyphen can be tricky. If two (or more) words form a single adjective, a hyphen is required; otherwise, it should not be used. For example, (a) A short-channel device shows a finite output conductance.
    - This is a good example of mixed-signal simulation. (c)Several devices with short channels were studied.
- If you are using Latex, do not use the quotation marks to open. If youdo that, you get "this". Use the single opening quotes (twice) to get "this".
- Do not use very informal language. Instead of "this theory should be taken with a pinch of salt," you might say, "this theory is not convincing," or "it needs more work to show that this theory applies in all cases."
- Do not use "&"; write "and" instead. Do not write "There're" for "There are" etc.
- If you are describing several items of the same type (e.g., short-channel effects in a MOS transistor), use the "list" option; it enhances the clarity of your report.
- Do not use "bullets" in your report. They are acceptable in a presentation, but not in a formal report. You may use numerals or letters instead.
- Whenever in doubt, look up a text book or a journal paper to verify whether your grammar and punctuation are correct.
- Do a spell check before you print out your document. It always helps.
- Always write the report so that the reader can easily make out whatyour contribution is. Do not leave the reader guessing in this respect.

Above all, be clear. Your report must have a flow, i.e., the reader must be able to appreciate continuity in the report. After the first reading, the reader should be able to understand (a) the overall theme and (b) what is new (if it is a project report).

Plagiarism is a very serious offense. You simply cannot copy material from an existing report or paper and put it verbatim in your report. The idea of writing a report is to convey in your words what you have understood from the literature.

The above list may seem a little intimidating. However, if you make a sincere effort, most of the points are easy to remember and practice.

A supplementary exercise that will help you immensely is that of looking for all major and minor details when you read an article from a newspaper or a magazine, such as grammar, punctuation, organization of the material, etc.

PRESENTATION OF A REPORT

In this section, we will look into the issues associated with presentation of a Research Report by the Researcher or Principal Investigator. While preparing for the presentation of a report, the researchers should focus on the following issues:

What is the purpose of the report and issues on which the Presentation has to focus?

Who are the stakeholders and what are their areas of interest?

The mode and media of presentation.

Extent of Coverage and depth to address at.

Time, Place and cost associated with presentation.

Audio – Visual aids intended to be used.

Conclusion

If your paper was written to argue a point or to persuade the reader, then your conclusion will summarize the main points of your arguments presented in the paper. You will also want to restate your thesis and conclude with a statement of your position on the topic.

Questions

1. What are the types of report?
2. How the research report are essential?
3. What are the Do’s and Don’t of report writing?
4. How do you present a report?

Lesson – 21

CHARACTERISTICS OF A GOOD RESEARCH REPORT

Introduction

The introduction to a research paper can be the most challenging part of the paper to write. The length of the introduction will vary depending on the type of research paper you are writing. An introduction should announce your topic, provide context and a rationale for your work, before stating your research questions and hypothesis. Well-written introductions set the tone for the paper, catch the reader's interest, and communicate the hypothesis or thesis statement.

QUALITIES AND CHARACTERISTICS OF GOODS RESEARCH REPORT

A lot of reports are written daily. Some of them are intended to document the progress of some activities, feasibility reports, investigation reports, some of the reports are for monitoring purposes, some are evaluation reports but it is clear that all the reports have some objective and purpose behind it. That objective and purpose can only be achieved if a report has the following qualities and characteristics:

1. It should be factual: Every report should be based on facts, verified information and valid proofs.
2. Clear and Easily understandable: Explained below
3. Free from errors and duplication
4. Should facilitate the decision makers in making the right decision:
5. Result focused and result oriented
6. Well organized and structured
7. Ethical reporting style

Reader-Friendly

Readers are various stakeholders who receive reports generated by M&E. If reports are reader-friendly, they are likely to be read, remembered and acted upon. Following decisions need to be made by CSOs to make their reports reader-friendly:

- What do they need to know?
- When do they need to know?
- How do they like to know?

