UTTARAKHAND OPEN UNIVERSITY
Teenpani Bypass Road, Transport Nagar, Haldwani - 263 139
Phone No. : (05946) - 286002, 286022, 286001, 286000
Toll Free No. : 1800 180 4025
Fax No. : (05946) - 264232, email : <info@uou.ac.in>
http://www.uou.ac.in

ENS - 602

Research Methodology for Environmental Studies

ENS - 602

Research Methodology for Environmental Studies

Department of Forestry and Environmental Science
School of Earth and Environmental Science

Uttarakhand Open University
Haldwani, Nainital (U.K.)

ENS 602

Research Methodology for Environmental Studies

UTTARAKHAND OPEN UNIVERSITY

SCHOOL OF EARTH AND ENVIRONMENTAL SCIENCE

University Road, Teenpani Bypass, Behind Transport Nagar, Haldwani - 263 139

Phone No. : (05946) - 286002, 286022, 286001, 286000

Toll Free No. : 1800 180 4025, Fax No. : (05946) - 264232,

e-mail: info@uou.ac.in, Website: http://www.uou.ac.in

Board of Studies

Programme Coordinator

Dr. H.C. Joshi
Department of Forestry and Environmental Science,
SoEES, Haldwani, Nainital (U.K.)

Editors

Dr. Rajiv Pandey,
Division of Forestry Statistics,
Indian Council of Forestry Research and Education (ICFRE), Dehradun (U.K.)

H.C. Joshi, Beena T. Fulara, Krishna K. Tamta,
Neha Tiwari, Preeti Pant, Deepti Negi & Khashti Dasila
Department of Forestry and Environmental Science
Uttarakhand Open University, Haldwani (U.K.)

Unit Writers

Unit Writers	Unit No.
Dr. Sushil Bhadula, Department of Zoology, Swami Vivekanand college of education. Roorkee, (U.K.)	**1,2,3,4 and 5**
Dr. Vikas Saini, Department of Zoology, Swami Vivekanand college of education. Roorkee, (U.K.)	**6, 7 and 10**
Adapted from E-PG Pathshala, EVS/SAES-XIV/1,6,7,8,9, Statistical Applications in Environmental Sciences	**8**
Dr. Samiksha Bahukhandi, HEC group of institutions, Haridwar (U.K)	**9,11, 12,13,14**

Cover Page Design and Format Editing

Dr. Beena Tewari Fulara, Dr. Krishna Kumar Tamta, Dr. H.C. Joshi
Department of Forestry and Environmental Science
Uttarakhand Open University, Haldwani

Title : Research Methodology for Environmental Studies
ISBN : XXXX-XXXX
Copyright : Uttarakhand Open University
Edition : 2024 (Restricted Circulation)
This is the first copy of the contents subjected to final editing later.

Published By :
Printed at :

Disclosure: This is the first copy of the contents subjected to final editing later. Unit no. 8 is adapted from E-PG Pathshala under Creative Commons License.

Table of Content

Unit 1: Research Methodology: Meaning, Objective, Types; approaches and process; criteria of good research

Unit Structure

1.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is research?
- What is meaning and definitions of research?
- What is research methodology?
- What is importance of the research in environmental studies?
- What are the objectives of research?
- What are the types of Research?
- What are research approaches and process?
- What are the criteria of good research?

1.1. Introduction

The word “research” has been taken from French word “recherche” which means quest, search, pursuit and search for truth i.e., “to go about seeking”. In common understanding, research refers to a search for knowledge and includes the creative and systematic work conducted to increase the knowledge about nature, human,

culture and society. It is a careful investigation or inquiry especially through systematized effort to gain new knowledge or new facts in any branch of knowledge. Research is applicable in all disciplines such as chemistry, zoology, botany, social sciences, political sciences, environmental sciences, economics, businesses. Research not only provides base line information and data of the concern subject rather also provides the solutions of the particular problem. By using these information and data, governments and other decision-making agencies seeks the remedial measures against that particular problem. Research also helps in understanding various processes, phenomenon and issues. As you know there are various environmental processes, phenomenon and issues such as global warming, climate change, ozone layer depletion, pollution, depletion of natural resources, over exploitation of natural resources, biodiversity degradation, solid waste generation. The question is that how we can elaborate these environmental issues? Answer is simple! Because of research, research provides the data and information on these critical issues; therefore, we are aware about these environmental issues. Scientists/researchers provide the data based on long term research on these issues. Research also solves the local socio-economical, environmental, political and business issues. Various technologies have been developed through the research like bioremediation, sewage treatment, eco-friendly technologies etc... On the view of above account, we can explicitly say that, research provides knowledge and technologies for the welfare of human and society, therefore research is boon to human being. In this unit you will learn about concept of research and research methodology, meaning approaches and process of research and criteria of good research.

1.2. Definitions of Research

- According to Clifford Woody "Research comprises defining and redefining problems, formulating hypothesis or suggested solutions, collecting, organizing and evaluating data, making deduction and reaching conclusions and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis.

- According to D. Slesinger and M. Stephenson "the concepts or symbols for the purpose of generalizing to extend, correct or verify facts, whether that knowledge aids in construction of theory or in the practice of an art".
- In simple words, research can be defined as "an original contribution to the existing stock of knowledge making for its progression. It is the search of truth with the help of investigation, study, observation, comparisons and experiments".

1.3. Research methodology

Research may also be defined as "the search of knowledge or facts through objectives and systematic methods of finding solution to a research problem is called as research".

Meaning of Research: Generally, research refers to a search for facts, information and observations. Research means scientific and systematic search for pertinent information on a particular topic. Scientific research is a systematic, controlled, empirical and critical investigation of hypothetical propositions about the natural phenomenon.

Research is a careful investigation as a movement from the known to the unknown. It is literally a passage or path of discovery. Human being has the instinct of curiosity. When the unknown tackles human, more and more our inquisitiveness or curiosity make us probe, which results into learning and attaining understanding of the unknown things. This curiosity or inquisitiveness is mother of all the knowledge and the invention of methods, processes, etc which one employs for obtaining the knowledge of whatever the mysterious, can be termed as research.

Loosely, we can say the research is an academic activity and term research should be used in a technical sense.

Meaning of Research methodology: As you know that research methods are the methods which used by researchers in performing research. Various methods and technique are used for conducting research and these methods and techniques are collectively called as research methods. Science or study of the research method is called research methodology.

There are two terms related to research methodology: first is Research techniques and second is methods. Research techniques involves the behaviors and instruments we use in performing research operations such as making observation, recording data etc. research methods refer to the behavior and instruments used in selecting and constructing research technique.

Difference between techniques and methods are given below:

Table-1: Difference between research techniques and research methods (Source: Book on Research methodology methods and techniques: CR Kothari and Gaurav Garg)

Type	Techniques	Methods
Library Research	1. Recording of notes, content analysis, tape and film listening and analysis. 2. Statistical compilations and manipulations, reference and abstract guides, content analysis.	1. Analysis of historical records. 2. Analysis of Documents.
Field Research	1. Observational behavioural scales, use of score cards etc. 2. Interactional recording, possible use of tape recorders, photographic techniques 3. Recording mass behavior, interview using independent observers in public places. 4. Identification of social and economic background of respondents. 5. Use of attitude scales, projective techniques, use of sociometric scales. 6. Interviewer uses a detailed schedule with open and closed questions. 7. Interviewer focuses attention upon a given experience and its effects. 8. Small group of respondents are interviewed simultaneously. 9. Use as a survey technique for information and for discerning opinion: may also be used as follow up of questionnaire.	1. Non-participant direct observation 2. Participant observation 3. Mass observation 4. Mail questionnaire 5. Opinionnaire 6. Personal interview 7. Focused interview 8. Group interview 9. Telephone survey 10. Case study and life history

	10. Cross sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.	
Laboratory Research	1. Use of audio-visual recording devices, use of observers etc.	1. Small group study of random behavior, play and role analysis.

Research methodology is a scientific method to solve the research problem. In research methodology we study the different steps that are adopted by scientists/researchers in studying his research problem. Researchers not only require the knowledge of instruments but should also know the following methods:

- How to handle instruments?
- What are the principles of instruments?
- What are the standard books and documents related to his research?
- What are the prescribed standard methods of data collection for his study?
- How to do data analysis as calculate mean, mode, median, standard deviation, chi square, correlation?
- Where different tests are applied?
- What should be the interpretations of that research?

All this mean that it is necessary for the researcher to formulate and design research methodology for his problem. Research methodology may vary from problem to problem. For example, an environmentalist who works in ecology on the problem "impact of sewage on aquatic system"; has to design the research methodology for his problem. He has to analyze the sources of sewage, water quality parameters (ex: temperature, pH, dissolved oxygen, biological oxygen demand etc.), plankton (phytoplankton and zooplankton) diversity, fish diversity etc. The environmentalist should require the following methodology.

- Standard book of the American Public Health Association (APHA), manuals of instruments etc.

- Chemicals and glassware to conduct the analytic work in the laboratory
- International and National Standards of water quality parameters for different uses
- Instruments like thermometer, pH meter, dissolved oxygen meter, BOD incubator etc.
- Research papers related to research problem (impact of sewage on aquatic ecology in this case).
- Standard books for identification of plankton (Freshwater biology-Edmondson) and fishes (Day Fauna) etc.
- Preservatives such as alcohol, formalin etc. and other chemicals
- Camera, computer, printer etc.
- Instruments such as Microscope.
- Knowledge of biostatistics and other relevant tests for this research problem.
- Collaborators and other experts from the research domain

After careful study of the above mention methods/techniques, researcher may develop the research methodology for scientific research problem. Therefore, when we discuss about the research methodology, we not only consider about the research methods but we also consider the hypothesis, reason behind the research problem and logic behind the method use in the research problem. Research methodology also includes for following questions.

- How the research problem has been defined?
- Why this research problem is important?
- Why the hypothesis has been framed and undertaken?
- Why specific technique has been adopted?

1.4. Role and Importance of Research in Environmental Studies

It is well said that “all progress is born of curiosity or inquiry” “Doubts are even better than the over confidence”. Curiosity leads to invention. Research inculcates systematic thinking and inductive philosophy and also supports development of habits for logical thinking. Environmental problem is one of the great issues at national and international level. As you know, our natural resources are depleting

very fast and air, water, and soil are being contaminated. Biological diversity is decreasing and many species are at the verge of extinction. Therefore, research in environmental science has a great opportunity to scientists, researchers and others, who are working on different aspects of environmental science. Research can be a great tool to solve all the environmental problems. As you know that scientists invented LED bulb which is eco-friendly and use less amount of energy as compared to traditional bulb. These bulbs have long durability, therefore generate less amount of solid waste. The invention of LED bulb is results of research. There are various roles of research in environmental science, some of them are being discussed below:

In Biodiversity Management: As you know that biodiversity is variety and variability among all species on a spatial scale. This biological diversity degrading day by day due to habitat loss, pollution, poaching, over exploitation etc. Research provides the basic data of biodiversity. By using these data and other associated information, scientists seek the solution for biodiversity degradation. Have you heard the term "Endangered species"? This term is used for those species, which are less in numbers and may become extinct in near future. The number of individual species is estimated by research. IUCN categorize the species as: Extinct, critically endangered, Endangered, Vulnerable, Threatened, Data deficit etc. This categorization of species is possible through research. Therefore, research is very important tool in biodiversity assessment and management. We can conserve, protect and preserve the biodiversity by generating data and information through research.

In Air quality analysis: Air is most essential components of life. Without food and water, we can survive for few days but without air we cannot survival for several minutes. Research not only provides the air quality data but also identify the sources and remedial measures of air pollution. Researchers conducted research on air pollution and provide the Air Quality Index. Researchers generally measure air pollutants as RSPM, SPM, CO_2, O_3, Oxides of Sulphur and oxide of Nitrites. As you know that Government of Delhi banned the 10 years old vehicles to minimize the air pollution. Research also provides the solutions to mitigate the air pollution.

In Water quality analysis: You are well aware about the importance of water. It is used for variety of purposes as drinking, bathing, irrigation, industrial purpose etc. When it becomes polluted, it causes numerous health problems. Research provides the data of water quality by analyzing parameters of water quality by using different methodologies and interprets the quality of water. The drinking water quality parameters are: pH, total dissolved solids, chlorides, total coliforms, heavy metals (Mercury, Arsenic, Chromium, Cadmium etc.) and iron. After the research, scientist has recommended that which water is good for drinking and which water is good for bathing and irrigation purposes. Besides the drinking water quality, researches also provide the water quality of pond, lake, river and other aquatic bodies. On the basis of research Governments and other reputed agencies declared the most polluted rivers, most polluted lakes, most polluted wetlands of the world. Research also provides remedial measures for water pollution.

In Soil Quality analysis: Soil is upper most layer of earth. Human being depends on soil for food, fodder and several other components. Soil quality is also an important factor. Researcher analyze the soil for its various physical and chemical properties as organic matter, moisture content, water holding capacity, chlorides, temperature, pH, heavy metals etc. Research provides the data and information about the quality of soil. Using these data, farmers practiced the different methods of agriculture and also sow matching crops on their farm lands. Various modern agriculture technologies such as bioremediation, use of bio-fertilizers, vermicomposting and organic farming are results of researches.

In Policies making: Researches also analyze various components requiring policy formulation and modification. Ecologists, environmentalists, sociologists etc. conduct various researchers such as Environmental Impact Assessment (EIA) to mitigate the environmental problems. Any project which is being implemented required EIA and SIA which are also result of research.

Conservation of Natural Resources: As you know the natural resources are ecologically, environmentally, socially and economically very important. Researches on natural resources provide basic data to conserve the natural resources at local and global level. Researches also estimate the contribution of

natural resources to the communities. Government of different countries declared some spatial natural patches as the national parks, wildlife sanctuaries, biosphere reserves, conservation sites etc. on the basis of researches and investigations.

Besides the environmental studies research is an important tool in other sciences and technologies. Research also used in economics, businesses, marketing and social sciences. In addition to above importance, research can be beneficial in following ways:

- Research is important to those researchers who are doing Ph.D. and it is may mean a careerism or a step to attain high position in the society.
- Research is important for generating new ideas, insights and thoughts.
- Research may be useful in development of new innovative and creative work.
- Research may develop new theories and hypothesis.
- Research may generate new data for variety of application.
- Research may develop new technologies.

1.5. Objectives of Research

The purpose of research is to discover unknown to known and answers to questions through the application of scientific procedures. The main aim of research is to find out the hidden truth about the inquiry or question under investigation. The research objectives may broadly classify into the following based on the nature of research inquiries:

1. Exploratory or formulative research - deals with gaining familiarity with a phenomenon or to achieve new insights;
2. Descriptive research – deals with description of the phenomenon or about the inquiry;
3. Diagnostic research – deals with evaluation of relationship or causes of happenings of the phenomenon or the inquiry;
4. Hypothesis-testing research – deals with testing of a hypothesis about the inquiry.

1.6. Types of Research

There are various types of research which are summarized in Fig-1 and also discussed below:

1. **Descriptive vs Analytical Research:** Descriptive research comprises surveys and facts- finding enquiries of different kinds. The major objective of descriptive research is description of the state of affairs of the enquiry as it exists at present. In social science and business research, we quite often use the term ex post facto research for descriptive research studies. The main feature of this descriptive method is that the researchers have no control over the variables. Researchers can only describe and report what has happened or what is happening. Most ex post facto research projects are used for descriptive studies in which the researchers seek to measure the items, for example: frequency of species at particular area, food preferences of species, or related data. Ex post facto studies also include the cause-and-effect description even when they cannot control the variables. As you know, the air pollution leads to health hazards. Research can describe this cause-and-effect relationship. The methods of research utilized in descriptive research are survey methods, including comparative and correlation methods. On the other hand, in analytical research the researchers have to use facts or information already available, and analyze these to make a critical evaluation of the inquiry. In environmental science or studies generally, analytical methods are adopted.

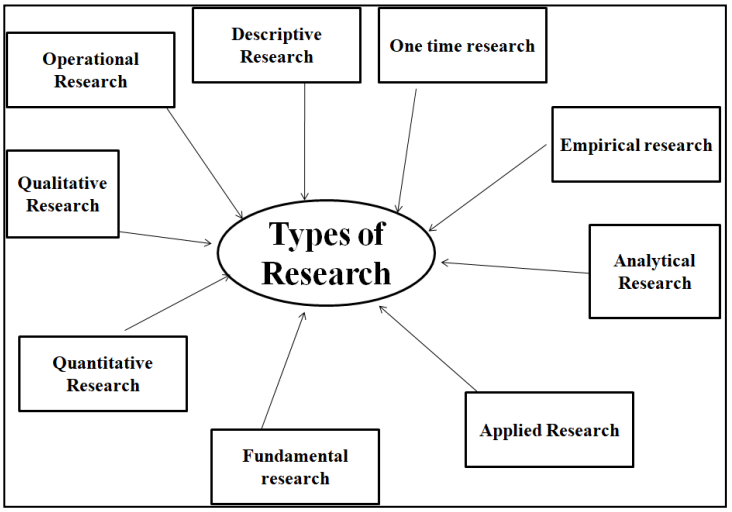


Fig.1. Different types of Research

Researchers analyze the air, water and soil quality and interpret the causes, impacts and suggest the control measure for air, water and soil pollution.

2. **Applied vs Fundamental Research:** Applied research emphasizes at findings solutions for immediate problems facing a society or an industrial/business organization, whereas fundamental research is mainly concern with generalizations of results and also formulation of a theory. Gathering knowledge for generation of generalized knowledge is termed as "fundamental research". Research concerning some natural phenomenon such as happening of earthquake, energy transfer from sun or relating pure mathematics and physics are examples of fundamental research. In the similar way, research studies, concerning human behaviours carried on with a view to make generalizations about human behaviours, are also examples of fundamental research. However, research focused at certain conclusions leading to solution for concrete social or business issues is an example of applied research. Research to identify social, economic or political trends that can affect particular institution, marketing research, evaluation research are examples of applied research such as social dynamics and exclusion processes in tribal community. Research to suggest mitigation plan for reduction of air pollution, management of forests for improved biodiversity are examples of applied research. Precisely, the main aim of applied research is to find out a solution for some pressing practical problems, whereas basic research is directed towards finding facts or information that has a board base of future applications and thus, adds to the already exiting knowledge.
3. **Quantitative Vs Qualitative Research:** Quantitative research is based on quantitative measurements and analysis of some variables of the inquiry under evaluation. It is applicable to phenomenon that may be expressed in terms of quantities. On the other hand, qualitative research is about the analysis of the qualitative feature of the variables of the inquiry under evaluation. For example, when researchers are interesting in investigations the reasons for human behaviours/animal behaviours, the research is qualitative research. We can understand the quantitative and qualitative research for evaluation of the plankton of aquatic ecosystem. When we observe the quantity (Plankton/ml of water) of

plankton, the research is regarded as quantitative research. In quantitative research we analyze the total numbers of plankton in 1 milliliter of water. On the other hand, when we analyze the diversity of plankton (chlorophyceae, bacillariophyceae, cyanophyceae, protozoa, cladocera, rotifera), the research is example of qualitative research. When we analyze the total number of plants in a forest, it is an example of quantitative research and if we analyze the colour of flower, straightness of bole, shape of seeds, the researchers are qualitative in nature and therefore examples of qualitative research.

4. **Conceptual Vs Empirical Research:** Conceptual research is generally used by philosophers or thinkers to develop new concepts or reinterpret concept which already exists. It is totally based on ideas or theory.

   On the other hand, empirical research is based on observations and data for the enquiry under consideration. In the empirical research, researchers bring the conclusions which can be verified by new observation or new experiments. Empirical research also called experimental type research such as evaluation of waste degradation under some chemical treatments, improving fertility of soil by adding some fertilizers. Researches of environmental studies generally conduct empirical research in which researchers conduct different experiments such as analysis of water quality (analysis of TDS, Dissolved oxygen, heavy metals etc.), analysis of air quality (respirable suspended particulate matter, particulate matter, ozone, CO_2), analysis of soil quality (moisture content, water holding capacity, pH, organic matter etc.).

   All other type of research falls in one or more of the above-mentioned types of approaches based on either the aim of research or the mechanism required to accomplish research, on the environment in which research is conducted. Form the point of view of time; we can think have research either as one-time research or longitudinal research. The onetime research is confined to a single time period on the other hand longitudinal research is conducted over many time periods. Depending upon the environment in which research is to be carried out research can be field setting research or laboratory research or simulation research. Research can be clinical or diagnostic research. These researches follow case

study methods or in-depth approaches to reach the basic casual relations or desired conclusions. Exploratory research is based on development of hypothesis rather than their testing. Formalized research deals with specific hypothesis to be tested. Historical research utilize historical sources like documents, remains etc. to study events or the past. Research may be conclusion–oriented in which researcher is free to pick a problem, redesign the enquiry as he proceeds and is prepare to conceptualize as he wishes. On the other hand, decision-oriented research is always for the need of a decision maker and the researcher is not free to embark upon research according to his own leaning. Operations research deals with the application of advanced analytical methods to help make better conclusion.

1.7. Approaches to Research

Research Approaches: There are two basic approaches of research viz. quantitative and qualitative approach. The quantitative approach involves the generation of data in quantitative form which may be subjected to rigorous quantitative analysis in a formal and rigid pattern. This approach may be further categorized in inferential, experimental and simulation approaches to research. The aim of inferential approach is to form a data base to draw inference about the characteristics or relationship of population. This generally means survey research where a representative sample of population is studied (questioned or observed) to determine its characteristics, and it is then inferred that the population has the same characteristics for example, evaluation of air population of a city based on survey at different locations on the city. Experimental approach is characterized by control over the research environment and in this case some variables are manipulated through some intervention to observe their effect on the response for example, evaluation of some fertilizers for increasing the fertility of the soil.

Simulation approach involves the construction of an artificial environment within which relevant facts or observations and data can be generated. This permits an observation of the dynamic behavior of a system under controlled conditions. The term simulation in the context of business or social sciences application refers to "the operation of numerical model that represents the structure of dynamic process.

Given the values of initial conditions, parameters and exogenous variables, a simulation is run to represents the behaviors of the process over time". Simulation approach can also be useful in building models for understanding future conditions.

Quantitative approach to research is referred with subjective assessment of attitudes, opinions and behaviours. Research in a situation is a function of insights and impressions of researcher. Such an approach to research gives results either in non-quantitative form or in the form which is not subjected to rigorous quantitative and analysis. Generally, the techniques of focus group interviews, projective techniques and depth interview are used in qualitative research, for example, evaluation of tree species based on farmers preference for tree planting for agroforestry in their farmland.