Easy, Simple Language

M&E reports are meant to inform not impress. Using easy, simple language, be it Urdu or English makes the report friendly on reader. To do this, here are some useful tips:

- Write only what is necessary
- Avoid repetition and redundancy
- Give interesting and relevant information
- Avoid preaching or lecturing
- Compose short and correct sentences

Purposeful Presentation

Each report has some objective(s) to meet. The “objective” comes from analyzing the needs of the reader. A CSO is working for a project that has several donors, and is channeled through an agency that needs to be informed about some specific things going on in the field. CSOs reports are the main pathways or channels of information to the people who decide to fund this and other such projects. Similarly, field reports are the amin vehicles for the management of the CSOs to make decision regarding the project itself. A good report presents facts and arguments in a manner that supports the purpose of the report.

Organized and Well-Structured

Each CSO comes up with a format of internal reporting to suit its requirements. Reporting to donors is done on their prescribed formats. The M&E system should be able to generate information that can be organized using different formats. In the annex, this manual provides some useful formats that can be customized by a CSO.

Result-Focused

In general, all readers are interested in the RESULTS. Therefore, one over-riding principle that CSOs should aim for in all report writing is to report on the results of their activities. This requires some analysis on their part that goes beyond a mere description of their activities. Result-focused means that description of activities is liked with the project objectives. This aspect must be addressed especially in the project progress reports. According to Phil Bartle, “A good progress report is not merely a descriptive activity report, but must analyze the results of those reported activities. The analysis should answer the question, "How far have the project objectives been reached?"

Timely Prepared and Dispatched

M&E generate “Information Products”, a customized set of information according to needs to a defined group of users. M&E’s information products are time-bound for both internal and external stakeholders. Reports, in suitable formats, need to be timely produced and made available to the readers. It is useful to develop an Information Product Matrix (IPM) like the one described below:

Straightforward

A good report is straight forward, honest description. It contains no lies, no deception, no fluff. It is neat, readable and to-the-point. It is well spaced, has titles and subtitles and is free of language errors.

TOP 10 QUALITIES OF GOODS ACADEMIC RESEARCH

Academic Research is defined as a process of collecting, analyzing and interpreting information to answer questions or solve a problem. But to qualify as good research, the process must have certain characteristics and properties: it must, as far as possible, be controlled, rigorous, systematic, valid and verifiable, empirical and critical. The main characteristics for good quality research is listed below:

1. It is based on the work of others.
2. It can be replicated and doable .
3. It is generalisable to other settings.
4. It is based on some logical rationale and tied to theory. In a way that it has the potential to suggest directions for future research.
5. It generates new questions or is cyclical in nature.
6. It is incremental.
7. It addresses directly or indirectly some real problem in the world.
8. It clearly states the variables or constructs to be examined.
9. Valid and verifiable such that whatever you conclude on the basis of your findings is correct and can be verified by you and others.
10. The researcher is sincerely interested and/or invested in this research.

Meanwhile, bad research has the following properties:

1. The opposites of what have been discussed.
2. Looking for something when it simply is not to be found.
3. Plagiarizing other people’s work.
4. Falsifying data to prove a point.
5. Misrepresenting information and misleading participants.

Report provides factual information depending on which decisions are made. So everyone should be taken to ensure that a report has all the essential qualities which turn it into a good report. A good report must have the following qualities:

1.Precision: In a good report, the report writer is very clear about the exact and definite purpose of writing the report. His investigation, analysis, recommendations and others are directed by this central purpose. Precision of a report provides the unity to the report and makes it a valuable document for best usage.

2.Accuracyof Facts: Information contained in a report must be based on accurate fact. Since decisions are taken on the basis of report information, any inaccurate information or statistics will lead to wrong decision. It will hamper to achieve the organizational goal.

3.Relevancy: The facts presented in a report should not be only accurate but also be relevant. Irrelevant facts make a report confusing and likely to be misleading to make proper decision.

4.Reader-Orientation: While drafting any report, it is necessary to keep in mind about the person who is going to read it. That's why a good report is always reader oriented. Readers knowledge and level of understanding should be considered by the writer of report. Well reader-oriented information qualify a report to be a good one.

5.SimpleLanguage: This is just another essential features of a good report. A good report is written in a simple language avoiding vague and unclear words. The language of the report should not be influenced by the writer's emotion or goal. The message of a good report should be self – explanatory.

6.Conciseness: A good report should be concise but it does not mean that a report can never be long. Rather it means that a good report or a business report is one that transmits maximum information with minimum words. It avoids unnecessary detail and includes everything which are significant and necessary to present proper information.

7.GrammaticalAccuracy: A good report is free from errors. Any faulty construction of a sentence may make its meaning different to the reader's mind. And sometimes it may become confusing or ambiguous.

8.UnbiasedRecommendation: Recommendation on report usually make effect on the reader mind. So if recommendations are made at the end of a report, they must be impartial and objective. They should come as logical conclusion for investigation and analysis.

9.Clarity: Clarity depends on proper arrangement of facts. A good report is absolutely clear. Reporter should make his purpose clear, define his sources, state his findings and finally make necessary recommendation. To be an effective communication through report, A report must be clear to understand for making communication success.

10.AttractivePresentation: Presentation of a report is also a factor which should be consider for a good report. A good report provides a catchy and smart look and creates attention of the reader. Structure, content, language, typing and presentation style of a good report should be attractive to make a clear impression in the mind of its reader.

The inclusion of above factors, features or characteristics, make a good report to be effective and fruitful. It also helps to achieve the report goal. A reporter who is making the report, always should be careful about those factors to make his report a good one.

It Has Been Proven That Visual Aids (Charts, Graphs, Diagrams, Images, Etc.) Increase The Effectiveness Of Research Posters

Be sure to use the right visual aids to have the right effect! An attractive poster will catch a reader's eye at first glance and draw them in to read your poster or listen to your presentation.

Whenever possible use an image or graphic to portray information rather than a paragraph or text.

A Few Things To Consider When Inserting Images And Graphs:

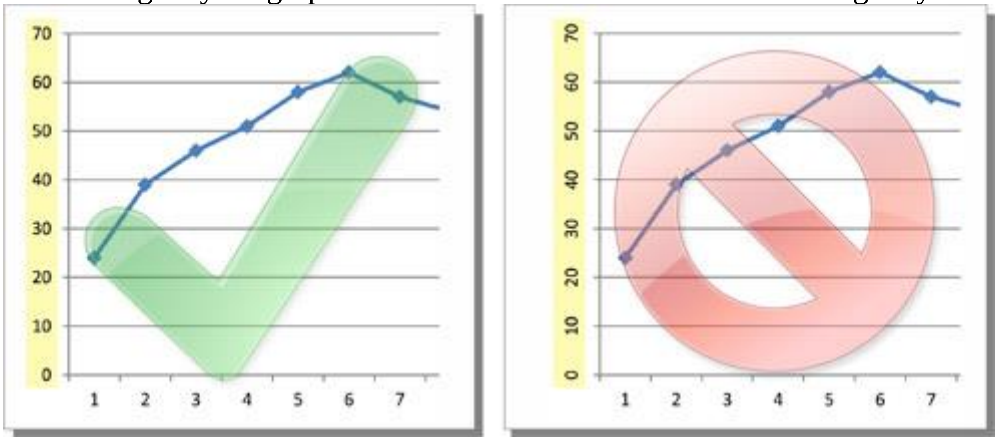
Respect the fact that the viewer has a limited amount of time to spend looking at posters by making sure the time spent at your poster is effective and informative. Communicating a concept visually will help you to communicate the message faster, and will help the reader retain the information.

Be sure to include a brief caption for your figures, and explicitly refer to the figure in the text.

Charts and graphs must include a title and labels for each axis in order to be meaningful. Failure to include these will require you to constantly explain what the chart or graph means. Consider using QR codes to link to supplemental materials1. They are a great way to get your viewer more information. You can use it to include your contact information, a link to a webpage, or even an online video that further explains your research!

Effective Graphs & Charts

Be sure to give your graphs and charts clear and concise titles that give your reader direction.



Interpreting legends is sometimes very difficult, and you should do anything in your power to make your graphs clearly understood by the viewer. Don’t just take you graph “as is” from the program you have created it in. Add ‘arrows’ or ‘callouts’ to directly label various elements. Giving the reader additional visual cues (e.g. below) will help the reader better understand the information on the graph without having to get into nitty gritty of the data they contain.