1.8. Process of Research

Research Process: Research does not complete in a single step; it has a series of steps or processes. Research is a process, which includes various steps such as defining or formulating the research problem, conducting review of literature, developing the hypothesis, preparation of research design, collecting the data, analysis of data and report preparation. Each process of research is being elaborated in details in the following paragraphs. The important steps of research process are summarized in Fig-2 and also given below:

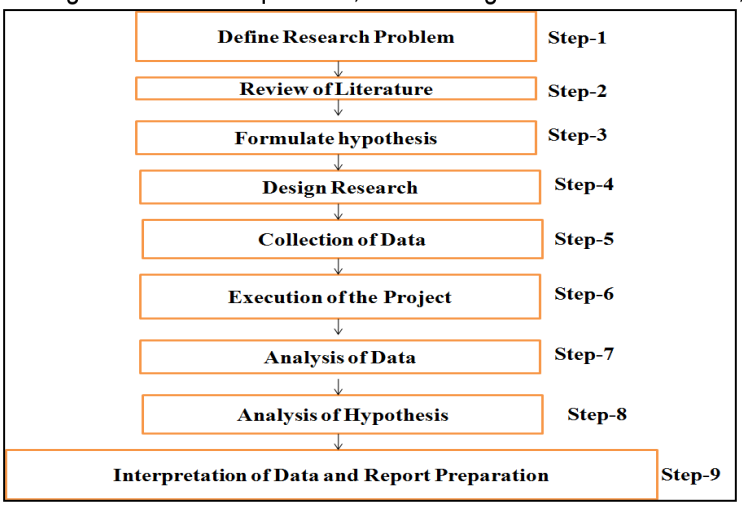


Fig-2: Different steps in the Research Process

Step-1: - Define or formulating Research Problem: This is one of the most important step of research process. When any researcher starts their research work,

he must choose the work in which he will carry out the research. In case of environmental studies, we generally seek the local problems such as air pollution, water pollution, soil pollution, impacts of anthropogenic activities on environment, biodiversity degradation etc. There must be background and rational behind selection of each research problem, which must be intended to provide answers to the posed problem. Experienced scientists may help you to formulating the research problem because they know the local, regional, national and even international problems. We should formulate the research problem in terms of understanding and keeping in view of analytical methods and available facilities required to conduct the research. Suppose we choose to research on "impact of sewage on aquatic ecology" we must assure that we have all the necessary equipment, chemicals, literature, other standard books and laboratory facilities to conduct the research. After discussing with senior researchers, mentors, guides, supervisors, we can formulate the research problem. Researcher must also carefully study the previous research work on the related problem. Each research problem must have certain objectives, which should be clear and focused.

Step-2: - Survey of Review of literature: After defining the research problem and objectives, the researcher should survey the literature so that the research should understand that what research work on the similar problem has already been completed. Review of literature provides the ideas and interpretation of the data of earlier researches on the problem. For the survey of review of literature, researcher should search authentic documents, reports and research papers from international and national journals. Conference proceedings, abstract books, government reports are also sources of literature used for collecting review of literature. In the era of internet, it now becomes very easy to collect the adequate amount of literature. Researcher should also cite the referred literature so that due credit should be attached to the previous researches.

Step-3: - Formulate Hypotheses: After the collection and study the sufficient amount of relevant literature the researcher has to set the working hypothesis or hypotheses. Hypothesis is a statement about the enquiry. Researcher can formulate the research hypothesis on the basis of experience and guidance of senior

researchers according to objective of research enquiry. Hypothesis is all about the tentative assumption. By using the above-mentioned example (Impact of sewage on aquatic ecosystem) we can formulate following two hypotheses:

i. There is no impact of sewage on aquatic ecology (Null Hypothesis)

ii. There is impact of sewage on aquatic ecology (Alternative hypothesis)

These hypotheses will provide clear idea about the significance of the research work. If our null hypothesis is rejected after testing the data it means that sewage is impacting the aquatic ecology. Rejection of null hypothesis ultimately supports that the data favours the alternative hypothesis. On the other hand, if the data supports the null hypothesis, it means that there is no impact of sewage on aquatic ecology.

Step-4: - Preparing Research Design: The research problem having been formulated in clear terms, then the researcher will be required to prepare a research design. Precisely, researcher will have to develop the conceptual structural framework within which research would be conducted. The preparation of a design guides the researcher to achieve the results of the enquiry under evaluation in precise manner. In other words, the function of research design is to provide ways for the collection of relevant facts with optimum effort, time and expenditure. But how all these can be attained depends mainly on the research purpose. Research purposes may be grouped in to following four categories:

a. Exploration
b. Description
c. Diagnosis
d. Experimentation

- Exploration of an enquiry requires a flexible research design which should be capable to provide opportunity for considering many different aspects of a problem. The purpose of an accurate description of a situation or of an association between variables, the design is focused to minimis bias and maxims the reliability of the data collected and analyzed.
- There are various research designs, such as experimental and non-experimental hypothesis testing. Experimental design can be either informal designs (such as before and after without control, after only with control, before and after with control)

or formal designs, which are suitably designed based on principles of design (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial design).

- The preparation of the research design, suitable for a particular research problem, must include the followings:
  a. The means of obtaining the data, information or facts.
  b. The availability and skills of the researchers and his staffs (if any).
  c. Explanation of the way in which selected means of obtaining information will be organized and the reasoning leading to the selection.
  d. The available of time for research i.e., time period required for the research.
  e. The cost factor relating to research i.e., the finance available for the research.

Step 5: Determining sample design: Determining sample design is crucial for unbiased result for the enquiry under evaluation. A sample design is a proper definite planning for getting the sample from any population so that data is collected from the selected units of the sample for generalization of the results about the population from which, these units were selected. The selecting units for a sample from the all units of the population for generalization of results are popularly known as sample design. Population in research is the collection of all units about which the researcher wishes to draw his conclusion. Sample is collection of some units from the population, which must be representative to the population i.e., the units of the sample must contain each characteristic that entire population possesses. In case of environmental studies, we take the sample of water from a pond to evaluate the water quality of the pond. We collect soil samples from the nearby forests to draw information about the soil property of the forests.

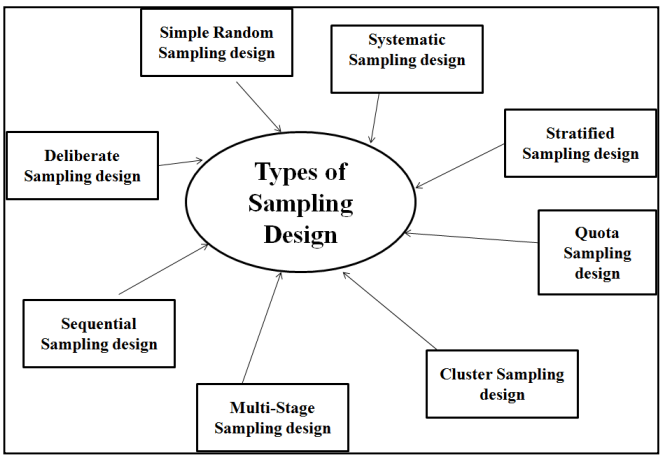


Fig-3: Showing types of sampling design

The units of the sample can be selected either through probability sampling or non-probability sampling. Probability sampling considers all units of the population to be chosen in the sample and attach some probability to each and every unit of the population to be chosen in the sample i.e., each unit has a known probability of being included in the sample. The non-probability sampling does not consider all units of the population to be included in the sample rather allow researcher to select the units as per his choice for the unit i.e., every unit of the population is not considered in non-probability sampling. There are various sample designs in probability and non-probability sampling scheme as shown in Fig-3 and discussed in following paragraphs:

1. **Simple Random sample design:** In this sampling every unit of the population has equal chance of inclusion in the sample. For example, if researchers of environmental studies wish to select 500 planktons from a population containing 2000.
2. The researcher can attach a number from 1 to 2000 to all the 2000 units (planktons) and write on small piece of paper. The 2000 number marked papers would be placed in a box and after thoroughly reshuffling, one by one 500 times, one paper would be picked and the unit bearing the number would be selected in the sample. By using such lottery method, the researcher would get 500 units, which constitute the sample. Once the selected unit i.e., paper bearing the number is again put into the box, that means, the unit have chance to be included in the sample again, and such sampling is known as with replacement method. However, if the selected paper (unit) is not returned back to box, then the unit does not have chance to be included in the sample and thus such sampling is known as without replacement method. Random sampling can also be possible by using the random number tables. To select the sample, each unit of population for above example is assigned a number from 1 - 2000, and 500 numbers from random number table is selected from the table. The number from the table may be selected by selecting a starting point and then a systematic pattern is used in proceeding through the table till the 500 numbers selected. We might start in 4th row, second column and proceed down the column to the bottom of the table and

then move to the top of next column to the right. When the number exceeds the limit of the number in a frame, in this case over 2000, it is simply passed over and next number is selected. Since the numbers were placed in table in completely random fashion, the resulting sample is random. This procedure gives each item an equal probability of being selected in the sample. In case of infinite population, the selection of each item in a random sample is controlled by the same probability and that successive selections are independent of one another.

3. **Deliberate sample design:** This method includes purposive or deliberate selection of particular units of the population for constituting a sample which represents the population. It is also known as purposive or non-probability sampling. Whenever population units are chosen in the sample based on the ease of access, it may be called convenience sampling. Judgment sampling, another non-probability sampling, considers the units of sample based on researcher judgement i.e., as per his choices. For example, if a researcher wishes to evaluate the pollution status of a city and select the samples from bus stand, at waste dumping sites, industrial areas, then such sampling is non-probability sampling. In fact, the researcher does not consider all the areas of the city for sampling rather selected the samples from polluted areas. This type of sampling may give biased results and may be used for exploratory analysis or to have some preliminary information about the population.
4. **Systematic sampling design:** This sampling scheme is useful once, information about the units of population is not known. In this scheme, first unit of sample is selected by some random mechanism and then after each unit of the sample is being selected based on a system. Suppose in case of environmental studies. we wish to estimate water quality of a river. The sample units for the enquiry may get by selecting a random point and then by a system say after every 10 km, samples may be drawn till the point, the all units of the sample is collected.
5. **Stratified sampling design:** This sampling design is used, if the population from which a sample is to be drawn does not constitute a homogenous group i.e.,

there is variability in the units of the population. In this design, the population is stratified into a number of non-overlapping subpopulation or strata so that each stratum should be as homogeneous as possible i.e., the units of the stratum is more or less similar. The units of the sample are selected from each stratum. If the item selected from each stratum is based on simple random sampling, the entire procedure, first stratification and then selection of units through simple random sampling is known as stratified random sampling. For example, if you wish to evaluate the soil from a forest patch containing vegetation in some areas and rest are open areas, then you have to divide the forest patch in two strata i.e., first stratum is part of area containing vegetation and the second stratum is open area, then after you have to draw samples from the two strata.

6. **Quota sampling design:** In this sampling method, the researcher is simply fixing the quota for different stratum, the actual selection of units for sample being left to the judgment of the researcher. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population. Quota sampling is a non-probability sampling. Quota sampling is easy to administer, fast to create and complete and inexpensive.
7. **Cluster sampling design:** It includes grouping the population based on some affinities and the selecting the groups or the cluster rather than individual for inclusion in the sample. The sampling scheme easy to operate and relatively less costly in terms of resource uses. Cluster sampling can be applicable for biodiversity analysis, natural resources monitoring. In this, the nearby areas of the forests are considered as cluster, and number of such clusters can be selected for biodiversity analysis through random mechanism, as termed as cluster random sampling.
8. **Sequential sampling design:** This sampling used where the ultimate size of the sample is not fixed in advance rather determine according to mathematical decision. This type of sampling gives the researcher boundless probability of finding his research methods.
9. **Multi stage sampling design:** This sampling is also a cluster sampling containing more than one stage for sampling. This is used for large geographical

areas such as estimation of agriculture productivity of India. In this, first, a few states would be selected out of all states of India in random mechanism, then in second stage, a few districts from each selected states would be selected, and at the third stage some villages would be selected from the selected districts, and in final stage, farm land would be selected for the survey i.e., termed as primary sampling units.

Step-6: - Collecting the data: Collection of data is an important step in research process, as based on the data, results of the study has been derived. In general, data may be categorized as primary data, in which researcher collects the data through experiments or by surveys and uses for his own researcher. The other category of the data is known as secondary data which may be obtained from the previous research papers, review articles, books, data book, statistics etc. If the researcher performs the experiments and investigations, he observes data with the help of research methodology. In case of survey, study data may be collected by the following methods:

a. **By observation:** This method implies the collection of information by investigator's experience and his own observation. The information obtained relates to what is the current situation of the area and what will be the future condition. Suppose an environmentalist survey the area where any developmental projects are in operation, he may observe the situation like deforestation, destruction of riparian zone of rivers, development of infrastructure, habitat destruction of wild animals etc. This method is generally expensive and information obtained by the investigator may be limited. This method is not suitable where large samples are desired to answer the enquiry.

b. **Through personal interview:** In this method researcher conducts the interview of the respondents as local people, farmers, socialists, politicians etc. In case of environmental studies, we should prepare the list of questions and then ask the questions to the concerned person as local people, farmer or stakeholders. These questions may include demographic information (such as age, education, family size, assets); specific question related to enquiry such as What is the benefits of this project? Which plants and animals are going to

extinct in this area? What are the causes of habitat degradation of wild animals? Which animals being poached by hunters? which plant being over exploited? etc.

c. **Through telephonic interviews:** This method is also important method to collect the data. This type of methods generally uses in industrial survey. This is time saving method of data collection. However, this method is rarely used in environmental studies.

d. **Through Email:** In this method researchers developed the questionnaire and mailed to the respondents with request to answer the question and returned back. This type of method is successful in obtaining feedback about the facilities of hotels, courses in schools, economic status of the communities etc.

e. **Through schedules:** In this method researcher provide the relevant questions to respondents with schedule.

f. **Execution of Project:** Researcher should ensure that the project is executed in systematic and timely manner. If the survey is conducted through questionnaire, then how the data can be processed. The data may be coded and digitized in computer before processing. In the case of field research, investigator may collect data from field and from lab. Once the data has been collected, it has to be digitized and processed to arrive at the conclusion. Occasional field check should also be made to ascertain validity of data. Report preparation is also an important part of execution of project and must contain details of methods, relevant survey, observations and result of data analysis and conclusion along with recommendation, if needed.

Step-7: - Analysis of data: After the collection of data, the data should be analyzed. Analysis of data needs coding, tabulation and statistical analysis. Researcher must categorize the raw data in some meaningful categories. Classified data should be placed in tabulated form in computer. The table should have clear title/caption along with appropriate title of column and row. We can use excel sheet to calculate various statistical tests to analyze the data. Analysis of data depends on the objective and requirement of research. In environmental studies and related science, there are various statistical methods, which can be applicable for data analysis such as:

mean, mode, median, standard deviation, ANOVA, correlation, regression, t-test, z-test, Chi (X^2) square test, Lotakka Voltera model, Gaussian plume model, box model, stream source point model, Shannon index, biodiversity index etc... The analysis protocol depends on the objective as well as the requirement of the method. Suppose we collected the data of water quality and we wish to express the average. For this purpose, we have to calculate the mean of water quality parameters and express it in table form. The researcher may also desire to test the hypothesis, if the objective of enquiry needs it. There are several tests such as chi-square test, t-tests, F-test have been developed by statistician to calculate the hypothesis. These tests calculate the acceptance and rejection of the hypothesis based on their test formulas, therefore called tests of significance.

Step 8: Interpretation of data and Report Preparation: Explanation of the results of research work is known as interpretation. In environmental studies we interpret the data on the basis of data analysis, theories and our findings. Such as if we find high biological oxygen demand in aquatic ecosystem, we can interpret this as: biological oxygen demand was higher in site -XYZ (6.5mg/l) which may be due to high organic load, entry of sewage etc. A high value of BOD indicates the high pollution level in the aquatic ecosystem. We should compare our study with previous studies during the interpretation of data. After interpretation of data researcher has to prepare the whole report of the research work. Report of the research work should contain the following main parts:

1. **Preliminary Pages**: It contains list of all the chapters with page number, followed by acknowledgements and certificates and termed as Content.
2. **Introduction**: Introduction should contain background, clear statement of the research problem and objectives of the research. Most of the introduction should be written in present tense.
3. **Review of literature**: It should include previous researches related to the research problem. It would be better if you give your review of literature in two parts namely: International review of literature and National review of literature. The concerned literature can be searched through the software for databases

as web of science, Scopus. This database is collection of literature from many fields, theme, topics, subjects etc.

4. **Materials and methods**: This chapter contains all the material and methods used in research work. The important parts of this chapter may categorize as study area i.e., description of latitude and longitude of study area, climate, soil etc.; parameters to be analyzed, data collection methods, details of analytical methods, principles of instruments, procedure of experiments, data analysis methods etc.
5. **Results and Discussion**: The results of the study should be presented in logical sequence and identifiable sections. The result should be based on data analysis as required. The data of research work should be crystal clear, and tabulated or represented in graphical forms, as deemed fit. The result should be described in narrative forms as per the data analysis. The result should be discussed that why such results has been occurred and supplemented by the earlier researches, if available. The discussion should be pronounced, focused and to the point.
6. **Conclusion**: Researcher should write the important findings of the research work in this section. It should be clear, precise and relevant. The section contains the findings of the research and concluding remarks about the research.
7. **References:** References i.e., the literature which has been used in the research, should be reported in the last of the report. The writing of reference should be based on standard form. There are many forms or style to write the reference. However, for a report, a uniform style of references should be adopted. The

1.9. Criteria of Good Research

As you know, research is systematic and focused study of the research problem therefore research should include the following important criteria.

1. The aim of the research should be clear.
2. Objectives of the research should be well defined.

3. Common but relevant concept should be used in research.
4. The design of the research should be planned to obtain results.
5. Description of analysis of methods should be given in details.
6. The validity and reliability of the data should be checked carefully.
7. The result should be discussed with the support of relevant literature or earlier work on the topic.
8. The research work should describe all the possible dimensions of the study so that other researchers may use this research for future use.

There are various qualities of research which are given below:

1. **Systematic:** In means that research is systematic i.e., structured with specific connected steps in specified sequence in accordance with the well-defined set of rules. Systematic research eliminates the scope of guess and intuition in arriving at conclusion.
2. **Logical:** Good research is always logical. This implies that research is controlled by the rules of logical reasoning. The conclusion of the research is derived by the logical process of induction and deduction reasoning. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from some premise to a conclusion which follows from that very premise. In fact, logical reasoning makes research more meaningful in context of decision making.
3. **Empirical:** Good research is empirical. It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.
4. **Replicable:** Replicability is important quality of research, which ensures that the research is repeatable. This quality allows research result to be verified by replicating the study and thereby building a sound basis for decision.

Problems related to research in India: There are various research related problems in India specially those engaged in empirical research. In India, there is lack of proper scientific training for research specially training for selecting and adopting appropriate research methodology for the research. The logical thinking is

also not inculcated in the academics, is another serious concerned. Various workshops, seminars, conferences are organized by different institutions but very few of them are dealing the issues of the research methodology. Moreover, a combined research methodology workshop catering to many fields of research is also not very appropriate, as the participants lose their interest. It is also found the there is insufficient interaction between the university and government departments therefore the academicians are not well aware to the actual problems to be researched. In India there is lack of code of conduct for researchers and inter-university and inter department rivalries are also quite common. There is also lack of basic facilities in universities and colleges to conduct the advanced research in terms of resources such as well equip laboratories, lack of sophisticated equipment etc. Many researchers in India also face financial crises to conduct their studies. Many universities of India don't have good libraries and printing facilities. The lack of quality material in library such as quality books, recent papers, digital database, is also a problem for demotivating the researchers. Therefore, it is very important to enhance the basic facilities of research, quality training for conducting research, interactive sessions on research methodology and provide great environment to the researchers to promote the research in the country.

1.10. Summary

In this unit we have discussed various aspects of research and research methodology. So far you have learnt that:

- The word "research" has been taken from French word "recherche" which means "to go about seeking". It comprises the creative and systematic work taken to increase the knowledge of human being, culture and society.
- Research comprises defining and redefining problems, formulating hypothesis or suggesting solutions, collecting data, organizing and evaluating data, making deduction and reaching conclusions.
- The concepts or symbols for the purpose of generalizing to extend, correct or verify facts, whether that knowledge aids in construction of theory or in the practice of an art is known as research.

- There are two terms related to research methodology first is Research techniques and second is research methods. Research techniques involves the behaviors and instruments we use in performing research operations such as making observation, recording data etc. research methods refer to the behavior and instruments used in selecting and constructing research technique.
- There are various roles of research in environmental science such as in Biodiversity Assessment and Management, in Water quality analysis, in Air quality analysis, in Soil Quality analysis, in Policies making, Conservation of Natural Resources, in Waste Management etc.
- The general objectives of research are to gain familiarity with a phenomenon or to achieve new insights into it (studies with this object in view are termed as exploratory or formulate research studies, to portray accurately the characteristics of a particular individual, situation or a group, to determine the frequency with which something occurs or with which it is associated with something else and to test a hypothesis of a causal relationship between variable.
- There are various types of research such as Descriptive, Analytical, Applied, Fundamental, Quantitative, Qualitative, Conceptual, Empirical, One-time research, Exploratory research, Historical research, Operations research etc.
- There are two basic approaches of research viz. quantitative and qualitative approach. The quantitative approach involves the generation of data in quantitative form which may be subjected to rigorous quantitative analysis in a formal and rigid fashion. This approach can be further classified in inferential, experimental and simulation approaches to research. The purpose of inferential approach is to form a data base to infer characteristics or relationship of population. This generally means survey research where a sample of population is studied (questioned or observed) to determine its characteristics, and it is then inferred that the population has the same characteristics.
- Research is a process comprises of various steps such as Defining or formulating the research problem, review of literature, developing the hypothesis, preparation of research design, collecting the data, analysis of data and report preparation.

- There are various sample design methods such as Simple Random sample design, Deliberate sample design, Systematic sampling design, Stratified sampling design, Quota sampling design, Cluster sampling design, Sequential sampling design and Multi stage sampling design.
- Criteria of good research includes many as : the aim of the research should be clear, common concept should be applied in research, the research work should describe all the possible dimension of the study so that other researchers may use present research for future use, the design of the research should be planned and focused to obtain results, objectives of the research should be well defined, the validity and reliability of the data should be checked carefully and description of analysis of methods should be elaborated in details.