Avoid using labels on your graphs that are aligned vertically. It’s not comfortable or natural to have to tilt you head to the side to read something, so don’t make your readers do this. The less painful the interaction with your poster is, the better.

Do not use acronyms or abbreviations unless they are widely used within your field of study. If you do use abbreviations, be sure to first write out the full word or phrase followed by the abbreviated term in parenthesis. Afterwards, you can use the abbreviation alone. This will provide a reference to readers who may not be familiar with the abbreviation.

Create a Pattern

No we're not suggesting you use a plaid background or fill your boxes with polka dots; we're talking about the flow of colors on your poster.

When using different colors for your text, headlines, etc. be sure to use the colors in the same way throughout your poster. If you use blue for the heading of one section, make sure to use blue for all of the other headings as well. The goal is to cause recognition and aide in the navigation of your poster, not to introduce chaos.

Consider the Color Blind

Sometimes the colors in a given image are unavoidable, but when choosing the colors you use in charts and graphs keep in mind that not everyone can see the same variations in color. There are two simple things to keep in mind when considering visibility your viewers who may have color deficiencies. You can learn more about designing with partial sight and color deficiency in mind here

1. Increase the contrast of the colors that will be used next to each other. For example yellow on red would be a bad color combination, but if you use a light yellow and a deep burgundy it will be much easier to see the difference.
2. Pick colors from opposite sides of the color wheel. The further away the colors are from each other on the color wheel, the easier it is to tell them apart.

Conclusion

A conclusion is the last paragraph in your research paper, or the last part in any other type of presentation. You restate your thesis and summarize your main points of evidence for the reader.You can usually do this in one paragraph.

Questions

1. What are the top 10 qualities of good academic research ?
2. Explain the qualities and characteristics of good research report.
3. What are the features of good report?
4. Explain the characteristics of pictures and graphs.
5. How do you draw the effective graphs and charts?

Lesson – 22

INTRODUCTION TO SPSS

Introduction

SPSS Statistics is a software package used for logical batched and non-batched statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009. The current versions

(2015) are officially named IBM SPSS Statistics. Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection, now divested under UNICOM Intelligence), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services).

The software name originally stood for Statistical Package for the Social Sciences (SPSS),[2] reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing.

Abbreviation of SPSS

SPSS is the abbreviation of Statistical Package for Social Sciences and it is used byresearchers to perform statistical analysis. As the name suggests, SPSS statistics software is used to perform only statistical operations. ... Interpreting SPSS output. Statistical analysis of SPSS data output.

Benefits of SPSS

Statistical analysis can be conducted using two main methods. One is simply by using a generalized spreadsheet or data management program such as MS Excel or through using a specialized statistical package such as SPSS. Here are key reasons why SPSS is the best option to use.

1. Effective data management

While it is spot on that a spreadsheet program offers more control with regards to the data organization , this can also be seen as a demerit. In contrast, you cannot move data blocks in SPSS as it is meant for organizing data in an optimal manner. A row represents one case, whereas a column denotes one variable. SPSS makes data analysis quicker because the program knows the location of the cases and variables. When using a spreadsheet, users must manually define this relationship in every analysis.

2. Wide range of options

SPSS is specifically made for analyzing statistical data and thus it offers a great range of methods, graphs and charts. General programs may offer other procedures like invoicing and accounting forms, but specialized programs are better suited for this function. SPSS also comes with more techniques of screening or cleaning the information in preparation for further analysis.

Furthermore, normal spreadsheet programs may only support data analysis immediately following installation, with extra plug-ins being required for accessing more intricate techniques.

3. Better output organization

SPSS is designed to make certain that the output is kept separate from data itself. In fact, it stores all results in a separate file that is different from the data. However, in programs like Excel, results of an analysis are placed in one worksheet and there is a likelihood of overwriting other information by accident.

Even though Excel still offers a good way of data organization, using dedicated software like SPSS is more suitable for in depth data analysis.

IMPORTANCE AND BENEFITS OF SPSS IN RESEARCH

SPSS statistics is a software package used for logical batched and non-batched statistical analysis. This software is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts and plots of distributions and trends, descriptive statistics and conduct complex statistical analyses. This packages of program is available for both personal and mainframe computers.