TERMINAL QUESTIONS

1. Fill in the blank spaces with appropriate words.

Explanation of the results of research work is called…….. In environmental studies we interpret the data on the basis of theories and our findings. Such as if we find high biological oxygen demand in aquatic ecosystem, we can interpret this parameter like as: biological oxygen demand was higher in site -XYZ (6.5mg/l) which may be due to high organic load, entry of sewage etc. A high value of BOD indicates the high ……. level in the …….. ecosystem. We should compare our study with previous studies during the ……of data. After …… of data researcher has to prepare the whole …….. of the research work.

2. (a) Define the research.

(b) Explain, what is research methodology?

3. (a) Describe the role of research in Environmental studies.

(b) Give the Process of Research in detail.

4. Define Sample design? Write a note about the various types of sampling design.

5. (a) Discuss Approaches to Research in details.

(b) Differentiate between qualitative and quantitative research?

(c) Explain, how you will collect the data to estimate the soil property across the Botanical Graden?

6. (a) Fill the blank spaces with appropriate words.

After defining the…….. ……., researcher should survey the adequate amount of review of literature. A brief summary of the research problem should be written which helps the researcher to survey the related……. Review of literature provides the ……and interpretation of the data. For the survey of……. … ……., researcher should search authentic …… ……..from international and national journals. Conference proceedings, abstract books, government reports also good sources of colleting review of literature. In the era of internet it now becomes very easy to ……..the adequate amount of literature. Researcher should also make sure that the ISBN number, ISSN number, name of journals, volume number, issue number and impact factor of the journal by which review of literature being collected.

(b) If researcher conducted experiments in research work to arrive at conclusions, this type of research is called as (Empirical research/Descriptive research)

(c) Primary data is obtained from **(**Personal research/review articles/thesis/reference books**)**

(d) From which language the word "research" has been taken? (Latin/French/Greek/Sanskrit)

7. What are the qualities of research?

(a) Discuss the criteria of good research.

(b) Describe the various steps for report preparation in Research process.

(c) Explain, what do you mean by objectives of research**?**

ANSWERS

1. (a) Interpretation, pollution, aquatic, interpretation, interpretation, report

2. (a) see section 1.2.

(b) See section 1.3 under heading meaning of research methodology

3. (a) See section 1.4.

(b) See section 1.8.

4. (a) See section 1.8 under heading determining sample design.

5. (a) See the section 1.7

(b) see the section 1.8 under heading qualitative and quantitative research.

(c) See section 1.8 under heading collecting the data in step-5

6. (a) research problem, literature, ideas, review of literature, research papers, collect

(b) Empirical research

(c) Personal research

(d) French

7. See the qualities of research in section 1.9.

(a) See the section 1.9

(b) See the section 1.8 under heading interpretation of data and report preparation.

(c) see the section 1.5.

Unit 2: Research Problem: Definition; Selection and Techniques of defining a problem

Unit Structure

2.0. Learning Objectives

After studying this unit, you will be able to:

a. What is research problem?

b. Meaning of research problem.

c. Current environmental research problems in India

d. How we can select or choose research problem?

e. What are the criteria and techniques of defining research problem?

2.1. Introduction

As you know that research comprises defining and redefining problems, formulating hypothesis or suggested solutions, collecting, organizing and evaluating data, making deduction and reaching conclusions. Researcher must select or identify and choose the appropriate research to start the research. As you have learnt in unit-1 (Research Methodology) of this course that selecting and defining the research problem is first and foremost step in research process. It can be understood by the example. When we approach to doctor for treatment of a health problem i.e., disease, the doctor, in the very first step, tries to identify the problem of patient and this process is known as diagnosis. After the diagnosis, doctor can prescribe appropriate medicines to cure or treat the particular disease. In the

same way, when researchers/scientists do the research, the first step is identification of the research problem as precisely as possible. After the defining research problem, researcher may collect the required data and information to perform the appropriate tests and data analysis to conclude that how this problem is originated and how we can mitigate that problem. Researcher always seeks the solution to particular problem. There are various problems according to subject, event, process, function and place. In case of environmental studies, there are many research problems specific to environment such as pollution, depletion of natural resources, land degradation, global warming, climate change, waste disposal etc. You must aware about the air pollution, moreover, on annual basis, World Health Organization (WHO) release the list of most polluted cities, most polluted countries etc. The question is that how WHO released the list of polluted cities or countries? Answer is simple, WHO collect the data from countries and processes the data for doing so i.e., WHO conduct research to release the list of cities and countries. However, most importantly, they identify the problem of pollution. Without defining a research problem, the research cannot be conducted. Therefore, defining the problem is first and foremost step in research process. In this unit, you will learn about the research problem, current environmental research problems in India, selection of research problem and criteria and techniques involved in defining research problem.

2.2. What is research problem?

In general, a research problem refers to some difficulty or problem which a researcher/scientist experiences in the context of either a theoretical or practical situation and wishes to obtain remedial measures for that particular or specific problem. For example, suppose in case of environmental science, we experience that air borne diseases are increasing day by day in Delhi due to vehicular emission, industrialization, urbanization, deforestation etc. On the basis of this, we can define the research problems like as:

- Impacts of vehicular emission on air quality within Delhi City
- Impacts of deforestation on air quality within Delhi City
- Impacts of industrialization on air quality within Delhi city.
- Air borne diseases in the population of Delhi

- What mechanisms can be adopted to reduce the pollution in Delhi city

In the same way, we can identify the other environmental problems such as water pollution, noise pollution, soil pollution, bio-diversity degradation, global warming, ozone layer depletion, acid rain, green house effects etc. at particular place. Thus, we can classify the components of a research problem as under:

i. There must be an individual or a group which has some difficulty or the problem.

ii. The problems must have objectives to be attained at. Without objectives, there is no meaning of research problem.

iii. There must be alternative means for obtaining the objectives. This means there must be at least two means available to a researcher and if he has no choice of means, he cannot have a research problem.

Therefore, the research problem is one which requires a researcher to find out the best solution for the given problem i.e., to find out by which way of action the objective can be attained optimally in the context of a given environment. There are various factors which may result in making the problem complicated. For example, the environment may change affecting the efficiencies of the courses of action or the values of the outcomes, the number of alternative courses of action may be very large, person not involved in making the decision may be affected by it and react to it favorably or unfavorably, and similar other factors. All such elements, which can directly or indirectly influence the objectives, may be considered in context of a research problem.

2.3. Meaning and definitions of Research Problem

As you know that we cannot start the research without having research problem. There may be two types of research problem. The first research problem is related to state of nature and second problem is dealing to relationship between variables. The researcher must be clear to the problem he or she wishes to study. This means that the researcher must decide the general area of interest or aspect of a subject that he would wish to analyze for drawing the conclusion. At the start, the research problem may be stated in a broad general way and then the ambiguities, if any, relating to problem must be resolved to make the problem more

focused and pronounced. The broad area in environmental sciences may include life science, chemical sciences, physical sciences, computer sciences, political sciences etc.

Definitions of Research Problem:

- It is a statement about an area of concern, a difficulty to be removed or in practice that point need for meaningful understanding for deliberate investigation.
- A research problem is a question, which a researcher wishes to answer
- A research problem is a problem that a researcher wishes to solve.
- According to Kerlinger "A problem is an interrogative sentence or statement that asks what relation exists between two or more variable".
- According to R.S. Woodworth "A situation for which we have no ready and successful response by instinct or by previous required habit".
- In other words, research problem may be defined as "It is an area of concern where there is a gap in the knowledge base needed for professional practices".

2.4. Sources of research problem

There are various sources of research problem which are given below:

a. Personal Experiences and previous research
b. Practical Experience
c. Existing theories
d. Social issues
e. Brainstorming
f. Environmental issues
g. Feedback from community
h. Local problems
i. Discussion with mentor guide or supervisor
j. Personal observation

2.5. Current Environmental research problems in India

In India, numerous research problems related to environment are existed and Researcher may choose one of these environmental problems for his/her research. However, the research problems vary according to place, time and need. Therefore, it would be better, if researcher select the research problem, which has potential uses to the society. Some of the important environmental research problems of India are summarized in Fig.1. and also, being detailed in subsequent paragraphs. However, these are not the comprehensive list of problems in the environmental aspects.

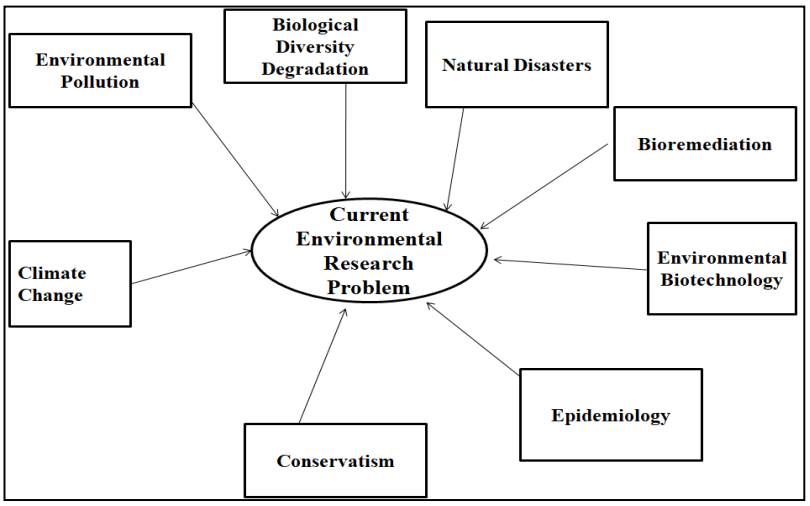


Fig.1: Current research problems related to environmental aspects in India

1. **Environmental Pollution:** As you know that environmental pollution is one of the biggest problems of the present time. Many researchers conduct the research on environmental pollution such as air, water, soil, noise etc. Researcher may identify the environmental pollution and choose the appropriate research problem for their research. For example, the researcher may wish to know the level of air and water pollution; to know the causes of the pollution, to suggest remedial measures for reduction of pollution.
2. **Biodiversity degradation:** Biodiversity degradation is also an important issue at the global level. Many researchers doing the research on the issue. Habitat fragmentation, environmental pollution, poaching, over exploitation, introduction of invasive species etc. are main major concerned for biodiversity researches.

Researcher may seek the research problem on causes of biodiversity loss, cause of over exploitation, cause of forest degradation. Researches may also evaluate the fragmentation and species shift.

3. **Natural disasters:** As you know that earthquake, flood, volcanoes, cyclones, landslides are the forms of natural disasters. Many scientists and researcher do researches on such natural disasters. The impacts of natural disasters are immeasurable and countless. Researcher may wish to conduct researches on the issues such as evaluation of impacts of natural disasters; causes of natural disasters; vulnerability of community due to disasters; adaptation mechanism to counter or cope the impacts of the disasters.
4. **Bioremediation:** Bioremediation is a new technique in which biological components specially bacteria and fungi used to clean the environment. Research problems related to bioremediation has great scope in present time such as role of bacteria for treatment of water pollution.
5. **Environmental biotechnology:** Various studies have already conducted on environmental biotechnology in which scientists manipulate the genes of certain bacteria and other microbes to produce new variety of organisms. Genetic modified crops are results of biotechnology. Therefore, researchers of environmental science may seek such type of research problem for their study such as disease resistant variety of pea, rice etc.
6. **Epidemiology:** As you know that certain bacteria, viruses, protozoans, fungi etc. are responsible for various diseases in human being. Typhoid, Cholera, Malaria, dysentery, African sleeping sickness, kala azar etc. are common epidemic diseases. The studies on these diseases are need of hour and therefore, studies on these aspects may be chosen for research problem.

2.6. Selection of Research Problem

The research problem undertaken for study must be carefully selected though, the task of selecting a problem is difficult. For selecting research problem, researcher may discuss or seek advice from his mentor, supervisor or peers. Nevertheless, every scientist/researcher

must find out his own salvation for research problem cannot be rented. A problem must spring from the mind of researcher like a plant springing from its own seed. As you know if our eyes need glasses, it is not the optician alone who decide about the power of the lens, we require. We have to see ourselves and enable him to prescribe for us the right number (power) by cooperating with him. Therefore, a researcher guide can at the most only assist a researcher choose a subject or problem. The following points may be analyzed by a researcher in selecting a research problem or a subject for research.

1. Subject which has been researched thoroughly should not be normally chosen, as it will be difficult to deduce new knowledge on the subject.
2. Controversial topic, theme and subject should generally be avoided by an average researcher.
3. Too narrow or too broad research problem should be avoided.
4. The subject selected for research should be familiar and practicable so that the related research material or sources of research are available to researcher for his research. Even then it is quite difficult to supply definitive ideas concerning how a researcher should obtain ideas for his research. For this purpose, a researcher should contact an expert or a professor in the university who is already engaged in related research. He may as well-read quality articles published in current literature available on the subject. He may enlighten him that how the techniques and ideas discussed in literature might be applied to the solution of other problems. He may discuss with renowned scientists about his concerns related to the research problem. In this way, the researcher should make all possible efforts in selecting a research problem.
5. The importance of the subject/theme/topic, the qualifications and training of a researcher, the costs included, the time factor and few other criteria such as availability of laboratory, equipment that must also be included in a selecting a problem. In other words, before the final selection of a problem, a research must ask himself the following questions
    i. Whether he is well equipped in terms of his background to carry out the particular research?

ii. Whether the study falls within the budget he can afford?

iii. Whether the necessary cooperation can be obtained from those who must participate in research as subject?

iv. Whether necessary laboratory facilities and equipment are available?

If the answers to all these questions are in the confirmatory, one may become sure for conducting the research so far as the practicability of the study is concerned.

6. The selection of a research problem must be preceded by a preliminary study. This may not be necessary if a research closely similar to the proposed research has already been conducted. But when the field of inquiry is relatively new and does not have a set of well-defined techniques, a brief feasibility study would be better to undertake.

If the subject for research is selected properly by observing the above-mentioned points, the research will not be a unnecessary tedious, haphazard, rather it will be focused and matching with the resources. In facts, besides these above, enthusiasm for research work is a must. The subject or the research problem selected must be highly prioritized by researcher so that he may undertake all efforts needed for the research work.

2.7. Criteria and techniques of defining a Research problem

There are various criteria and techniques of defining research problem which are discussed below:

Necessity of Defining Problem: It is stated that a research problem clearly stated is a research problem half solved. This statement signifies the need for defining a research problem in very clear and precise manner. The problem to be monitored must be defined unambiguously so that it will help to discriminate relevant data from the irrelevant ones. A proper definition of research problem will enable the researcher to be on the track whereas an ill-defined problem may lead to confusion. Questions like: what type of data is to be collected? What characteristics of data are relevant and need to be monitored? What relations are to be explored? What techniques are to be used for the purpose? and similar other questions crop up in the mind of the researcher. The researcher can well plan the strategy and find answers to all such questions only when the research problem has been

defined in clear manner. Therefore, defining a research problem properly is prerequisite for any research and is a most importance step of research. In facts, formulation of a research problem is often mandatory than its solution. It is only on the careful detailing of the research problem that researcher can work out the research design and can smoothly carry on all the important steps of research while conducting research.

Technique involved in Defining a Research Problem: There are various techniques, which are discussed in this chapter. Let us start with the question: what does one mean when researcher wants to define a research problem? The answer may be that one wants to state the problem along with the bounds within which it is to be studied. In other words, defining a research problem includes the task of laying down boundaries (domain) within which a researcher will study the problem with a pre-determined objective in consideration. How to define a research problem is obviously a phenomenal task. However, it is a task that must be tackled intelligently to avoid confound encountered in a research process. The general approach is that the researcher should himself pose a question (or in case someone else who wants the researcher to carry on research, the concerned individual, organization or an authority should pose the question to the researcher) and setup techniques and procedure for answering the question concerned for formulating or defining the research problem. But such as approach generally does not produce definitive results because the question phrased in such a manner is typically in broad general terms and as such may not be in a form suitable for testing.

Defining a research problem properly a crucial part of a research study and must not be accomplished quickly. However, in practice this is frequently overlooked which cause a lot of problems during the research. Hence, the research problem should be defined in a systematic and appropriate manner, giving due weightage to all concerns of the research.

Undertaking of the following steps generally one after the other may support to qualify a better research problem:

1. Writing a statement of the research problem in a general method.
2. Understating the nature of a research problem.
3. Surveying the available review of literature.

4. Developing the ideas through discussion.
5. Rephrasing the research problem in to a working proposition

There are various important techniques of defining research problem which are summarized in Fig-2 and also described below:

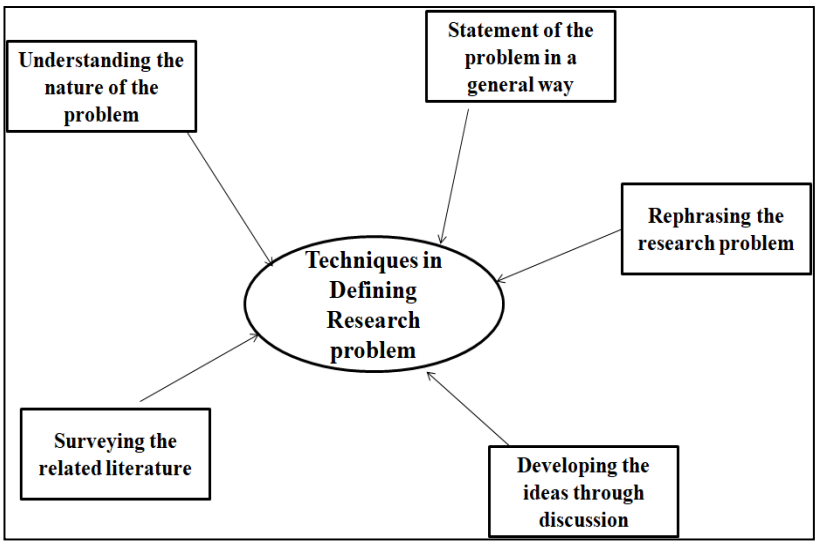


Fig-2: Different steps in defining the research problem

1. **Statement of the problem in a general way:** In defining research problem, the first of the entire problem should be stated in a broad general way, keeping in view either some practical concern or some scientific or intellectual interest. For this purpose, the researcher must immerse himself thoroughly in the subject concerning which he wishes to pose a research problem. In case of social research, it is considered advisable to do some field analysis and as such the researcher may undertake some sort of preliminary filed survey or what is often called **pilot survey**. Then the researcher can himself state the problem or researcher can seek the guidance of the supervisor or the subject expert in accomplishing this research problem. Often, the guide puts forth the problem in general terms, and it is then, the researcher, who narrow down and phrase of the problem in operational terms. In case there is some directive from an organizational authority, the problem, then can be stated accordingly. The research problem stated in a broad general way may contain various ambiguities which must be resolved by focused and directed thinking and rethinking over the problem. At the same time the possibility of a particular solution

has to be considered and the same should be kept in view while stating the research problem.

2. **Understanding the nature of the problem:** The next step in defining the research problem is to understand its origin and nature in explicit manner. The best way of understanding the research problem is to discuss the problem with those who first raised the problem in order to find out the origin and cause of problem along with the objectives in view. If researcher has stated the problem himself, he should consider once again all those points that induced him to make a general statement concerning the research problem. For better understanding of the nature of the research problem involved, he can discuss with those who have a good knowledge and understanding of the concerned or similar research problems. The researcher should also keep in view the environment within which the problem is to be studied and analyzed.

3. **Surveying the related literature:** All available literature related to research problem at hand must necessarily be surveyed and examined before a definition of the research problem is achieved. It means that researcher must be well-conversant with applicable theories in the field, reports and records and also all other appropriate literature. He must devote adequate time and energy in reviewing of researches already undertaken on the related research problem. This is undertaken to find out what data and materials, if any is available for operational purpose. "Knowing what data is available frequently provides to narrow the problem itself as well as technique that might be used", this would also help a researcher to know if there are certain gaps in the theories or whether the existing theories valid to the problem under study or conflicting with each other, or whether the findings of the different studies do not follow a pattern consistent with the theory. This will enable a researcher to consider new strides in the field for furtherance of knowledge i.e. researcher can move up starting from the existing idea. Studies on related problems are useful for indicating the type of difficulties that may be encountered in the present study as also the possible analytical shortcomings. At times such studies

may also suggest useful and even new lines of approach to the present research problem.

4. **Developing the ideas through discussion:** Discussion concerning a problem often produces useful information. Various new ideas can be developed through discussion. Hence, a researcher must discuss his problem with his colleagues and other scientists who have enough experience in the similar areas of research or in working on similar research problem. This is pretty known as experience survey. People with vast experience are in a position to explain to the researcher on different aspects of his proposed study and their advice and comments are usually precious to the researcher. The discussions assist the researcher to sharpen his focus of concentration on specific aspects within the field. Discussion with experienced person should not only be confined to the formulation of the specific problem at hand, but should also be concerned with the general approach to the given problem, technique that might be used to obtain the possible solutions etc.
5. **Rephrasing the research problem:** In the last, the researcher should rework to rephrase the research problem into working proposition. Once the nature of problem has clearly been understood, the environment (within which the problem has got to be studied) has been defined, discussions over the problem have taken place and the available literature has been surveyed and examined, rephrasing the problem into analytical or operational terms is not a difficult task. Through rephrasing, the researcher puts the research problem in as specific terms as possible so that it may become operationally feasible and may help in the development of working hypothesis.

In addition to above mentioned points, the following points must also be observed while defining a research problem:

1. Technical terms and words or phrases, with special meanings used in the statement of the research problem, should be defined explicitly.
2. Basic assumptions or postulates (if any) relating to the research problem should be clearly mentioned.

3. A straight forward statement of the value of analysis (i.e., criteria for the selection of the research problem) should be provided.
4. The suitability of the time period and the sources of data available must also be considered by the researcher in defining the research problem.
5. The scope of the investigation or the limits within which the problem is to be studied must be noted or mentioned explicitly in defining a research problem.

An Illustration: The techniques of a defining a problem outlined above can be illustrated for better understanding by taking an example as under:

Let us suppose that a research problem in a broad general way is as follows:
'Increasing air borne diseases on population of Delhi'
In this form the question has a number of ambiguities such as:

- What sort of diseases is being referred to?
- What are the causes of air pollution in Delhi?
- Which types of pollutants are high in concentrations in atmosphere at Delhi?
- What is relationship between the disease and the pollutants?

In view of all such ambiguities, the given statement or the question is too general to be amenable to analysis.
Rethinking and discussions about the problem may result in narrowing down the question to:

- What factors are responsible for bad air quality at Delhi?

This latter version of the problem is definitely an improvement over the earlier version as various ambiguities have been removed to the extent possible. Further rethinking and rephrasing might take place to the problem on a still better operational basis as shown below:

- What was the condition of air quality before 20-30 years ago?
- What changes are responsible for air borne diseases at Delhi?

With this short of formulation, the various terms involve such as "Air borne diseases" "Responsible factors" "Sources of air quality degradation" must be explained explicitly. The researcher must also consider whether the necessary data is available on the related

research problem. Precisely, all relevant factors must be considered by a researcher before finally defining a research problem.

We may conclude by saying that the task of defining a research problem, very often and follows a sequential pattern-the problem is stated in general way, the ambiguities are resolved, thinking and rethinking process results in a more specific formulation of the problem so that it may be realistic one in term of available data and resources and is also analytically meaningful. All this results in a well-defined research problem that is not only meaningful from an operational point of view but is equally facilitate for formulating of working hypotheses and for means of solving the problem.

2.8. Summary

In this unit we have discussed meaning, definitions of research problem and different environmental research problem. So far you have learnt that:

- Research comprises defining and redefining problems, formulating hypothesis or suggested solutions, collecting, organizing and evaluating data, making deduction and reaching conclusions and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis.
- In general, a research problem refers to some difficulties or problems which a researchers/scientist's experience in the context of either a theoretical or practical situation and wishes to obtain remedial measures for that particular or specific research problem.
- There may be two types of research problem, first problem is which related to state of nature and second problem is related to relationship between variables. The researcher must single out the problem he wishes to study. It means that the researcher must decide the general area of interest or aspect of a subject matter that he would like to analyze.
- At the start, the research problem may be stated in a broad general way and then the ambiguities, if any, relating to problem should be resolved. The broad area in environmental sciences includes life science, chemical sciences, physical sciences, computer sciences, political sciences etc.

- In India, there are various environmental research problems and researcher may choose one of these problems for his research. However, the research problems vary according to place, time and need.
- Some of the important environmental research problems of India are: Environmental Pollution, Biodiversity degradation, Natural disasters, Bioremediation, Environmental biotechnology, Epidemiology etc.
- The research problem undertaken for study must be carefully selected though the task is difficult. For selecting research problem, researcher may take assistance from guide. Nevertheless, every scientist/researcher must find out his own salvation for research problem cannot be rented. A problem must spring from the mind of researcher like a plant springing from its own seed.
- A research problem or a subject for research which has already been sufficiently researched should not be normally chosen. It would be difficult to add new knowledge to the research problem. Controversial topic, theme and subject should not become the choice of an average researcher. Moreover, too narrow or too broad research problem should be avoided.
- There are various criteria and techniques of defining research problem which are as: Writing statement of the research problem in a general method, Understating the nature of a research problem, Surveying the available review of literature, Developing the ideas through discussion, Rephrasing the research problem in to a working proposition.
- The points must also be observed while defining a research problem are : technical terms with special meanings used in the statement of the research problem, basic assumptions relating to the research problem should be clearly mentioned, straight forward statement of the value of analysis should be provided, the suitability of the time period and the sources of data available must also be considered by the researcher in defining the research problem and the scope of the investigation or the limits within which the problem is to be studied must be noted or mentioned explicitly in defining a research problem.

Terminal Questions

1. (a) Fill in the blank spaces with appropriate words.

The research problem undertaken for study must be carefully ……. or ……... The task is a difficult one, however, it may not seen be so. For selecting …….. ……., researcher may take help from ……. in this connection. Nevertheless, every scientist/researcher must find out his own salvation for research problem cannot be rented. A …….. must spring from the mind of researcher like a plant springing from its own ……... As you now if our eyes need glasses, it is not the optician alone who decide about the number of the lens we require. We have to see ourselves and enable him to prescribe for us the right number by cooperating with him. Therefore, a researcher guide can at the most only help a researcher choose a …… ……..

2. (a) Define the research problem.

(b) What do you understand by research problem? Explain with example.

3. (a) Describe current research problems related to environmental studies in India.

(b) How will you select a research problem?

4. (a) What are the criteria of defining research problem.

5. (a) Discuss the techniques involved in defining research problem.

6. (a) Fill the blank spaces with appropriate words.

All available ……..related to …….problem at hand must be necessarily be ……..and examined before a definition of the research problem is given. It means that …….must be well-conversant with applicable theories in the field, reports and records and also all other appropriate…… He must devote adequate time in reviewing of research already undertaken on related research problem. This is done to find out what data and other material, if any is available for …….purpose. “knowing what data is available frequently provides to narrow the problem itself as well as technique that might be used”, this would also help a researcher to know if there are certain gaps in the theories or whether the existing theories valid to the problem under study are conflicting with each other, or whether the findings of the different studies do not follow a pattern consistent with the theoretical expectations and so on. This will enable a researcher to take new strides in the field for

furtherance of knowledge i.e. researcher can move up starting from the…….. ……. Studies on related problems are useful for indicating the type of difficulties that may be encountered in the present study as also the possible analytical shortcomings.

(b) Defining the research problem is first and foremost step in research process

(Yes/No**)**

(c) Can we start research work without defining research problem? (Yes/No)

(d) Which research problem is related to environmental studies in India?

(Environmental Pollution/Soil erosion/Natural disasters/all of the above)

(e) What do you understand by bio-remediation?

7. (a) What are the sources of defining research problem?

ANSWERS

1 (a) Selected, identified, research, problem, guide, problem, seed, research problem**.**

2 (a) see section 2.3.

(b) See section 2.2

3 (a) See section 2.5

(b) See section 2.6

4 (a) See section 2.7

5 (a) See the section 2.7. including fig-2

6 (a) Literature, research, surveyed, researcher, literature, operational, existing idea

(b) Yes

(c) No

(d) All of the above

(e) See the section 2.4 under heading bioremediation

7 (a) See the section 2.4

Unit 3: Research Design: Meaning, Needs and Features of Good Design; Important Concepts Related to Research Design; Different Research Design; Principles of Experimental Design and Important Experimental Design

Unit Structure

3.0. Learning Objectives

After studying this unit, you will be able to know the following:

- What is research design?
- Meaning of research design
- Why research designs are needed?
- What are the concepts related to research design?
- Different types of research designs
- What are principles of experimental designs?
- What are important experimental designs?

3.1. Introduction

For good quality of research, we have to choose appropriate research design. After defining a research problem, researcher should have to choose an appropriate research

design to complete the research work. As you have learnt in Unit-2 (Research problem) defining a research problem is first and foremost step in research process, after the defining research problem investigator needs to select appropriate and suitable research design to perform the study to achieve the research objectives. Research design is the procedure which helps in collecting, monitoring, investigating and analyzing the data specified in research. In very simple words we can stated that a research design is framework that is used to find the answer to research problem. As you know researchers of environmental studies do the research on different aspects such as environmental pollution, ozone layer depletion, global warming, greenhouse effect, acid rain, biodiversity degradation, natural resources etc. They need specific research design for specific purpose. Suppose any researcher conducting research on water quality of river and he has defined the research problem then for answering the question under consideration, he should choose appropriate research design. In water quality research he needs research design which may include standard methods of water quality such as APHA, Trivedi and Goel (FULL REFFXXX), sampling techniques, procedure of measuring certain water quality parameters (dissolved oxygen, biological oxygen demand, chlorides, nitrates, phosphates etc.). On the other hands if researcher wants study on air quality, the research design would be different and depend on the objective and the characteristics of the experimental material. In this unit you will learn about meaning, definitions, need and type of research design.

3.2. Meaning and definitions of research design

The formidable problem that follows the task after defining the research problem is the preparation of a design of the research project to answer the research question, popularly known as “research design”. A research design is the arrangement of approaches for collection and analysis of data in a fashion that aims to combine significance to the research purpose with economy in procedure. In fact, the research design is the conceptual structure within which research is conducted; it contains the blue print for collection, measurement and analysis of the data. As such a design comprises an outline of what the researcher will do from writing the hypothesis and its operational implications

to the final analysis of data. Precisely, the research design decisions should consider the following:

- What is the research about?
- Why is the research being conducted?
- Where will the research be carried out?
- What type of data is necessary?
- Where can be required data be collected?
- What period of time will the study include?
- What will be the sample design?
- What techniques of data collection will be applied?
- How will the data be analyzed and monitored?
- In what style will the report be prepared?

Keeping the view, the above stated research design decisions; one may divide the overall research design into the following category:

a. **The sampling design:** It deals with the method of selecting units to be observed for the given study.
b. **The observational design:** It relates to the conditions under which the observations are to be made about the units.
c. **The statistical design**: It concerns the number of items is to be observed and the way information and data gathered from the units is to be analyzed**.**
d. **The operational design:** It deals with the techniques by which the procedures specified in the sampling, statistical and observational design can be carried out.

The important features of a research design are reported below:

1. It is a plan that specifies the sources and types of information relevant to the research problem.
2. It is strategy specifying the approach for collecting and analyzing the data.
3. It also comprises the time and cost to be incurred during the process of research.

In brief, research design must, at least, contains

- A clear statement of the research problem
- Procedures, methods and techniques to be used for obtaining the data and facts.
- The population to be studied
- The variable and their measurement technique
- Methods to be used in processing and analyzing data.

Definitions of Research Design:

- According to Burns and Grove a research design may be defined as "A blue print for conducting a research study maximum control over variables which could interfere with the validity of findings"
- According to Parahoo "A plan which explains, how when and where data are to be collected and analyzed is called research design"
- According to Sellitz (1962). "A research design is the arrangement of conditions for collecting and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure".

3.3. Need for research design

Research design is required as it guide the smooth sailing of the various research processes with minimal expenditure of effort, time and money, thereby making research as efficient as possible obtaining maximal facts or information. Researcher needs a research design or a plan in advance for data collection and analysis of the research problem. It is similar to a blue print (generally known as map/model of the house) prepared by an expert architect for better, economical and attractive construction of a building/house, Research design stands for advance planning of the methods to be adopted for collecting the data and the techniques to be used for data analysis in the researches, keeping in view of the research objectives and the availability of staff, time and money. The research design should be prepared carefully because any error crept in research design may distort the significance of the research. Research design in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the whole structure of the research work.

The need for a well thought research design is at times not recognized by many researchers, as researcher does not recognize the significance attach to the designing the research. As a result, the researchers do not get the reliable results to serve the purpose of the research they are considered to achieve the research problem. In fact, under the circumstances, the research may mislead conclusion of the research. Designing the research project without adequate consideration of various aspects of the research may occur in rendering the research exercise useless and unproductive. Therefore, it is imperative that an efficient and suitable design must be managed and considered before starting research. The design assists the researcher to organize the research processes in a form whereby it will be possible for him to look for errors and insufficiencies during the research. It would aid further, if research design is vetted critically by other experts. Moreover, in the absence of an appropriate design, it will be difficult for the expert to provide a complete review of the study.

3.4. Features of Good Research Design

A good design is generally characterized by flexibility, appropriateness, efficient, economical etc. Generally, a good design should minimize bias and maximizes the reliability of the data collected and analyzed. The design gives smallest experimental error is supposed to be a good research design. Design which yields maximal facts and provides an opportunity for considering many different aspects of a problem is considered appropriate and efficient design in respect of many research problems. Therefore, the good research design must consider the purpose or objective of the research problem and also the nature of the problem to be studied. A design may be quite appropriate in one case, however, may not suit for the context of some other research problem. One single design is not applicable for all kinds of research problems.

A research design suitable for particular research problem includes the following factors:

- The means of the collecting facts or information.
- The availability and skills of the manpower conducting the research.
- The objectives of the research problem to be analyzed.
- The nature of the problem to be analyzed.

- The availability of time and money for the research work.

The exploratory or formulative research focuses on discovery of ideas and insights, and must be flexible enough to permit the consideration of many different aspects of a phenomenon. However, if the aim of a research is accurate description of a situation or relationship among variables, accuracy becomes a prime concerned and a research design which minimizes bias and maximizes the reliability of the evidence is considered a good research design. Researches involving in the testing of hypothesis of a causal relationship between variables require a design which support for drawing inferences about causality in addition to the minimization of bias and maximization of reliability. In practice, it is difficult to include a specific study in a particular group as a given research may attempts to act for two or more functions of different studies. Therefore, the design may be categorized either as an exploratory or descriptive or hypothesis testing based on the primary function of the study. The availability of resources to be used in research and the means of obtaining the data must be given due weightage while designing the research such as experimental design, survey design, sample design etc.

3.5. Important concepts relating to research design

Before explaining the various research designs, it will be require to explain the various concepts applicable to design so that these may be better and easily understood. There are various important concepts relating to research design which are summarized in fig-1 and also described below:

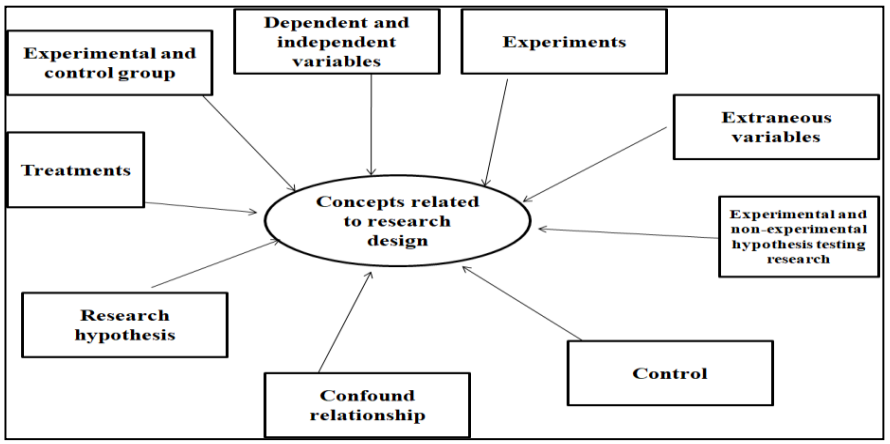


Fig-1: Concepts related to research design

Dependent and independent variables: An entity which varies with respect to unit, space, time or in combination of these is termed as variable. The variable can be quantitative and qualitative. As in environmental research, the quality of water can be assessed based on qualitative variables as color etc.; and quantitative variable as pH, dissolved oxygen, temperature etc. Qualitative phenomena (or the attributes) are also quantified on the basis of the presence or absence of the concerning attributes say absence or presence of bacteria in water. Variables whose values can be even in decimal points are called continuous variables as age, humidity and variables whose values can only be expressed as whole number is known as discrete variables as number of insects in a forest. A variable which depends on or is a consequence of the other variable, is termed as a dependent variable and a variable which influence the dependent variable is termed as an independent variable. For example, if we say that pH depends on sewage then pH is dependent variable and sewage is independent variable.

Extraneous variable: Variables that are not considered in the study, however may affect the dependent variable are known as extraneous variable. Suppose the researcher want to test the hypotheses that there is impact of domestic sewage on water quality of river. In this research sewage is independent variable and water quality parameters (pH, Temperature, total solids, total dissolved solids, total suspended solids, turbidity, transparency, dissolved oxygen, nitrates, phosphates etc.) are dependent variables. Livestock near to river may as well affect the water quality parameters, but since it is not considered to the purpose of the study by the researcher, it will be termed as extraneous variable. The effect on dependent variable due to extraneous variable (s) is known as "experimental error". A study must always be so designed that the effect on the dependent variable is attributed entirely to the independent variable (s) and not to some extraneous variable or variables. However, designing of research must consider all sort of extraneous factors so that experimental error can be minimized as much as possible.

Control: One important character of the research design is to minimize the influence of extraneous variables. It is better to consider a "control" for comparison purposes. "Control" is a situation, where all experimental variation or condition remains similar except the applied or used treatments i.e., the differences between control and treated should be

compared at least in relative manner. In experimental researches, the term "control" is used to refer to restrain experimental conditions. For example, a researcher wants to know the impact of sewage on water quality of river. He must select a site where sewage is not mixing with river water, and, therefore, researcher can compare the water quality with sewage and without sewage site. The site where sewage is not mixing with river water is known as "control" site.

Confounded relationship: When the dependent variable is not free from the influence of extraneous variables, the relationship between the dependent and independent variables is known to be confounded by extraneous variables.

Research hypotheses: When a prediction or a hypothesized relationship is to be tested by scientific ways, it is termed as research hypothesis. It is a predictive statement that relates independent variable to dependent variable. Generally, a predictive research hypothesis should contain, at least one independent and one dependent variable and state the relationship between the two. Predictive statement which is objectively not verified is not considered as research hypothesis.

Experimental and non-experimental hypothesis-testing research: When the purpose of research is to test a research hypothesis, it is termed as hypothesis testing research. The testing of hypothesis may be based on the experimental design or non-experimental design. Study in which the independent variable is manipulated through some stimuli or treatment is termed as "experimental hypothesis testing research" and a research in which an independent variable is not manipulated is called "non-experimental hypothesis testing research". For example, suppose a researcher wants a study whether sewage affects the water quality of river and for this purpose he randomly selected 5 sewer drains sites. This is an example of non-experimental hypothesis testing research because herein the independent variable i.e., sewage is not manipulated or treated. The hypothesis may be "sewage affects water quality". However, if the researcher randomly selects 5 sewer drains from a particular area and classify the five sewer drains in 2 or 3 categories based on volume of sewage. In this case, hypothesis may be "higher sewage volume makes the water unusable for drinking", this is an example of experimental hypotheses testing research.

Experimental and control group: In an experimental hypothesis-testing research when a group is exposed to usual conditions, it is termed as control group, but when the group is exposed to some novel or special condition, it is termed as "experimental group". If you wish to evaluate the effect of growth regulators on plant growth and laid an experiment with application of five growth regulators on plants and one plant without growth regulator, then the plants without growth regulator is control group and other plants with growth regulator are termed as experimental group.

Treatments: The different conditions, which experimenter wishes to compare are usually referred to as "treatments". If researcher wants to study on impact of three types of fertilizers on yield of wheat, in this case the three types of fertilizers will be treated as "treatments".

Experiment: The process of analyzing the facts of the statistical hypothesis, relating to research problem, is called as experiment. For example, researcher can conduct an experiment to examine the usefulness of a newly prepared drug. Experiments may be categorized in to two types: absolute experiment and comparative experiment. If researcher wants to determine the impact of sewage on water quality of river, it is a case of absolute experiment, but if he wants to compare the impact of volume of sewage it is termed as "comparative experiments".

3.6. Different research designs

There are different research designs which are described below:

1. **Research design in case of exploratory research studies:** Exploratory research studies are also known as formulative research studies. The aim of these studies revolves around formulating a problem for more precise investigation or of developing the working hypothesis from an operational point of view. The major emphasis in such studies is on the discovery of ideas and insights. As such the research design suitable for these studies must be flexible enough to provide opportunity for considering different aspects of a problem under study. Inbuilt flexibility in research design is required because the research problem, broadly defined initially, is transformed in to more precise meaning in exploratory studies, which in fact may necessitate changes in the research procedure for data

collection. Generally, the following three methods in the context of research design for these studies are considered:

a. **The survey of concerning literature:** The survey of related literature is the most important for formulating the research problem or developing hypothesis. Hypothesis stated by earlier workers on the research should be reviewed and accordingly further research should be planned. It can also be evaluated that the already stated hypothesis may lead to new hypothesis. With this approach, the researcher should review and build upon the available knowledge, however, if hypothesis have not yet been formulated, the researcher must review the available literature for deriving the appropriate hypothesis. The researcher should also make an attempt to apply concepts and develop theories in different research contexts to the area in which the researcher is himself working.

b. **The experience survey:** It means the survey of people who have had practical experience with the problem to be analyzed. The object of such a survey is to obtain insight into the relationships between variables and new ideas relating to the research problem. For such type of survey people/communities who are some knowledge and competent and can add new ideas may be carefully chosen as respondents to make a representation of different types of experience. The selected respondents may be interviewed by the researcher based on schedule. The interview should be flexibility so that the respondents should raise those issues and questions which the investigator has not been considered. Usually, the experience interview is likely to be long and may last for few hours. Hence, it is generally referred desirable to send copy of the questions to be discussed to the respondents well before in time. This will also give an opportunity to the respondents for some advance thinking over the various aspects of the schedule and facilitates to respond effectively at the time of interview. Therefore, an experience survey may enable the researcher to define the research problem more concisely and help in the formulation of the research hypothesis. This

survey may also provide facts about the practical possibilities for doing various types of research.

c. **The analysis of insight stimulating examples:** It is also useful method for suggesting hypothesis for research. It is specifically suitable when there is little knowledge is available to serve as a supervisor. This method intends to have intensive study of selected instances of the phenomenon in which researcher is interested. For this purpose, the existing records can be evaluated, or unstructured interview may be conducted, or some other approach may be adopted, which provide information about the phenomenon. Attitude of the researcher, the intensity of the study and the ability of the researcher to draw information into a unified explanation are the main characters which make this method suitable procedure for suggesting insights.

What sorts of examples are to be selected and studied, is not clearly defined? Experience shows that for particular problems certain types of instances are more suitable than others. One can note few examples of "insight stimulating" cases such as the reactions of the strangers, the study of individuals who are in transition from one stage to another, etc. In general cases that provide sharp contrasts or having striking features are regarded relatively more appropriate for this method of hypothesis formulation.

An exploratory or formulative research merely leads to insights or hypothesis for a research design, and must be flexible so that many different facets of a problem may be considered depending upon the requirement.