SPSS is beneficial for both qualitative and quantitative data equal importance is been given for both data sets, about 85% of the research scholars carry quantitative data for their further analysis, so the layman thinks that the SPSS software is more beneficial in quantitative data than qualitative one but no SPSS gives equal weight age for both data sets. Even if there are several software's available in market to analyze quantitative data SPSS is more preferable than other software's. Since SPSS is user friendly software & ease to use for the beginners and also helps in analysis even when the data set goes larger.

How it is benefit for research scholars?

Mainly for the research scholars who are versatile in their own core papers and quite weak in software part can prefer SPSS software for their further analysis as SPSS gives a perfect graphical representation and also an appropriate result for the data that has been entered. SPSS is just a drag and drop process which has almost all basic and some advanced statistical analysis which helps the research scholars to easily adapt to this software and can do the analysis part and attain their result.

Applicability of SPSS statistics for data analysis

There are 15 modules IBM SPSS statistics for data analysis according to the research needs.

- SPSS **Exact Tests module** enables one to use small samples and still feel confident about the results

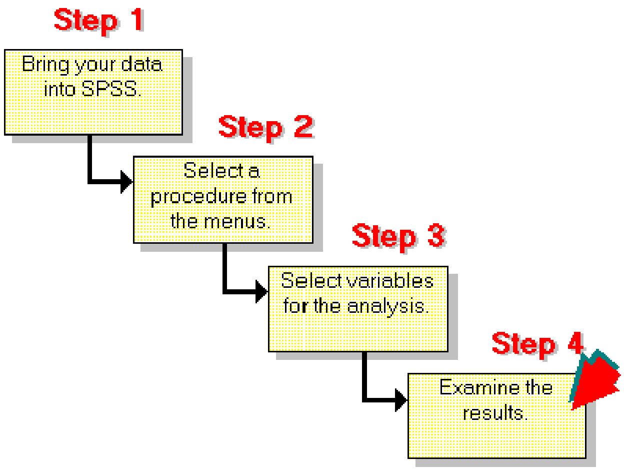
- SPSS **forecasting module** enables analysts to predict trend and develop forecasts quickly and easily-without being an expert statistician
- SPSS **Missing Values module** finds the relationships between any missing values in your own data and other variables. Missing data can seriously affect your models and results. It is used by survey researchers, social scientists, data miners and market researchers to validate data.

The SPSS Model

SPSS has similarities to both spreadsheet applications and databases. SPSS requires you to create a database, and then perform statistical manipulations with the data. The key difference between SPSS and these other applications is SPSS is designed for statistical analysis. Thus, it has far better and more efficient statistical capabilities than most spreadsheet and database applications.

Analyzing data with SPSS is relatively easy. You'll need to complete several steps to perform your analysis:

1. Get your data into SPSS.You can open a previously saved SPSS data file; read a spreadsheet, dBASE, or text data file; or enter your data directly in the Data Editor.
2. Select a procedure from the menus to calculate statistics or create a chart. You can select a procedure to perform statistical analysis or select a procedure to create high-resolution charts.
3. Select the variables you want to use in the analysis.The variables in the data file are displayed in a dialog box for the procedure.
4. Run the procedure and look at the results.


Step 1
Bring your data into SPSS.
Step 2
Select a procedure from the menus.
Step 3
Select variables for the analysis.
Step 4
Examine the results.

5. It is difficult to adequately describe how SPSS works without the benefit of the software. Thus, we are meeting in the SSIL lab for the SPSS sessions.

Conclusion

SPSS Statistics is a quick way to get others on board for more analysis. You can use RFM modeling to gain deeper insight into your customers' behavior, whether it is in retail, e-commerce, distribution, or other commercial industries. Even charities can apply this model to improve interactivity with donors.

RFM (Recency, Frequency and Monetary value)analysis is, relatively speaking, an easy modeling process to understand. Business users can see the value quickly. Use it to leverage a deeper use of analytics in your organization. It is a great starting point for finding more and interesting ways to bring data mining and predictive analytics into your company.

Questions

1. Expand SPSS.
2. What are the benefits of SPSS package?
3. State the importance of SPSS package.
4. How the SPSS is benefit of Research scholar?
5. State the SPSS models.