Research design in case of descriptive and diagnostic research studies: Descriptive research studies are referred the description of the characteristics of a particular individual or of a group, whereas diagnostic research studies determine the frequency of occurrence of phenomenon or its relation with something else. The studies concerning the relationship between variables are example of diagnostic research studies. However, studies dealing the specific predictions, narration of observations and characteristics concerning individual, group or situation are examples of descriptive research. From the point of view of the research design, the descriptive and diagnostic studies have common requirements and

can be group into a single group of research studies. In descriptive and diagnostic studies, researcher must be able to define population explicitly and also what he wants to observe along with suitable method for measure it. Since the objective is to obtain complete and accurate information in the research, the procedure to be used should be carefully planned. The research design should make enough provision for bias reduction and maximize reliability, with due concern for the economics of the research. The design in these studies should be rigid and must focus attention on the following points:

- Formulating the objective of the study (what is the study and why is it important?)
- Designing the methods of data collection (what techniques of collecting data will be adopted?)
- Selecting the sample (how much material (chemicals, instruments, glassware) will be needed?)
- Collection of the data (where can the required data be found and with what time period should the data be associated?)
- Processing and analyzing the data
- Reporting the findings

In descriptive/diagnostic study, the first step is to define objectives precisely so that it can be ascertained that whatever data has been collected that is relevant for the research. This would ensure that the research leads to the desired conclusion. The second steps revolve around the selection of methods for data collection. Various methods viz. observation, questionnaires, interviewing, examination of records etc. are available for collection of data and research may be used any one depending upon the merit of the method. Designing for data collection procedure should ensure unbiasedness and reliability. Questions must be well defined and explicit, interviewer should not express their own opinion, and observers must be ensured to record the actual behavior of the given item. to the data collection instruments must be well tested before they are subjected to data collection. Precisely, “structured instruments” should be used for data collection in such studies.

Most of the descriptive/diagnostic studies are based on the samples and based on the data analysis of sample, researcher wishes to make conclusive statement about the population. The designing of sample should be made in such a fashion so that representative samples may be drawn and the accurate conclusion can be drawn for the population. The data collection process should be monitored effectively so that the data is free from error. Data collection should be examined for completeness, comprehensibility, consistency and reliability. The collected data must be processed and analyzed. The processing of the data includes coding the interview replies and observations, classification and tabulation of the data and statistical analysis. The data processing and analyzing procedure should be detailed before the data collection. Coding should be performed carefully so that error should not be crept. Statistical analysis should be performed as per plan. The suitable statistical calculations, along with the use of suitable tests of significance may be carried out to draw conclusions about the study. of the finding should be communicated in an efficient manner through report. The layout of the report needs to be well planned, so that all important component of the research should be included in simple and effective manner.

Thus, the research design in case of descriptive/diagnostic studies includes all the steps of a survey therefore referred as survey design

The difference between research design for the exploratory and descriptive/diagnostic research is elaborated below:

Table-1: Difference between research designs **(Source: Book on Research methodology methods and techniques: CR Kothari and Gaurav Garg)**

Research design	Type of study	
	Exploratory or formulative	Descriptive/diagnostic
Overall design	Flexible design (design must provide opportunity for considering different aspects of the problem)	Rigid design (design should be unbiased and reliable)
1. Sampling design	Non-probability sampling design (purposive or judgments sampling)	Probability sampling design (random sampling)
2. Statistical design		Pre-planned design for analysis

3. Observational design	No pre-planned design for analysis	Structured instruments for data collection
4. Operational design	Unstructured instruments for data collection No fixed decision about the operational procedures	Advanced decisions about operational procedures

3.7. Principles of Experimental Designs

Sir Fisher has enumerated three principles of experimental designs.

i. The Principle of Replication,

ii. The principle of Randomization and

iii. The Principle of Local Control.

The Principle of Replication: Replication means repetition i.e. Principle of Replication ensures the repletion of the experiments more than once. Replication ensures higher statistical accuracy of the experiments. For example, suppose researcher is to examine the effect of sewage on water quality. Researcher may take one sample in a quarter and can compare quarterly effect of sewage on water and draw conclusion. However, if the researcher applies replication and collects six water samples in each quarter and draw conclusion about the quarterly effect of sewage on water quality. The result of later experiment will be more reliable than the former. Replication in experiment is introduced to increase the precision of the results of the research by increasing the accuracy of the main effects and interactions.

The principle of Randomization: Randomization ensures minimization of bias in the experiment. This principle ensures that research design should be implemented in such a way that the variations caused by extraneous factors should be minimized by providing equal opportunity of allocation of the treatments into any plot of the experiment. For example, if researcher grows one variety of rice, say, in the first half of the parts of a field and the other variety is grown in other half, then it is possible that the soil fertility may be different in first half in comparison to the other half. If this is soothing case, the results

derived by such experimentation would not be realistic rather biased. In such a circumstance, researcher may assign the variety of rice to be grown in different parts of the field on the basis of some random sampling technique i.e., researcher may apply randomization principle and protects against the effects of the extraneous factors (soil fertility differences in the given case). As such, through the application of the principles of randomization, the validity of the test is ensured.

The Principle of Local Control: The Principle of Local Control ensure minimization of error by reducing the heterogeneity in the experimental units by dividing the whole experimental units into homogeneous units (similar units) known as block. The treatments can be randomly assigned to these units of the blocks. Dividing the field into several homogenous parts is known as "blocking". In principle of local control researcher can eliminate the variability due to extraneous factors from the experimental material.

3.8. Important Experimental Design

Experimental design refers to the layout of an experiment and of many types depending upon the experimental materials or units. Experimental design may be categorized in to two categories viz. informal experimental designs and formal experimental designs. Information experimental design are designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental design offers relatively more control and use precise statistical procedures for analysis.

Informal designs:

i. before and after without control design,

ii. after only with control design and

iii. before and after with control design.

Formal experimental designs:

Completely Randomized Design (CR design),

Randomized Block Design (RB design),

Latin Square Design (LS design)

1. **Before and after without control design:** This design considers two situations and the response is measured pre and post treatment application. The effect of the treatment would be equal to level of the response after the treatment minus the level of the response before the treatment. However, in such design, the limitation is in regard to the passage of time, which may lead to changes in the variation in the extraneous factors, resulting into impacting the response.
2. **After only with control design:** This design considers two groups or areas (test area and control area) and the treatment is applied into the test area only. The response is measured from both the areas at the same time. Treatment impact is evaluated by subtracting the values of the response in the control area from the response value in the test area.

The basic assumption in the design is the selection of identical sites for test and control. Failing to which, may lead introduction of extraneous factors leading to distortion of the result.

3. **Before and after with control design:** In this design, two areas are selected and the response is measured in both the areas at a fixed time before the treatment. The treatment is applied thereafter in the test area only and the response is measured in both at a same time after the application of the treatment. The treatment effect is estimated by subtracting the change in the response in the control area from the change in the response in the test area. This design is superior to the above two designs for the simple reason that it avoids extraneous variation resulting both from the passage of time and from non-comparability of the test and control areas.
4. **Completely randomized Design (CR design):** Completely randomized design is the simplest design and based on only two principles viz. the principle of replication and principle of randomization of experimental designs. The essential characteristic of the design is the homogeneity in the experimental materials or units or plots and allocation of the treatment to the units or plots in random fashion. For example, if experimenter has 10 subjects and wishes to test two treatment A and B, through the randomization process, experimenter can allocate

treatment A to five persons and treatment B to remaining five persons with an equal opportunity of being assigned to treatment A and treatment B. one way analysis of variance (or one way ANOVA) is used to analyze such a design. Even unequal replications can also applicable in this design

5. **Randomized block design (RB design):** Randomized Block design considers all the three principles of experimental designs. In the RB design, the experimental material or subjects or units or plots are divided into homogeneous groups of units or plots, such that within each group the units or plots or subjects are relatively homogenous in some respects. The variable selected for grouping the subjects is one that is believed to be related to the measures to be obtained in respect of the dependent variable. The number of subjects or units or plots in a block should be equal to the number of treatments. One subject or one unit or one plot in each block would be randomly assigned to each treatment. In general, blocks are the levels at which we hold the extraneous factor fixed, so that the contribution of extraneous factors to the total variability of data can be measured. The main feature of RBD design is that in this treatment appears the same number of times in each block. The RB design is analyzed by the two-way analysis of variance technique.

Let us illustrate the RB design with the help of an example. Suppose four chemicals were tested for increasing fertility of wastelands in an area. The wastelands were divided into five categories based on the level of degradation. In this design, each treatment was randomized for one category of wasteland and similarly separate randomization were made for remaining four categories of the wastelands and accordingly the treatments were allocated in all five categories of wastelands. The randomization was made with the help of random number table.

Table-2. Example of Randomized block for evaluation of fertility improvement of wasteland for four treatments

Wasteland Category I	Wasteland Category II	Wasteland Category III	Wasteland Category IV	Wasteland Category V
Treatment A	Treatment C	Treatment A	Treatment B	Treatment C

Treatment D	Treatment A	Treatment C	Treatment D	Treatment B
Treatment B	Treatment D	Treatment B	Treatment A	Treatment D
Treatment C	Treatment B	Treatment D	Treatment C	Treatment A

Latin square design (LS design): Latin square design is suitable for the experimental material which has two ways variability. For example, an experiment has to be made for evaluating productivity of the five different varieties of agroforestry species on farm land under five different fertilizers. The varying fertility of the farm land in which the experiments have to be performed must be taken into consideration, otherwise the results would be biased because the output i.e., productivity of each species contains the effect of not only species and fertilizers, but it may also be the effect of fertility of soil. The problem of variability due to two factors i.e., species and fertilizers can be accounted by using LS design. In this design, each fertilizer appears once in each row and each species appears once in each column of the design. In other words, the treatments in a LS design are allocated among the plots in such a fashion so that no treatment occurs more than once in any one row or any one column. The two blocking factors may be represented through rows and columns (say fertilizer in row and species in column). In this design number of levels for the two factors and the number of blocks are same. The layout of LS design for five levels of factors are as follows. **Table-3:** Layout of Latin square design

	Block I	Block II	Block III	Block IV	Block V
Species A	A	B	C	D	E
Species B	B	C	D	E	A
Species C	C	D	E	A	B
Species D	D	E	A	B	C
Species E	E	A	B	C	D

The above diagram clearly shows that in a LS design the field is divided in to as many blocks as there are varieties of species and number of fertilizers and then each block is again divided into as many parts as there are number of fertilizers in such a way that each of the fertilizer is applied in each of the block (whether column-wise or row wise) only once. The analysis of the LS design is very similar to the two-way ANOVA technique.

This design suffers from one limitation that is the possibility of interaction between treatments and blocking factors. This defect, can however, be removed by taking the

means of rows and columns equal to the field mean by adjusting the results. Another limitation of this design is that it requires number of rows, columns and treatments to be equal. This reduces the utility of this design for larger number of treatments to be tested. If treatments are 10 or more then the design will be large in size and chances exits those rows and columns may not be homogenous. In case of (2x2) LS design, the error degrees of freedom become zero and therefore the design cannot be used. Therefore, LS design of orders (5x5) to (9x9) are generally used.

6. **Factorial experiments:** Factorial experiments are used in experiments where the effect of more than one factor is being evaluated. Factorial design may be classified in two categories viz. simple factorial design and complex factorial design. These factorial designs are discussed below:

   i. **Simple factorial designs:** In case of simple factorial designs, we consider the effects of different levels of two factors on the response variable. Simple factorial design may either be a 2 x 2 or it may be said 3 x 4 or 5 x 3 simple factorial design. We illustrate some simple factorial designs as under.

Table-4: An example of 2 x 2 simple factorial design i.e., two factors with two levels

	Treatment A	Treatment B
Level I	Cell 1	Cell 3
Level II	Cell 2	Cell 4

Table 7: An example of 4x3 simple factorial design i.e., four treatments each with three levels

	Experimental variable			
	Treatment A	Treatment B	Treatment C	Treatment D
Level I	Cell 1	Cell 4	Cell 7	Cell 10
Level II	Cell 2	Cell 5	Cell 8	Cell 11
Level III	Cell 3	Cell 6	Cell 9	Cell 12

This model of a simple factorial design includes four treatments viz. A, B, C and D each at three levels viz. I, II and III and has 12 different cells for one replication. In such a design the means for the columns provide the researcher with an estimate of the main effects for treatments and the means for row provide an estimate of the main effects for the levels. Such a design also enables the researcher to determine the interaction between treatment and levels.

ii. **Complex factorial design:** Experiments with more than two factors at a time are considered as complex factorial designs. Simple factorial design is also termed as a "two factor factorial design" whereas complex factorial design is known as "multifactor factorial design". In case of three factors having two treatments each with two conditions and each one of which having two levels, the experiment used to be termed 2 x 2 x 2 complex factorial design which will contain a total of eight cells as shown below:

Table-8: An example of 2 x 2 x 2 factorial experiment (Source: Book on Research methodology methods and techniques: CR Kothari and Gaurav Garg)

Level				
	Treatment A		Treatment B	
	Condition A	Condition B	Condition A	Condition B
Level I	Cell 1	Cell 3	Cell 5	Cell 7
Level II	Cell 2	Cell 4	Cell 6	Cell 8

On the basis of above account, we can summarize that there are several research designs and the researcher must decide the design in advance for collection and analysis of data depending upon the objective and resources. The experimenter must attaché due consideration to the various points such as the type of universe and its nature, the objective of the study, the resource list or the sampling frame, desired standard of accuracy for selecting the appropriate design for his research.

3.9. Summary

In this unit we have discussed meaning, definitions and types of research designs. So far you have learnt that:

- According to Burns and Grove a research design may be defined as "A blue print for conducting a research study maximum control over variables which could interfere with the validity of findings"
- According to Parahoo "A plan which explains, how when and where data are to be collected and analyzed is called research design"
- Overall research design may include the sampling design, the observational design, the statistical design and the operational design.
- A research design suitable for particular research problem, generally includes the consideration of some factors such as: the means of the collecting facts or information, the availability and skills of the researcher, the objectives of the research problem to be analyzed, the nature of the problem to be analyzed etc.
- There are various concepts related to research design as dependent and independent variables, extraneous variable, control, confounded relationship, research hypotheses, experimental and non-experimental hypothesis-testing research, experimental and control group, treatments, experiment etc.
- There are three main principles of experimental design viz. the Principle of Replication, the principle of Randomization and the Principle of Local Control
- Experimental design may be categorized in two categories viz. informal experimental designs and formal experimental designs. Information experimental design are designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental design offer relatively more control and use precise statistical procedures for analysis.
- Informal designs categorized into: before and after without control design, after only with control design and before and after with control design.

- Formal experimental designs may categorize into: completely randomized design (CR design), randomized block design (RB design), latin square design (LS design) and factorial designs.

TERMINAL QUESTIONS

1. (a) Fill in the blank spaces with appropriate words.

   A good design is generally ……by different characters and it should be ……. appropriate, efficient, economical etc. Generally the design should minimize …….and maximizes the ……..of the data collected and analyzed. This type of design is considered a……. ……. The design gives smallest experimental error is supposed to be the best design in many researches. In the similar way, design which yield maximal…….. and provides an opportunity for considering many different aspects of a problem is considered most appropriate and efficient design in respect of many…….. ……... Therefore, the question of good research design is related to the purpose or objective of the research problem and also with the nature of the problem to be studied. A design may be quite appropriate in one case, but can be found wanting in one respect or the other in the context of some other research problem. One …….design cannot serve the aim of all kinds of research problems.

2. (a) Give the definition of research design.

   (b) What is need of research design in research?

3. (a) Describe the features of good research design.

   (b) Give the important concepts related to research designs

4. (a) Describe the different research designs.

5. (a) Discuss three principles of experimental designs.

   (b) What are informal experimental designs?

   (c) What are formal experimental designs?

6. (a) Fill the blank spaces with appropriate words.

In an experimental hypothesis-testing research when a group is exposed to usual conditions, it is termed as……. ……., but when the group is exposed to some novel or special condition, it is termed as **"…….. ……..."**. In the above illustration the river water (where no impacts of sewage) can be called as …….. group and other sewer drain 1, sewer drain 2 etc are termed as experimental group. If both groups (river water and sewer drain in this case) are exposed to special studies program, then both groups would be termed "experimental groups". It is possible to design studies which include only experimental groups or studies. Various conditions under which experimental and control groups are put are usually referred to as **"……."**. If researcher wants to study on impact of three varieties of fertilizers on yield of wheat, in this case the three varieties of fertilizers will be treated as **"…….. ……"**.

(b) Before and after without control is type of informal design (Yes/No)

(c) Latin square design is type of informal design (Yes/No)

(d) factorial design may divided in to…….. categories (one/two/three/four)

(e) What do you understand by latin square design?

7. (a) Describe the CR designs.

(b) Explain dependent, independent and extraneous variables with examples.

ANSWERS

1 (a) Characterized, flexible, bias, reliability, good, design, facts, research, problems, single

2 (a) see section 3.2

(b) See section 3.3

3 (a) See section 3.4

(b) See section 3.5

4 (a) See section 3.6 including Table-1

5 (a) See the section 3.7

(b) see the section 3.8 including fig-2

(c) See section 3.8 including fig-2

6 (a) Control, group, experimental, group, control, treatments, three, treatments
(b) Yes

(c) No

(d) Two

(e) See the Latin square design in section 3.8

7 (a) See the section 3.8 under heading completely randomized design.

(b) See the section 3.5 under heading dependent, independent and extraneous variables.

Unit 4: Design of Sample Surveys: Sample Design and Sampling and Non-Sampling Errors; Types of Sampling Designs; Non-Probability; Probability and Complex Random Sampling Designs

Unit Structure

4.0. Learning Objectives

After studying this unit, you will be able to know:

- What is Sample design?
- Meaning of Sampling and Non-sampling error
- What are the types of sampling designs?
- What is non-probability and probability?
- What are complex random sampling designs?

4.1. Introduction

You have learnt various research designs in Unit-3 of this course. As you know all items in any field of inquiry constitute a **"universe"** or **"population"**. A collection of all units or items under consideration for the enquiry constitute the population. A complete enumeration of all items in the "population" is known as census or census survey. It is obvious that for any inquiry or investigation of a large population, census survey is rather infeasible. For example, to have an idea of

average per capita monthly income of people in India, we will have to enumerate all the earning individuals in the country, which is very difficult task.

Census survey is impossible in the situations when population is infinite. In some cases when population is finite but the units are destroyed while inspected as per the need of the inquiry, census survey is not at all desirable, e.g., inspection of crackers, estimation of root biomass of shrubs. Further, many times it is not possible to evaluate every item in the population, and sometimes it is possible to obtain sufficiently accurate results by studying only a part of total population. In such cases there is no utility of census surveys.

However, when the universe is a small i.e., the total items of the population is not large, it is no use resorting to evaluate a part of the total population i.e., evaluation of only some items. When field studies are undertaken in practical life, consideration of time and cost almost invariably lead to a selection of respondents i.e., selection of only a few items. The respondents or items selected should be a representative of the total population as possible in order to produce a true picture of the population. The selected respondents or items from the population is called a “sample” and the selection process, by which these items or respondents are selected, is called “sampling technique”. The survey so conducted is known as “sample survey”. Mathematically, let the population comprises of N number of items also termed as population size and if a part of size say n (which is $< N$) of the population is selected according to some rule for studying some characteristics of population, the group consisting of these n items or units is known as “sample” and n is termed as sample size. Researcher must prepare a sample design for his study i.e.; he must plan how many numbers of items should be selected for the evaluation along with the process of selection of these items. In this unit you will learn about sample designs, sampling and non-sampling errors, types of sampling designs, non-probability and probability and complex random sampling designs.

4.2. Sample design

A sample design is a definite plan for obtaining a sample from a given population to provide the answer for the inquiry. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design also considers the number of items to be included in the sample i.e., the size of the sample. Sample design is developed before the data collection process. There are many sample designs depending upon the objective of the inquiry and the population under investigation and researcher can choose any one sample design according to their study or research. Some designs are relatively more precise and easier to apply than others. Researcher must select/prepare a sample design which should be reliable and suitable for his research study. The main steps of the sample design are summarized in fig-1 and also elaborated in subsequent paragraphs:

1. **Objective:** Defining the objectives is the first step of sample design. These objectives should be defined clearly and explicitly. Researchers should ensure that the objectives are commensurate with the money, manpower and time required for the study.
2. **Population:** This is second step of sample design. In order to meet the objectives of the survey, population should be defined clearly and comprehensively. Population is collection of total units or items, on which the result of survey would be generalized.

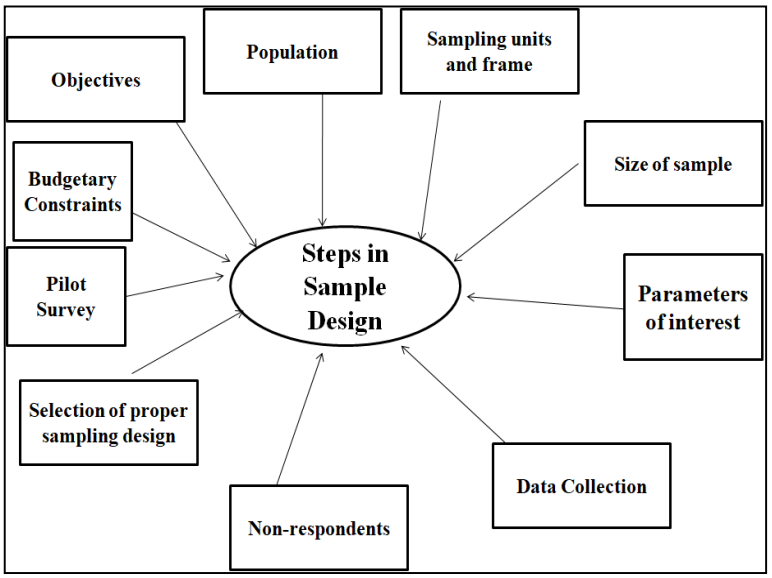


Fig-1: Important steps in sample design

Sampling units and frame: A decision has to make about a sampling unit before sample selection. Sampling unit may be a geographical entity such as state, district, village etc. or construction unit such as house, flat etc. or social unit such as family, club, school etc. or biological unit such as a tree, a shrub, or herb or part of tree. The researcher will have to decide one such unit that he has to select for data collection for his research. The listing of all sampling units of the population is known as "frame" or "sampling frame". Sampling frame contains some description of all the units of a population or universe (in case of finite universe only). Such a list must be comprehensive, correct, reliable, authentic and appropriate.

3. **Size of the sample:** Size of the sample refers to the number of items or units to be selected from the universe for the inquiry under study. The size of sample should neither be excessively large nor too small rather should be optimum. An optimum sample is one which fulfils the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must fix the desired precision and *acceptable confidence level for the estimate. The size of the population and the variability in population should be considered. Sample size is large, if the variability of the population is high. The parameters of interest in a research study must also be kept in view, while

deciding the sample size. The cost and manpower required to do the survey must also be considered for deciding the sample size.

4. **Parameters of interest:** Statistical constants of the population are called as parameters, e.g., population mean, population proportion, population variance etc. In census survey, we get the actual value of the parameters. On the other hand, in sample survey, we get the estimate of unknown population parameters in place of their actual values.

   In determining the sample design, one must consider the question of specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of the persons with some characteristics in the population or we may be interested to estimate average height of trees of a forest or the other measure concerning the population. There may also be important sub-groups in the population about whom the researcher would like to make estimates. All the above concern are important for designing the sample.

5. **Data collection:** In data collection, the focus should be collecting the essential information as per the inquiry and irrelevant facts must not be collected. The objective of the survey should be clear to the investigator or researcher while collecting the data.

6. **Non respondents:** Because of practical difficulties, sometimes data may not be collected for all the sampled units. This is known as non-response and non-response liable to alter the results. That is why the non-responses should be handled with care. The reason for non-response should be recorded by the researcher or investigator.

7. **Selection of proper sampling design:** The researcher must decide the technique to be used in selecting the items for the sample. There are several samples designs out of which the researcher must select one suiting for his study. Investigator must select the sampling design which, for a given sample size and for a given cost, has minimum sampling error.

8. **Pilot survey:** Piloting is also important step in sample design process. It is also helpful to apply the research design on a small scale before implementing the

study in full manner. Piloting is required to address the problem in survey well in advance. The process of conducting a survey on smaller units is called "pilot survey" or "pretest".

9. **Budgetary constraint:** Cost, budget and manpower are important components on decisions relating to size of the sample and sample designing. Proper budget and man power of the research work should be adjudged well in advance.

4.3. Sampling and Non-Sampling errors

The errors in the collection of data are categorized into two broad types: Sampling and non-sampling errors.

Sampling error: The sampling error arises due to the fact that only a part of the population has been used to analyze population parameters. The sampling error is absent in a census survey as all the units of the population is considered for analysis in census survey. Sampling errors can be measured for a given sample design and sample size. The measurement of sampling error is generally known as the "precision of the sampling plan". If sample size increases, the precision of the estimate is also increased. However, increasing the sample size increases the cost of collecting data and also enhances the systematic bias. Thus, the effective way to increase precision is usually to select a best sampling design which has a smaller sampling error for a given sample size at a given cost. In practice, however, researcher prefer a less precise design because it is easier to adopt and also because of the fact that systematic bias can be controlled in better way in such type of design.

In short while selecting a sampling procedure, researcher must ensure that the procedure causes a relatively small sampling error and helps to control the systematic bias in a better way.

Non-sampling error: These errors arise at the stage of collection and preparation and compilation of data and thus are present in both the sample survey as well as the census survey. Thus, the data obtained in census survey is free from sampling error, however subjected to non-sampling errors. Non- sampling errors can be

reduced by defining correctly the sampling units, frame and the population and by employing efficient people for data collection and analysis. Non –sampling errors arise due to number of factors such as: inefficient field workers, non-response, bias in data recording, data tabulation etc. These errors are likely to grow when the number of units inspected increases.

Sample survey Vs Census survey: In a sample survey, since researcher study only a part of the whole population, and thus the survey needs less money and less time. Most of the cases, non-sampling errors are so large that the result of sample survey are much more precise than those of census survey.

However, if the objective of study is very serious in nature and information is needed about each and every unit of the population, then there is no way out but to resort to census survey. Moreover, if time and money are not important factors or if population is not so large, a census survey may provide good results than any sample survey provided efficient and skilled staffs are employed.

4.4. Types of sampling Designs

The method of selecting a sample is fundamental importance and depends upon the nature of data and objectives of the investigation. The techniques of selecting a sample are classified in two broad categories: non-probability sampling and Probability sampling. There are several sampling designs which are summarized in Fig-2 and also described below:

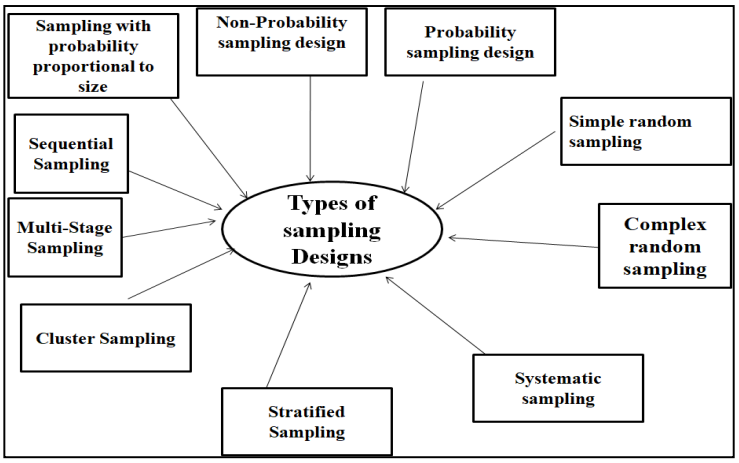


Fig-2: Different types of sampling designs

A. **Non-probability sampling:** non-probability sampling is that sampling procedure which does not attach any probability to the units for being included in the sample. Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgment sampling. In this type of sampling, items for sample are selected deliberately by the researcher; his choice concerning the selection of the items remains supreme. In other words, under non-probability sampling, the investigator of the inquiry purposively selects the specific units of the universe for constituting a sample. For example, if biodiversity of a state is to be studied, the investigator may select a few nearby forests, rivers, lakes etc. purposively for intensive study. The result of such study cannot be generalized for the whole population rather provide some information about the population. In such a design, personal choice has a great role for the selection of the sample. The researcher may choose a sample which shall yield result favorable to his choice and if that happens, the entire inquiry may get vitiated. Thus, there is always the danger of bias entering into this type of sampling technique. However, if the investigators are impartial, work without bias and have the necessary experience so as to take sound judgment, the result obtained from an analysis of deliberately selected sample may be tolerably reliable. However, in such type of sampling, there is no assurance that every unit of the population has some specifiable chance of being included. Sampling error in this type of sampling cannot be estimated and the element of bias is always present in such sampling. As such, this sampling design is rarely implemented in important inquiries. However, in small inquiries, this design may be adopted because of the relative advantage of time and money inherent in this method of sampling. This type of sampling is very convenient and is relatively inexpensive. But the samples so selected are not random samples. Quota sampling is also an example of non-probability sampling. Under quota sampling, the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled. In other words, the actual selection of the items for the sample is left to the judgment of interviewer. Quota samples are essentially judgment samples and

inferences drawn on the basis of these samples are biased and not agreeable to statistical treatment in a formal way.

B. **Probability sampling:** Probability sampling is also called as random sampling or chance sampling. Under this sampling design, every item of the population has some chance on inclusion in the sample. It is, so to say a lottery method in which individual units are picked up from the whole group not deliberately but by some mechanical process. Here the selection of sample item is based on probability alone that determines whether one item or the other is chosen. Researcher can estimate the error. Random sampling ensures the law of statistical regularity which states that if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the population. Therefore, the result of the random sampling can be generalized to the population. This is reason random sampling is considered as the best technique of selecting a representative sample. Random sampling is of many types depending upon the objectives of the inquiry and the nature of the units of the population. They are being described below:

i. **Simple random sampling:** This sampling from a finite population refers to that method of sample selection in which each item in the entire population has an equal chance of being included in the sample. Moreover, this method also gives each possible sample combination an equal chance of being picked up as sample. This sampling method is of two types i.e., sampling without replacement (SWoR) and sampling with replacement (SWR). In SWoR, once an item is selected for the sample, it cannot appear in the sample again. In SoWR, the selected item is not replaced in the universe, however, in SWR, the selected item is replaced in the universe before the next item is selected and therefore has chance to be chosen again. SWR is used less frequently. In such a situation the same element may appear more than once in the same sample. In brief, the important points of random sampling (or simple random sampling) are:

a. It attaches each element in the population an equal probability of getting selected in the sample, and all selections are independent of one another.

b. It gives an equal probability to all possible sample combination of being chosen.

Keeping this in view we can define a simple random sample (or simply a random sample) from a finite population as a sample which is chosen in such a way that each of the ${}^{N}C_{n}$ possible sample has the same probability $1/{}^{N}C_{n}$ of being selected (where N is population size and n is sample size). For example, let us have a certain finite population consisting of six elements (say a, b, c, d, e, f) i.e., N = 6. Suppose investigator is willing to take a sample of size n = 3 from the population. Then there are ${}^{6}C_{3} = 20$ possible distinct combinations of the samples of the required size of 3 as *abc, abd, abe, abf, acd, ace, acf, ade, adf, aef, bcd, bce, bcf, bde, bdf, bef, cde, cdf, cef* and *def*. Now, if researcher chooses one of these samples in such a way that each has the probability 1/20 of being chosen, then this is a random sample.

With regards to the question of how to take a random sample in actual practice, researcher could, in simple cases like the one above, write each of the samples on a slip of paper, mix these slips thoroughly in a container and then draw as a lottery either blindfolded or by rotating a drum or by any other similar device. The selected draw would be the sample for the inquiry. Such a procedure is clearly impractical, and impossible in complex problems of sampling. In fact, the practical utility of such a method is very limited.

Researcher can draw a random sample in a relatively easier method without taking the trouble of enlisting all possible samples on paper as discussed above. Instead of this, researcher can write the name of each element of a finite population on a slip of paper and put the slips of paper into a box or a bag and mix them thoroughly. Now for drawing sample of size three, the researcher can draw the required number of slips for the sample one after the other without replacement. In doing so researcher should make sure that in successive drawing each of the remaining elements of the population has the same chance of being

selected. This procedure will also result in the same probability for each possible sample. Since, researcher has finite population of 6 elements and he wants to select a sample of size 3, the probability of drawing one more element in the second draw is 2/5 (the first element drawn is not replaced) and in the same way the probability of drawing one more element in the third draw is ¼. Since, these draws are independent, the joint probability of the three elements which constitute sample is the product of their individual probabilities i.e., 3/6 x 2/5 x 1/4 = 1/20, which is same as discussed earlier and therefore verifies the earlier calculation.

This method of obtaining a random sample can be simplified in actual practice by the use of random number tables. Various statisticians like Tippett, Yates and Fisher have prepared tables of random numbers which can be used for selecting random sample. Generally, random number of Tippett's tables is used for drawing a random sample. Random number table is collection of random numbers in row and column format. The researcher can draw the required number of sample size by selecting the numbers from the Random Number table.

Table-1: A snapshot of Tppett's Random Number Table. **(Source: Book on Research methodology methods and techniques: CR Kothari and Gaurav Garg)**

2952	6641	3992	9792	7979	5911
3170	5624	4167	9525	1545	1396
7203	5356	1300	2693	2370	7483
3408	2769	3563	6107	6913	7691
0560	5246	1112	9025	6008	8126

Suppose, researcher is interested in taking a sample of 10 units from a population of 5000 units, bearing numbers from 3001 to 8000 (say). Researcher shall select 10 such figures from the random number table which should not be less than 3001 and not greater than 8000. If he randomly decides to read the table numbers from left to right, starting from the first row itself, he obtains the following numbers 6641, 3992, 7979, 5911, 3170, 5624, 4167, 7203, 5356 and 7483. The units bearing the above serial numbers would be selected and constitute the required random sample. One may note that it is easy to draw random samples from finite

populations however, it is often impossible to proceed in the way once the population does not have finite number of units. For example, if researcher wants to estimate the mean height of trees in a forest, it is not be possible to number all the trees, and choose random numbers to select a random sample. In such situations the investigator should select some trees for the sample and should treat the sample as a random sample for study purposes.

Simple random sample, we discussed above in "simple random sample without replacement". In case of infinite populations, it is relatively difficult to explain the concept of random sampling. However, a few examples will show the basic characteristics of such a sample. Suppose researcher considered the 20 throws of a fair dice as a sample from the hypothetically infinite population which consists of the results of all possible throws of the dice. If the probability of getting a particular number, say 1 is the same for each throw and the 20 throws are all independent, then we say that the sample is random. In the same way, it would be said to be sampling from an infinite population and sample would be considered as random sample if in each draw all elements of the population have the same probability of being selected and successive draws happen to be independent. In short, one can say that the selection of each item in a random sample from an infinite population is controlled by the same probabilities and that successive selections are independent of one another.

Complex random sampling design: Some complex random sampling design, which are the mixture of probability and non-probability sampling methods are discussed below:

1. **Systematic sampling:** In some cases, the most practical way of sampling is to select every i^{th} item on a list. This type of sampling is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random number to pick the first unit with which sampling to start. For example, if 4% sample is desired, the first item should be selected randomly from the first 25 items and thereafter every 25^{th} item would automatically be incorporated in the sample. Therefore, in systematic sampling only the first unit is

selected randomly and the remaining units of the sample are selected at fixed intervals.

Systematic sampling is very important and has as an improvement over a simple random sample as the systematic sample is spread more evenly over the entire population. This is an easy and less costly method of sampling and can be conveniently used even in case of large populations. However, this sampling has some problems with population having periodicity. If there is an unseen periodicity in the population, systematic sampling will prove to be an inefficient method of sampling. For example, say, every 25^{th} item produced by a certain production process in a factory is defective. If researcher select 4% sample of the items of this process in a systematic manner, he would either get all defective items or all good items in the sample depending upon the random starting position. If all elements of the universe are ordered in a manner representative of the total population i.e., the population list is in random order, systematic sampling is considered equivalent to random sampling. But if this is not so, then the results of such sampling may, at times, not be reliable. In practice, systematic sampling is used when lists of population are not available and they are of considerable length.

2. **Stratified sampling:** If a population from which a sample is to be drawn does not comprise a homogenous group, i.e., items of the population are heterogeneous, then stratified sampling technique is generally used in order to get a representative sample. In stratified sampling the population is divided into several sub-populations so that these sub-populations are individually more homogenous than the total population (the different sub-populations known as strata) and researcher select items from each stratum to comprise a sample. Since, each stratum is more homogenous than the total population, researcher will able to get more accurate estimates for each stratum than the total population, and by estimating more accurately for each of the sub-population, researcher get a better analysis of the whole population. In short, stratified sampling results in more reliable and provided

detailed observations. The following three questions are highly relevant in the context of stratified sampling.

i. How to prepare strata?

ii. How should items be selected from each stratum?

iii. How many items should be selected from each stratum?

In case of the first question, we can say that the strata should be prepared in such a way as to ensure elements being most homogenous within each stratum and most heterogeneous between the different strata. Therefore, strata are purposively prepared and are generally based on previous experience and personal judgment of the researcher. Careful consideration should be made for identifying the characteristics of the population to define the strata. At times, pilot study may be conducted for determining a more suitable and efficient stratification plan.

In respect of the second question, generally, selection of items for the sample from each stratum is simple random sampling. Systematic sampling can be used in certain cases if it is considered more suitable.

Regarding the third question, researcher generally follows the method of proportional allocation under which the sizes of the samples from the different strata are kept proportional to the sizes of the strata. That is, if Pi represents the proportion of total population items for stratum I and n represent the total sample size, the number of elements selected from the stratum i is n x Pi. For example, suppose we wish a sample of size n = 30 to be drawn from a population of size N = 8000 which is divided into three strata of size N1 = 4000, N2 = 2400 and N3 = 1600 then in proportional allocation, following sample sizes would be selected for the different strata:

For strata with N1= 4000, we have P1 = 4000/8000

And hence n1 = n. P1= 30 (4000/8000) = 15

Similarly, for strata with N2 = 2400, we have

n2 = n. P2 = 30 (2400/8000) = 9 and

for strata with N3 = 1600, we have

n3 = n. P3 = 30 (1600/8000) = 6

i.e., the 15, 9 and 6 items were selected from the three strata.

3. **Cluster sampling:** Cluster sampling used when the total area of interest happens to be large. It is a convenient way in which population is divided into a number of smaller non-overlapping areas and then a number of these smaller areas (usually called clusters) is selected randomly with the ultimate sample consisting of all (or samples of) units of these small areas or clusters.

   Thus, in cluster sampling, the total population is divided into a number of relatively small sub-divisions which are themselves clusters of smaller units and then some of these clusters are randomly selected for inclusion in the overall sample. Suppose, we want to estimate the proportion of machine –parts in an inventory which are defective and also assume that there are 20,000 machine parts in an inventory at a given point of time, stored in 400 cases of 50 machine parts each. Now, using a cluster sampling, we would consider the 400 cases as cluster and randomly select 'n' cases and examine all the machine –parts in each randomly selected case.

   Cluster sampling reduces cost by concentrating surveys in selected clusters. But certainly, it is less precise than random sampling for the estimate. There is also not as much information 'n' observations within a cluster as there happens to be in 'n' randomly drawn observations. Cluster sampling is used only because of the economic advantage it possesses; estimates based on cluster samples are generally more reliable per unit cost.

   If cluster happen to be some geographic sub-divisions, cluster sampling is better known as area sampling. In other words, cluster designs, where the primary sampling unit represents a cluster of units based on geographic area, are distinguished as area sampling. The merits and demerits of cluster sampling are also applicable to area sampling.

4. **Multi-stage sampling:** It is a further development of the principles of cluster sampling. Suppose, researcher wants to monitor the working efficiency of nationalized banks in India and he wants to take a sample of few banks for this

purpose. The first stage is to select large primary sampling unit such as states in a country. Then he may select certain districts from the selected states and interview all banks in the chosen districts. This would represent a two-stage sampling design with the ultimate sampling units being clusters of banks in districts.

If instead of taking a census of all banks within the selected districts, researcher selects certain blocks and interviews all banks in the chosen blocks. This would represent a three-stage sampling design. If instead of taking a census of all banks within the selected blocks, researcher randomly selects sample banks from each selected sub-blocks, then it is a case of using a four –stage sampling plan. If researcher select units randomly at all stages, researcher will have what is known as multistage sampling design.

Ordinarily multistage sampling is applied in big inquiries extending to a considerable large geographical area, say the entire country. There are two advantages of this sampling design which are given below:

(i) It is easier to administer than most single stage design mainly because of the fact that sampling frame under multistage sampling is developed in partial units.

(ii) A large number of units can be sampled for a given cost under multi-stage sampling because of sequential clustering, whereas this is not possible in most of the simple design.

5. **Sampling with probability proportional to size:** In cluster sampling, units do not have the same number, it is regarded suitable to use a random selection process where the probability of each cluster being included in the sample is proportional to the size of the cluster. For this purpose, researcher has to list the number of elements in each cluster irrespective of the method of ordering the cluster. Then researcher must sample systematically the suitable number of elements from the cumulative totals. The actual number selected in this way do not refer to individual elements, but indicate which clusters and how many elements from the cluster are to be selected by simple random sampling or by systematic sampling. The results of this type of sampling are equivalent to those of a simple random sample and the

method is less unwieldy and is also relatively less expensive. We can understand this with the help of an example.

Example: Following are the number of departmental stores in 15 cities: 35, 17, 10, 32, 70, 28, 26, 19, 26, 66, 37, 44, 33, 29 and 28. If researcher wants to select a sample of 10 stores, using cities as clusters and selecting within clusters proportional to size, how many stores from each city should be chosen? (Use a starting point of 10).

Solution: Let us put information as under (Table-2)

Since in the given problem, we have 500 departmental stores from which we have to select a sample of 10 stores, and wish to apply systematic sampling for selection of samples. The sampling interval for the systematic sampling is 50 (500/10). As we have to use the starting point of 10 so we add successively increment of 50 till 10 numbers (stores) have been selected. The numbers, thus, obtained are: 10, 60, 110, 160, 210, 260, 310, 410 and 460 which have been shown in the last column of the table (Table 4.1) against the cumulative totals. Therefore, 10 stores having the number as obtained, would constitute the sample. From this we can say that two stores should be selected randomly from city number five and one each from city number 1, 3, 7, 9, 10, 11, 12 and 14. This sample of 10 stores is the sample with probability proportional to size.

Table-2: Demonstration of probability proportional to size sampling (Source: Book on Research methodology methods and techniques: CR Kothari and Gaurav Garg)

City Number	No. of Departmental stores	Cumulative totals	Sample
1.	35	35	10
2.	17	52	
3.	10	62	60
4.	32	94	
5.	70	164	110 and 160
6.	28	192	
7.	26	218	210
8.	19	237	
9.	26	263	260
10.	66	329	310
11.	37	366	360

12.	44	410	410
13.	33	443	
14.	29	472	460
15.	28	500	

* If the starting point is not mentioned, then same can be randomly be selected.

6. **Sequential Sampling:** It is a complex sample design. The ultimate size of the sample under this technique is not fixed well in advance, but is determined according to mathematical decision rules on the basis of information obtained as survey progresses. This is usually adapted in case of acceptance sampling plan in context of statistical quality control. When a particular lot is to be accepted or rejected on the basis of a single sample, it is known as single sampling, when the decision is to be taken on the basis of two samples it is known as double sampling and in case of the decision rests on the basis of more than two samples, but the number of samples is certain and decided in advance, the sampling is known as multiple sampling. But when the number of samples is more than two but it is neither certain nor decided in advance, this type of system is often referred to as sequential sampling. Thus, in very short we can say that in sequential sampling one can go on taking samples one after another as long as one desires to do so.

   From a concise description of the various sampling designs presented above, we can state that normally one should resort to simple random sampling because in the sampling bias is generally eliminated and the sampling error can be estimated. Purposive sampling is considered more suitable when the universe is small and a known characteristic is to be studied intensively. There are situations in real life under which sample designs other than simple random samples may be considered better (say easier to obtain, cheaper or more informative) and as such the same may be used. At times, numerous methods of sampling may well be used in the similar researches depending upon the objectives and resources available for survey.

4.5. Summary

In this unit we have discussed about sample designs, sampling errors and types of sampling designs. So far you have learnt that:

- A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well as lay down the number of items to be included in the sample i.e., the size of the sample. Sample design should be decided before data collection.
- The main steps of the sample design are objective, population, sampling units and frame, size of the sample, parameters of interest, data collection, non-responses, selection of proper sampling design, pilot survey and budgetary constraint
- The errors the collection of data is categorized into two broad types: Sampling and non-sampling errors.
- The sampling errors arise due to the fact that only a part of units of the population has been used to analyze population parameters. These errors are absent in a census survey. Sampling errors can be measured for a given sample design and size. The measurement of sampling error is generally known as the “precision of the sampling plan”.
- Non-sampling error arise at the stage of collection and compilation and preparation of data and thus are present in both the sample survey as well as the census survey. Thus, the data obtained in census survey is free from sampling error, however subjected to non-sampling errors. Non- sampling errors can be reduced by defining the sampling units, frame and the population correctly and by employing efficient people in the data collection and analysis.
- The method of selecting a sample is fundamental importance and depends upon the objective, nature of data and investigation. The techniques of selecting a sample are classified in two broad categories: non-probability sampling and Probability sampling.
- There are several sampling designs such as non-probability sampling, probability sampling, simple random sampling, complex random sampling design, systematic sampling, stratified sampling, cluster sampling, multi-stage sampling, sampling with probability proportional to size and sequential sampling

TERMINAL QUESTIONS

1. (a) Fill in the blank spaces with appropriate words.

   A sample ……is a definite plan for obtaining a sample from given……. It refers to the technique or the procedure the researcher would adopt in selecting items for the…….. Sample design may as well as lay down the number of items to be included in the ……..i.e., the size of the sample. Sample design determined before…….. ……..... There are many sample designs and researcher can…….. any sample design according to their study or research. Some designs are relatively more ……..and easier to apply than others. Researcher must select/prepare a sample design which should be ……and suitable for his research study.

2. (a) Define sample design.

   (b) What are main steps of sample design?

3. (a) Describe sampling and non-sampling errors

   (b) Differentiate between non-probability sampling errors and probability sampling errors.

4. (a) Discuss the simple random sampling

5. (a) Give the different types of complex random sample design.

   (b) What do you understand by stratified sampling?

6. (a) Fill the blank spaces with appropriate words.

   The sampling ……arise due to the fact that only a part of the population has been used to analyze population parameters. These ……..are absent in a census survey. Sampling errors can be measured for a given sample design and size. The measurement of sampling error is generally known as the **"…… …….. …….. …….. ……".** If investigator ……the sample size, the …….can be improved. But increasing the size of sample has its own ……...viz. a large sized sample increased the cost of collecting data and also enhances the systematic bias. Thus the effective way to increase precision is usually to select a best

…… ……which has a smaller sampling error for a given sample size at a given cost. In practice, however, researcher prefer a less precise design because it is easier to adopt the same and also because of the fact that systematic bias can be controlled in better way in such type of design.

(b) Objective preparation is first step in sample design (Yes/No)

(c) Quota sampling is an example of **(**probability sampling/non-probability sampling**)**

(d) Non-probability sampling is also known as (Purposive sampling/chance sampling)

(e) Probability sampling is also known as (random sampling/deliberate sampling)

(f) What do you understand by cluster sampling?

7. (a) Describe the multi stage sampling

(b) What is sequential sampling? Explain in brief.

ANSWERS

1. (a) Design, population, sample, sample, data, collection, choose, precise, reliable
2. (a) see section 4.2.

   (b) See section 4.2
3. (a) See section 4.3

   (b) See section 4.3
4. (a) See section 4.4 under heading simple random sampling
5. (a) See the section 4.4 under heading complex random sampling

   (b) see the section 4.4 under heading stratified sampling
6. (a) Errors, errors, precision of the sampling plan**,** increases, precision, limitations, sampling design,

   (b) Yes

   (c) Non-probability sampling

(d) Purposive sampling

(e) Random sampling

(f) see the section 4.4. under heading cluster sampling

7. (a) See the section 4.4 under heading multi-stage sampling

(b) See the section 4.4 sequential sampling

Unit 5: Measurement and Scaling: Quantitative and Qualitative Data; Classification and Goodness of Measurement Scales; Sources of Error in Measurement, Techniques of Developing Measurement Tools; Scaling; Classification Bases; Techniques and Multi-Dimensional Scaling; Deciding the Scale

Unit Structure

5.0. Learning Objectives

After studying this unit, you will be able to:

- What is measurement?
- Meaning of quantitative and qualitative data.
- Classification of measurement scales
- What is goodness of measurement scales?
- What are sources of errors in measurement?
- Techniques of developing measurement tools
- What is scaling?

- What are scale classification bases?
- What are scaling techniques?
- Deciding the scales

5.1. Introduction

In our daily life, we use some standard to determine weight, height, and other features of a physical object. We also measure while judging a song, a painting or the personality of a friend. We, thus measure physical objects as well as abstracts concepts, depending upon the requirements. Measurement of these objects is a relatively complex and challenging task, especially when it concerns qualitative and abstract phenomena. Examples of qualitative characteristics are taste, honesty, intelligence, customer's perception and brand loyalty etc. These characteristics are also known as constructs or parameter.

Measuring qualitative features or characteristics as social conformity, intelligence or marital adjustment is not obvious and requires closer attention than measuring physical weight, biological age or a person's financial assets. In other words, parameters such as weight, height, etc. can be measured directly with some standard unit of measurement through comparing, but it is not that easy to measure properties like motivation to succeed, ability to stand stress etc. The meaningful assessment of the qualitative characteristics can be achieved by essentially measuring the characteristics. Scaling facilitates for measurements of qualitative characteristics and is one of the most important for research. In this unit, you will learn about measurement scaling, quantitative and qualitative data, classification and measurement of goodness of scales, sources of error in measurement, techniques for developing measurement tools, classification-based techniques and multidimensional scaling and deciding the scale.

5.2. Quantitative and Qualitative Data

Measurement is defined as a process of associating numbers or symbols to data obtained in a research study. These observations or data could be qualitative or quantitative. Generally, the research analysis is based on quantitative data. For example, mean, standard deviation, etc. can be computed for quantitative characteristics. Qualitative data can be

counted and cannot be computed. Therefore, the researcher should have a clear meaning of the type of data or variable before collecting the data. The result on qualitative variables may also in the form of numbers. For example, we can note the shape of tree stem as 1,2, or 3 depending on whether the three stem is single, bi-furcated, or twisted. We can as well as record "yes or no" answers to a question as "0" and "1" (or as 1 and 2 or perhaps as 59and 60). For example, we can represent survivalist of a seedling as 1 or 2 for survival or non-survival Categorical data (qualitative or descriptive) can be made in to numerical data by the artificial way. The coding of the various categories is known as nominal data.

Nominal data is numerical in name only, because they do not share any of the mathematical properties of the numbers we deal in ordinary arithmetic as greater than, less than, addition, subtraction, multiplication and division. For example, if we note the shape of tree as 1,2, or 3 as stated above, we cannot write 3 >2 or 3<1 and we cannot write 3-2 =2-1, 1+3=4 or 3 ÷2= 1.5 as this has no meaning. The only property holds are equality i.e., color of a flower may be exactly same to the other flower.

In situations, some characteristics can be represented as inequalities, we refer to the data as ordinal data. For example, if one mineral can scratch another, it receives a higher hardness number and on Mohs' scale the numbers from 1 to 10 are assigned respectively to talc, gypsum, calcite, fluorite, apaptite, feldspar, quartz, topaz, sapphire and diamond. In ordinal data, some of mathematical (arithmetic) properties of the numbers holds as greater than, less than, or equality, however, addition, subtraction, multiplication and division do not hold for ordinal data. For example, with the numbers in above example of mineral, we can write 5>2 or 6<9 as apatite is harder than gypsum and feldspar is softer than sapphire, but we cannot write for example 10– 9=5-4, because the difference in hardness between diamond and sapphire is actually much greater than that between apatite and fluorite. It would also be meaningless to say that topaz is twice as hard as fluorite simply because their respective hardness numbers on Mohs' scale are 8 and 4. The greater than symbol (i.e., >) in connection with ordinal data may be used to designate "higher than" or "preferred to".

In some situations, characteristics can be employed differences in addition to setting up inequalities, we refer to the data as interval data. For example, we are having temperature (in degrees Fahrenheit): 58^0, 63^0, 70^0, 95^0, 110^0, 126^0 and 135^0. In this case we can write

$100^0 > 70^0$ or $95^0 < 135^0$ which simply means that 110^0 is warmer than $70^0$and that 95^0 is cooler than 135^0. We can also write for example $95^0-70^0 = 135^0 - 110^0$, since equal temperature differences are equal in the sense that the same amount of heat is required to raise the temperature of an object from 70^0 to 95^0 or from 110^0 to 135^0. On the other hand, it would not be meaningful if we said that 126^0 is twice as hot as 63^0, even though $126^0 \div 63^0 = 2$, as it does not make sense. To show the reason, we have only to change to the centigrade scale, where the temperature becomes 5/9 (126-32) $=52^0$, the temperature 63^0becomes 5/9 (63-32) $=17^0$ and now the first figure is more than three times the second. This difficulty arises from the fact that Fahrenheit and centigrade scales both have artificial origins (zeros) i.e., the '0' of neither scale is indicative of the absence of whatever quantity (temperature) we are trying to analyze i.e., no absolute zero for temperature exits. The other example of the interval scale is humidity.

In some situations, characteristics can be employed quotients (multiplication or division) in addition to setting up dissimilarities and forming differences (i.e., when we may perform all the customary operations of mathematics), we refer to such data as ratio data. In this logic, ratio data includes all the usual measurement (or determinations) of length, height, money, weight, volume, area, pressures etc.

The above stated distinction between nominal, ordinal, interval and ratio data is important for the understanding about the nature of a set of data and may determine the use of particular statistical techniques. A researcher has to be clear about the measurement scale while measuring characteristics or properties of objects or of abstract concepts.

Table1: Characteristics and comparison of quantitative and qualitative data

Basis for classification	**Quantitative**	**Qualitative**
Meaning	It can be measured and expressed numerically	Objects is based on attributes and properties and can be measured numerically
Research methodology	Conclusive	Exploratory
Analysis	Statistical	Non-statistical
Collection	Structured	Unstructured

Examples	Total number of species in ecosystem, number of birds in any given area, number of plankton in pond ecosystem, height, weight of leaf, soil pH, etc.	Specific species in ecosystem such as plankton species (bacillariophyceae, cyanophyceae, rhodophyceae, chlorophyceae), intelligence, honesty etc.

5.3. Classifications of Measurement Scales

On the basis of above account, four scales of measurement can be considered in terms of their mathematical properties. The most widely used classification of measurement scales are discussed below:

1. Nominal scale
2. Ordinal scale
3. Interval scale
4. Ratio scale

1. **Nominal Scale:** It is simply a system of assigning number as symbols to events or outcomes in order to label them. The common example of nominal scale is the assigning numbers to all individuals of trees in a small patch of forest in order to identify the trees of the forest. Such numbers cannot be measured to be related with ordered scales for their character rather the numbers are just convenient label for the particular class of events or outcome and as such have no quantitative value. Nominal scales provide convenient methods of keeping track of individuals, objects and events. No further analysis of assigned numbers to the objects or events can be meaningful. For example, the color of stem of various tree species says 1,2,3 and 4 has been represented for brownish, black, white, and yellowish. In this example, one cannot usefully average the numbers for the colors of stem. The average calculated will not be meaningful. Neither one can usefully compared the numbers assigned to one group with the numbers assigned to another. The counting of members in each group is the only possible arithmetic operation for a nominal scale. Accordingly, nominal scale is useful only for mode as the measure of central

tendency. There is no measure for nominal scale for dispersion (Standard deviation, quartile deviation, mean deviation, range etc.) for nominal scales. With nominal data, if test of association is required for analysis, then chi-square test can be utilized. Correlation in nominal data is measured through the contingency coefficient.

Nominal scale is least powerful level of measurement. The scale does not indicate order or distance relationship and also has no arithmetic origin. A nominal scale simply describes differences between things by assigning them to categories. Therefore, nominal data is counted data. The scales cannot capture information for varying degrees of attitude, skills, understanding etc.. Nominal scales are useful for variety of purposes and are widely used in surveys and other ex-post facto research when data is being classified by major sub groups of the population.

2. **Ordinal scale:** The ordinal scale signifies the order in the objects or events. The scale places events in order, but there is no attempt to make the intervals of the scale equal by any rule. Rank order represents ordinal scales and frequently used in qualitative research. A rank of dominance of species in a forest involves the use of an ordinal scale. Researcher has to be very careful in making statement about scores based on ordinal scales. For example, if the dominance of tree has been assigned a value of 3; co-dominance as 2 and sub-dominance as 1 then it cannot be deduced that dominance is three times as good as sub-dominance. This statement would make no sense. Ordinal scales only permit the ranking of items is some order say highest to lowest or lowest to highest. Ordinal scale measures have no absolute values, and the real differences between adjacent ranks may not be identical.

   Therefore, ordinal scale implies a statement of “greater than” or “less than” (an equality statement is also acceptable) without our being able to state how much lesser or greater. There may be difference between rank 1 and 2 may be more or less than the difference between ranks 5 and 6. Median is best measure of central tendency for ordinal scale. A percentile or quartile measure can be used for measuring dispersion. Correlations can be estimated through Spearman's Rank

correlation. Measures of statistical significance can be adjudged by the non-parametric methods for ordinal scale.

3. **Interval Scale:** In the interval scale, the intervals are adjusted in terms of some rule that have been established as a basis for making the units equivalent. The units are equal only in so far as one accepts the prediction on which the rule is based. It can have an arbitrary zero, but it is not possible to determine an absolute zero or the unique origin. The main limitation of this scale is the lack of a true zero; it does not have the capacity to analyze the complete absence of a characteristic. Fahrenheit scale for temperature is a perfect example of an interval scale and shows similarities in what one can and cannot do with it. Researchers of Environmental studies generally measure temperature of atmosphere as well as aquatic ecosystems. One can say that an increase in temperature from 30^0 to 40^0 includes the same increase in temperature as an increase from 60^0 to 70^0, but one cannot say that the temperature as an increase $60^0$is twice as warm as the temperature of 30^0. As both numbers are dependent on arbitrarily zero at the temperature as freezing point of water. The ratio of the two temperatures, 30^0 and 60^0, means nothing because zero is an arbitrary point.

   This scale provides more powerful measurement than ordinal scales. Interval scales also incorporates the concept of equality of interval. Mean is the suitable measure of central tendency, while standard deviation is most widely used measure of dispersion. “t” test and F-test are appropriate tests for statistical significance for ordinal data.

4. **Ratio Scale:** It has an absolute or true zero of measurement. We can conceive of an absolute zero of length and in the same way we can conceive of an absolute zero of time. For example, the zero point on a centimeter scale shows the complete absence of length or height. But an absolute zero of temperature is theoretically unobtainable and it remains a concept existing only with investigators. The number of dead and disease tress in a forest and the number of leaves in a tree represents scores on ratio scales. Both these scales have absolute zeros. With ratio scales involved one can make statements like growth of Poplar trees was twice as good as

that of growth of the Shisham trees. The ratio involved does have significance and facilitates a kind of comparison which is not possible in case of an interval scale.

This scale represents the actual amount or value of the variables under consideration. Measures of physical properties such as weight, height, distance etc. are some common examples. Usually, all statistical techniques are applicable for ratio scales and all manipulations that one can carry out with real numbers can also be worked out with ratio scale values. Geometric and harmonic means can also be used as measures of central tendency and coefficient of variation may also be computed.

Therefore, proceeding from the nominal scale to ratio scale (the most precise), relevant information is obtained increasingly. If the nature of the variables permits, the researcher should use scale that provides the most clear-cut description. Researchers in physical sciences have the advantage to describe variables in ratio scale form but the behavioral sciences are generally limited to describe variables in interval scale form.

5.4. Qualities of Goodness of Measurement Scales

A measurement scale has to have certain desirable qualities to judge their goodness in measuring the characteristics under study. These qualities are summarized in Fig-1 and described below:

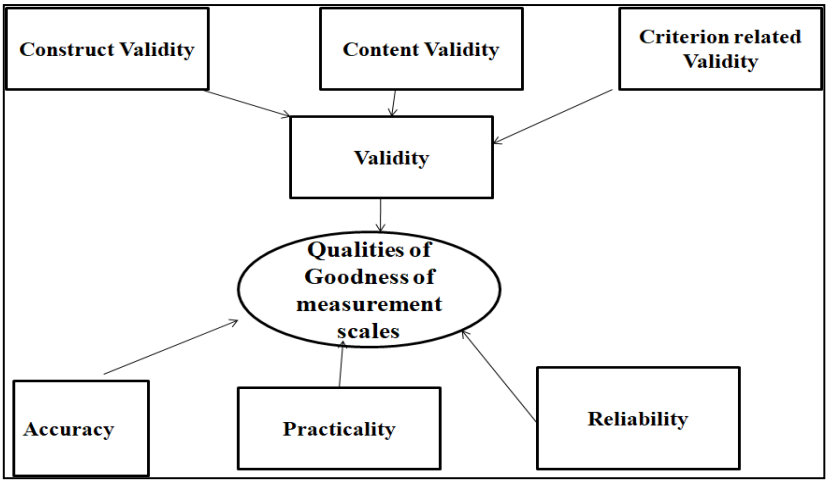


Fig-1: Showing qualities of goodness of measurement scales

1. **Validity:** Validity is the most critical quality and indicates the degree to which an instrument measures the different parameters. In other words, validity is the extent to which differences found with a measuring instrument and reflect true differences among those being analyzed. Validity determination is difficult in the absence of non-availability of direct confirming knowledge. Validity can be determined by seeking other relevant evidence that confirms the answers we have found with our measuring tool. The proof often depends upon the nature of the research problem and the judgment of the researcher. But one can certainly consider as three types of validity in this association.

    i. **Content validity:** It is the extent to which a measuring instrument provides adequate coverage of the topic under research. If the instrument contains a representative sample of the universe, content validity is good. Its determination is primarily judgmental and intuitive. It can also be measured by experts, who can judge how well the measuring instrument meets the standards, but there is no numerical way to express content validity.

    ii. **Criterion-related validity:** It relates to our ability to assume some outcome or analyze the existence of some current condition. This form of validity reflects the success of measures used for some empirical estimating objective. The concerned criterion must possess the following qualities:

        a. **Relevance:** (A criterion is relevant if it is defined in terms what we judge to be the proper measure)

        b. **Freedom from bias:** (freedom from bias is attained when the criterion gives each subject an equal opportunity to score well)

        c. **Reliability:** (A reliable criterion is stable)

        d. **Availability:** (The information specified by the criterion must be available)

In fact, a criterion-related validity is a broad term that actually refers to:

a. Predictive validity

b. Concurrent validity

The predictive validity refers to the usefulness of a test in assuming some future performance whereas the concurrent validity refers to the usefulness of a test in a closely relating to other measures of known validity. Criterion –related validity is expressed as the "coefficient of correlation" between test scores and some measure of future performance or between test scores and scores on another assess of known validity.

iii. **Construct validity:** It is most complex and abstract. A measure is said to possess construct validity to the degree that it confirms to assumed correlation with the other theoretical propositions. It is degree to which scores on a test can be accounted for by the explanatory constructs of a sound theory. For measuring construct validity, we associate a set of other propositions with the results received from using our measurement instrument. If measurements on our devised scale correlate in an expected way with these other propositions, we can conclude there is some construct validity.

If the above stated criteria and tests are met with, we can say that our analyzing or measuring instrument is valid and will result in correct measurement; otherwise, we shall have to look for more facts and/or resort to exercise of judgment.

2. **Reliability:** The test of reliability is very important test of measurement. A measuring instrument is reliable if it provides consistent results. Reliable measuring instrument does contribute to validity, but a reliable instrument may not be a valid instrument. For example, a scale that consistently overweighs objects by 5kg, is a reliable scale, but it does not give a valid measure of weight. But the other way is not true i.e., a valid instrument is always reliable. Accordingly, reliability is not as valuable as validity, but it is simpler to assess reliability in comparison to validity.

   Two aspects of reliability are stability and equivalence. The stability aspect is concern with securing consistent results with repeated measurement by the same person and with the same instrument. We generally determine the degree of stability by comparing the results of repeated analysis. The equivalence aspect considers how much error may get introduced by several investigators or various samples of the items being studied. A good method to test for the equivalence of measurement

by two researchers is to compare their results of the same events. Reliability can improve in the following two ways.

i. By standardizing the situations under which the measurement takes place i.e., researcher must make sure those external sources of variation are minimized to the extent possible, this will improve stability aspect.

ii. By carefully designed directions for the measurements with no variation from group to group, by using trained and inspired persons to conduct the research and also by broadening the sample of items used, will improve equivalence.

3. **Practicality:** Practicality of a measuring instrument may be analyzed in terms of economy, convenience and interpretability. The measuring instrument such as questionnaire should be economical, convenient and interpretable. Economy must be decided based on the objective of the investigation and the length of the question of questionnaire. A convenience test suggests that the measuring instrument should be easy to administer and utilizes less time. For this purpose, one should give due attention to the proper layout of the measuring instrument. For example, a questionnaire, with clear instructions is certainly more effective and easier to complete than one which lacks these features. Interpretability consideration is especially important when persons other than the developer of the test are to understand the results. The measuring instrument, in order to be interpretable, must be complemented by:

   a. Full instructions for administering the test
   b. Scoring keys
   c. Confirmation about the reliability
   d. Guides for using the test and for interpreting results.

4. **Accuracy:** The characteristics of accuracy of a measurement scale means it should be a true representative of the observation of underlying characteristics. For instance, measuring with an “inch” scale will provide accurate value only up to one

eight of inch, while measuring with "centimeter" scale will provide more correct value.

5.5. Sources of Error in Measurement

Measurement should be accurate and unambiguous in research and, is often not met with in entirely. The researcher must be aware about the sources of error in measurement. The following are some possible sources of error in measurement.

a. **Respondent:** Respondent is important for measurement while surveying. While surveying the respondent should not be uneasy and non-interested. The lack of attention and concentration of respondents have severe consequences for the responses of the questions of the questionnaire. Temporary or transient factors like fatigue, boredom, anxiety etc. may limit the ability of the respondent to respond accurately and fully.

b. **Situation:** Situation is important for right measurement. A condition which places a strain on interview can have serious effects on the interviewer-respondent rapport. For example, if someone else is present while interviewing, his presence can distort responses of the respondents by joining in or merely by being present. If the respondent feels that mystery is not assured, he may be unenthusiastic to express certain feelings.

c. **Measurer:** The interviewer may misrepresent responses by rewording or reordering questions. His behavior, style and looks can encourage or discourage certain relies from respondents. Careless mechanical processing may misrepresent the findings. Errors may also creep due to incorrect coding, faulty tabulation and or statistical calculations, mainly during data analysis.

d. **Instrument:** In analytic work, errors may also happen because of the defect in measuring instruments. The use of complex words, beyond the comprehension of the respondent, ambiguous meanings, poor printing, insufficient space for replies etc. are led to defective measuring instrument. The defective instrument may result in measurement errors. Another type of instrument deficiency is the poor sampling of the population of items.

In empirical work, researcher must calibrate the instruments, which is used for generation for the data. Various instruments are used in environmental studies researches such as spectrophotometer, BOD incubator, dissolved oxygen meter, turbidity meter, TDS meter, respirable dust sampler, anemometer, thermo-hygrometer etc...

Researcher must know that correct measurement depends on successfully meeting all of the associated issues and problems listed above. He should, to the degree possible, try to abolish, defuse or otherwise deal with all the possible sources of error so that the final results may not be defective.

5.6. Techniques of developing measurement tools

The technique of developing measurement tools engrossed a four-stage process, consisting of the following;

a. **Concept development:** Concept development means that the researcher should arrive at thoughtful of the major concepts pertaining to his study. Concept development is more obvious in theoretical studies than in the more pragmatic research, as the fundamental concepts are often already prepared in pragmatic researches.

b. **Specification of concept dimensions:** The second step requires the researcher to specify the dimensions of the concepts. This task may either be accomplished by deduction i.e., by adopting a more or less intuitive approach or by empirical correlation of the individual dimensions.

c. **Selection of indicators:** Indicators are essentially required for measuring each concept element. Indicators are definite questions, scales, or other devices by which respondent's knowledge, opinion, expectations etc. are analyzed. As there is infrequently a perfect measure of concept the researcher should consider several alternatives for the study. The use of more than one indicator gives stability to the scores and it also support their validity.

d. **Formation of index:** The last step is that of combining the various indicators into an index formation. We may require to combine all indicators into a single index if numerous dimensions of a concept or different measurements of a dimensions is

considered simultaneously. One simple method for getting an overall index is to provide scale values to the responses and then sum up the corresponding scores. Such an overall index would provide a better analysis tool than a single indicator because of the fact that an "individual indicator has only a probability relation to what we really want to know" and unable to capture the importance of the research. This way we must attain an overall index for the different concepts concerning the research.

5.7. Scaling

In research we fairly often face measurement problem, especially when the concepts to be analyzed and measured are complex and abstract and researcher do not have the standardized measurement tools. Alternatively, we can say that while measuring attitudes and opinions, we face the problem of their valid measurement. Researcher faces similar problem, of course in a lesser degree, while measuring physical or institutional concept. We should study some procedures which may enables us to analyze abstract concepts more accurately. Scaling has been defined as a "procedure for the assigning of numbers to a property of objects in order to impart some of the characteristics of numbers to the properties in question."

Scaling explains the procedure of assigning numbers to various degree of opinion, attitude and other concepts. This can be achieved by two ways viz.

1. Making a judgment about some characters of an individual and then placing him directly on a scale that has been defined in terms of those characteristics.
2. Constructing questionnaire in such way that score of individual's responses assigns him a place on a scale.

It may be said here that a scale is a continuum, consisting of the highest point (in terms of some characteristics e.g.: preference, favorableness etc.) and the lowest point along with several intermediate points between these two extreme points, as per the choice or requirement of the research. These scale point positions are associated to each other that when the first point happens to be the highest point, the second point shows a higher degree in terms of a given characters as compared to the third point and the third point indicates a

higher degree as compared to the fourth point and so on. Numbers for calculating the distinctions of degree in the attitudes or opinions are, thus, assigned to individuals corresponding to their scale-positions. All this is better understood when we talk about scaling techniques. Hence, the term "scaling" is used to the procedures for attempting to determine quantitative measures of subjective abstract concepts.

5.8. Scale Classification bases

The Number assigning procedures or the scaling procedures may be broadly classified on one or more of the following bases:

1. **Subject orientation:** Under the subject orientation, a scale can be designed to determine characteristics of the respondent who completes it, which is presented to the respondent. In respect of the former, we assume that the stimuli presented are sufficiently homogeneous so that the between-stimuli variation is small as compared to the variation among respondents. In latter approach, we ask the respondent to judge some specific object in terms of one or more dimensions and we assume that the between respondent's variation will be small as compared to the variation among the different stimuli presented to respondents for judging.
2. **Response form:** In the response form**,** we may classify the scales as categorical and comparative. Former are also called as rating scales. These scales are used when a respondent scores some object without direct reference to other objects. In comparative scales which are also called as ranking scales, the respondent is asked to compare two or more objects. In this sense the respondent may state that one object is superior to other or that three models of pen rank in order 1,2 and 3. The real meaning of ranking is, in fact, a relative comparison of certain property of two or more objects.
3. **Degree of subjectivity:** In the degree of subjectivity, the data may be based on whether we measure subjective personal preferences or simply make non-preferences judgment. In the former case, the respondent is asked to select which person he favors or which solution he would like to see employed, whereas in the latter case he is simply asked to judge which person is more effective in some

characteristic or which solution will take fewer resources without reflecting any personal preference.

4. **Scale Properties:** On the basis of scale properties, we can categorize scales as: nominal, ordinal, interval and ratio scales. Nominal scales only classify without indicating order, distance or distinctive origin. Ordinal scales indicate magnitude relationship of "more than" or "less than", but indicate no distance or unique origin. Interval scales have both order and distance values, but no unique origin. Ratio scale has all these characters.

5. **Number of dimensions:** In respect of this basis, scales may be categorized as "uni-dimensional and multidimensional scales". Under the former we determine only one characteristic of the respondent or object, whereas multidimensional scaling recognizes that an object might describe better by using the concept of an attribute space of 'n' dimensions, rather than a single dimension continuum.

6. **Scale's construction techniques:** There are five main techniques by which scales can be developed.

   i. **Arbitrary Approach:** It is an approach where scales are developed on adhoc basis. This approach is the most widely used approach, and presumed that such scales determine the concepts for which they have been designed, although there is very little evidence to support such an assumption.

   ii. **Consensus Approach:** In this approach, a panel of judges, who are expert of the subject, assesses the items selected for inclusion in the instrument in terms of whether they are applicable to the topic theme and unambiguous in implication.

   iii. **Item Analysis approach:** In analysis approach, a number of individual items are developed in to a test which is given to a group of respondents. After run the test, the total scores are computed for everyone. Individual items are then measure to determine which items discriminate between persons or objects with high total scores and those with low scores.

iv. **Cumulative Scales:** These are selected on the basis of their conforming to some ranking of items with ascending and descending discriminating power. For example, in such a scale the endorsement of an item showing an extreme position should also result in the endorsement of all items representation of a less extreme position.

v. **Factor Scales:** Factor scales may be constructed on the basis of interrelations of items which show that a common factor accounts for the relationship between items. This relationship is determined through factor analysis method.

5.9. Scaling Techniques

There are various scaling techniques; however, these are broadly classified into "comparative scaling techniques" and "non-comparative scaling technique". The former involves direct comparison of stimulus objects. Comparative scale data are explained in relative terms and are measured on ordinal scale. Comparative scaling technique is a non-numeric scaling technique as ordinal data cannot be used for numeric operations. These techniques are very easy to understand and apply. These techniques guide the respondent to choose between the stimulus objects. For example, respondents are asked to figure likeness to use toothpaste of brand "A" or brand "B". The respondents have to choose one out of A or B even if there is very small difference in the liking of the two brands. The disadvantage of comparative scales is the inability to generalize beyond the stimulus objects. For instance, if we want to compare a third brand of toothpaste "C" with the previous ones, we have to conduct a new study.

On the other hand, in non-comparative scales, each object is scaled independently of the others. For example, the respondents are asked to give preference scores on a 1 to 6 scale to brand "A" of the toothpaste. Here 1=no* preferred at all and 6= highly preferred. Similar scores can be obtained for brand "B" and "C". Because of the numeric data and wide applications, non-comparative scales are widely used in studies.

5.10. Multi-dimensional Scaling

Multi-dimensional Scaling (MDS) is relatively more complex scaling device, but with this sort of scaling one can scale objects, individuals or both with a less information. This scaling can be characterized as a set of procedures for portraying perceptual or affective dimensions of substantive interest. It is used when all the variables (whether metric or non-metric) in a study are to be examined simultaneously and all such variables happen to be independent. The underlying assumptions in MDS is that the respondent perceive a set of objects as being more or less similar to one another on a number of usually uncorrelated dimensions. Through MDS techniques one can represent geometrically the location and interrelationship among a set of points. In fact, the technique attempts to locate the points, given the information about a set of inter-point distances, in space of one or more dimensions such as to best summarize the information contained in the inter-point distances. The distance in the solution space then optimally reflect the distances contained in the input data. For instance, if objects say X and Y are thought of by the respondent as being simpler as compared to all other possible pairs of objects. MDS techniques will position objects X and Y in such a way that the distance between them in multidimensional space is shorter than that between any other two objects.

Two approaches of scaling i.e., the metric approach and the non-metric approach are applicable in the context of MDS, while attempting to make a space containing m points such that m (m-1)/2 inter-point distances reflect the input data. The metric approach to MDS treats the input data as interval scale data and solves applying statistical methods for the additive constant which minimizes the dimensionality of the solution space. This approach utilizes all the information in the data in obtaining a solution. The data are often obtained on a bipolar similarity scale on which pairs of objects are rated one at a time. If the data reflect exact distances between real objects in an r-dimensional space, their solution will reproduce the set of inter-point distances. But as the true and real data are rarely available, we require random and systematic procedures for obtaining a solution. Generally, the judged similarities among a set of objects are statistically transformed into distances by placing those objects in a multidimensional space of some dimensionality.

The non-metric approach first gathers the non-metric similarities by asking respondents to rank order all possible pairs that can be obtained from a set of objects. Such non-metric data is then transformed in to some arbitrary metric space and then the solution is obtained by reducing the dimensionality. The non-metric became popular during sixties with the invent of high-speed computers to generate metric solutions for ordinal input data.

The MDS techniques, in fact, do away with the need in the data collection process to specify the attribute (s) along with the several brands, say of a particular product, may be compared as ultimately the MDS analysis itself reveals such attribute (s) that presumably underlie the expressed relative similarities among objects.

In spite of all the merits stated above, the MDS is not widely used because of the computation complications of MDS. Many of its methods are quite complicated and methodological in terms of both the collection of data and the subsequent analyses. However, some progress has been achieved during the last few years in the use of non-metric MDS in the context of research problems.

5.11. Deciding the scale

Before deciding the scales following factors should be taken in to account:

1. **Data type:** Each data type can be used for only selected type of analysis. For example, nominal and ordinal data cannot be used for all basic statistical analyses. The research should know in advance what kind of analysis will be conducted and what type of data will be required for the analysis. The decision of scale will depend on that data type.
2. **Neutral Category:** If it is likely to get neutral or impartial response from some respondents, the neutral category may also be used.

5.12. Summary

In this unit we have discussed various aspects of Measurement of scaling. So far you have learnt that:

- Measurement is defined as a process of associating numbers or symbols to data obtained in a research study. These observations or data could be qualitative or quantitative in nature.
- Quantitative data measured and expressed numerically; its research methodology is conclusive. Some examples of quantitative data ate Total number of species in ecosystem, number of birds in area forest patch, number of plankton in pond ecosystem, height of tree, weight of mice etc.
- Qualitative data is based on attributes and properties; its research methodology is exploratory. Some common examples of qualitative data are Specific species in ecosystem such as plankton species (Bacillariophyceae, cyanophycean, Rhodophyceae, Chlorophyceae), intelligence, honesty, color of leaves, shape of crown etc.
- The classification of measurement scales is nominal scale, ordinal scale, interval scale and ratio scale. The nominal scale is a system of assigning number symbols to events in order to label them. The common example of nominal scale is the assignment of numbers to basketball players in order to identify them. Such numbers cannot be measured to be related with ordered scales for their order; the numbers are just convenient label for the particular class of events and as such have no quantitative value.
- Ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Ordinal scale implies a statement of “greater than” or “less than”. Interval scale has been established as a basis for making the units equivalent. The units are equal only in so far as one accepts the prediction on which the rule is based. It can have an arbitrary zero, but it is not possible to determine for them, an absolute zero or the unique origin. The main limitation of this scale is the lack of a true zero; it does not have the capacity to

analyze the complete absence of a characteristic. Fahrenheit scale for temperature measurement is a perfect example of an interval scale and shows similarities in what one can and cannot do with it.

- Ratio Scale has an absolute or true zero of measurement. The term "absolute zero" is not as accurate as it was once believed to be. This scale represents the actual number of variables. Measures of physical properties such as weight, height, distance etc. are some common examples. Usually, all statistical techniques are usable with ratio scales and all manipulations that one can carry out with real numbers can also be worked out with ratio scale values.
- A measurement scale has to have certain desirable qualities to judge their goodness in measuring the characteristics under study. These qualities are validity (content validity, criterion-related validity, predictive validity, concurrent validity), reliability, practicality and accuracy.
- There are some possible sources of error in measurement. These sources of errors may be respondent, situation, measurer, and instrument.
- The technique of developing measurement tools engrossed a four-stage process, consisting of the concept development, specification of concept dimensions, selection of indicators and formation of index
- Scaling has been defined as a "procedure for the assigning of numbers to a property of objects in order to impart some of the characteristics of numbers to the properties in question."
- The number assigning procedures or the scaling procedures may be broadly classified on one or more of the bases such as subject orientation, response form, degree of subjectivity, scale properties, number of dimensions and scales construction techniques.
- Scales can be developed by five main techniques. These techniques are arbitrary approach, consensus approach, item analysis approach, cumulative sales and factor scales.

- There are two broad categories of scaling techniques viz. into "comparative scaling techniques" and "non-comparative scaling technique". The former involves direct comparison of stimulus objects. Comparative scale data are explained in relative terms and are measured on ordinal scale. Comparative scaling technique is a non-numeric scaling technique as ordinal data cannot be used for numeric operations. On the other hand, in non-comparative scales, each object is scaled independently of the others.
- Multidimensional scaling is relatively more complex scaling device, but with MDS one can scale objects, individuals or both with a least information.
- Before deciding the scales two factors should be taken in to account. These factors are data type and neutral category.

TERMINAL QUESTIONS

1. (a) Fill in the blank spaces with appropriate words.

There are various scaling techniques; however, these are broadly classified into "……scaling techniques" and "…….. scaling technique". The former involves direct comparison of stimulus objects. Comparative scale data are explained in relative terms and are measured on …….. scale. It is a non-numeric scaling technique as ordinal data cannot be used for …….. operations. These techniques are very easy to understand and apply. These techniques force the …….. to choose between the stimulus objects. For example, respondents are asked the like to use toothpaste of brand "A" or brand "B". The respondents have to choose one out of A or B even if there is very small ……in their liking of the two brands. The disadvantage of …….. …….. is the inability to generalize beyond the stimulus objects. For instance, if we want to compare a third brand of toothpaste "C" with the previous ones, we have to conduct a new study. On the other hand, in …… scales, each object is scaled independently of the others. For example, the respondents are asked to give preference scores on a 1 to 6 scale to brand "A" of the toothpaste. Here 1=no* preferred at all and 6= highly preferred. Similar scores can be obtained for brand "B" and "C". Because of the numeric data and wide applications, non-comparative scales are widely used in studies.

2. (a) Discuss the quantitative and qualitative data.

   (b) What is classification of measurement scales?

3. (a) Describe the goodness of measurement scales.

   (b) Give the four qualities of goodness of measurement scales

4. (a) Write about sources of error in measurement.

5. (a) Discuss the techniques of developing measurement scales.

   (b) What is scaling? Explain in brief.

   (c) Write a short note on deciding the scales.

6. (a) Fill the blank spaces with appropriate words.

   The test of reliability is very important test of measurement. A measuring instrument is reliable if it provides consistent results. Reliable measuring instrument does contribute to……., but a reliable instrument requires not be a valid instrument. For example, a scale that consistently overweighs objects by 5kgs is a reliable scale, but it does not give a valid measure of weight. But the other way is not true i.e., a valid instrument is always …….Accordingly, reliability is not as valuable as validity, but it is simpler to assess reliability in ……to validity. If the …….is satisfied by an instrument, when while using it researcher can be confident that the transient and situational; factors are not inquisitive. Two aspects of reliability are……and ……. and these two aspects deserve special mention. The stability aspect is concern with securing consistent results with repeated measurement of the same person and with the same instrument.

   (b) Qualitative data cannot be computed (Yes/No)

   (c) Conclusive research methodology is used for (Quantitative data/Qualitative data)

   (d) What do you understand by scale classification bases?

7. (a) Describe the multi-dimensional scaling

ANSWERS

1. Comparative, non-comparative, ordinal, numeric, respondent, difference,

comparative scales, non-comparative,

2. (a) see section 5.2

   (b) See section 5.3

3. (a) See section 5.4

   (b) See section 5.4

4. (a) See section 5.5

5. (a) See the section 5.6

   (b) seethe section 5.7

   (c) See section 5.11

6. (a) Validity, reliable. comparison, reliability, stability, equivalence,

   (b) Yes

   (c) Quantitative data

   (d)See the section 5.8

7. (a) See the section 5.10

Unit 6: Data collection: Introduction, Collection of primary and secondary data, selection of appropriate method for data collection, case study method

Unit Structure

6.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is data collection?
- What are primary data and secondary data?
- About the data collection methods.
- About case study method.

6.1. Introduction

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem. It is one of the most important steps of research. Data collection methods can be divided into two categories: Primary methods of data collection and secondary methods of data collection. Researchers collect the data in numerous ways. After the selection of research problem, researchers gone through review of literature and then collect the data. In this unit you will learn about the data collection of primary and secondary data, selection of appropriate method for data collection and case study method.

6.2. Data collection

The task of data collection begins after a research problem has been defined and research design /plan chalked out. While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz. primary and secondary data. The primary data are those which are collected afresh and for the first time. This data is original in character. On the other hand, secondary data are those which have been already collected by someone else and which sort of data he/she would be using for his/her study and accordingly he/she will have to select one or the other method of data collection. The methods of collecting primary and secondary data the nature of data collection work is merely that of combination.

6.3. Collection of Primary Data

We collect primary data during the course of doing experiments in an experimental research. An experiment refers to an investigation in which a factor or variable under test is isolated and its effect measured. In an experiment the investigator measures the effects of an experiment which he conducts intentionally. However, in case we do research of the descriptive type and perform surveys, whether sample survey or census survey, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interview. Survey refers to the method of securing information concerning a phenomenon under study from all or a selected number of respondents of the concern universe. In a survey, the investigator examines those phenomena which exist in the universe independent of his/her action. The difference between an experiment and a survey can be depicted as under:

↑ Survey Can be studied through		↑ Experiments Determine
Possible relationships between the data and the unknowns in the universe		
Economic	Psychological	Others

There are several methods of collecting primary data, particularly in surveys and descriptive researches. Some important methods of primary data collection are summarized in fig-1 and also discussed below:

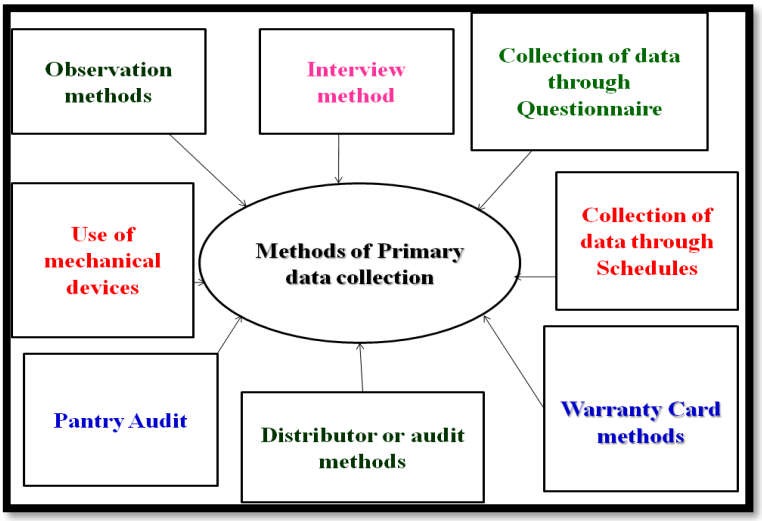


Fig-1: Showing Primary data collection methods

Observation method: The observation method is the most commonly used method specially in studies relating around us, but this sort of observation is not scientific observation. Observation becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability. Under the observation methods, the information is sought by way of investigator's own direct observation without asking from the respondents. For example, in study related to consumer behaviour, the investigator instead of asking the brand of wrist watch used by respondent, may himself look at the watch. The main advantages of this method are that subjective bias is eliminated, if observation is done accurately. Secondly, the information obtained under this method relates to what is currently happening it is not complicated by either the past behaviour or future intentions or attitudes. Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happen to be the case in the interview or the questionnaire method. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feeling for one reason or the other.

However, observation method has various limitations. Firstly, it is expansive method, secondly the information provided by this method is very limited, thirdly sometimes

unforeseen factors may interfere with the observational task. At time, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

There are several merits of the participant types of observation.

i. The researcher is enabled to record the natural behaviour of the group.
j. The researcher can even gather information which could not easily be obtained if he observes in a disinterested fashion.
k. The researcher can even verify the truth of statement made by informants in the context of questionnaire or a schedule.

Interview Method: The interview method of collecting data involves presentation of oral verbal stimuli and reply in term of oral verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews.

Personal interview: Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons. This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation. In case of direct personal investigation, the interviewer has to collect information personally from the sources concerned. He has to be on the spot and has to meet people from whom data have to be collected. This method is particularly suitable for intensive investigations. But in certain cases, it may not be possible or worthwhile to contact directly the person concerned or on account of the extensive scope of enquiry, the direct personal investigation technique may not be used.

The method of collecting information through personal interview is generally carried out in a structured way. This type of interview is called as structured interview. Such interviews involve the use of set of predetermined questions and of highly standardized techniques of recording. As against it unstructured interviews are characterized by flexibility of approach to questioning. Unstructured interviews do not follow a system of pre-determined questions and standardized techniques of recording information. Focused interview is other type of interview in which focus attention on the given experience of the respondents and its effects. Such interviews are used generally in the development of hypothesis and constitute a major type of unstructured interviews.

The clinical interviews are concerned with broad underlying feelings or motivations or with the course of individual's life experience. In case of non-directive interview the interviewer's function is simply to encourage the respondent to talk about the given topic with bare minimum of direct questioning. The interviewer often acts as a catalyst to a comprehensive expression of the respondents' feelings and beliefs and of the frame of reference within which such feelings and beliefs takes on personal significance.

The main advantages of interview methods are:

More information and that too in greater depth can be obtained.

Interviewer by his own skill can overcome the resistance. There is greater flexibility under this method as the opportunity to restructure question is always there, especially in case of unstructured interviews.

1. Observation method can as well be obtained easily under this method.
2. samples can be controlled more effectively as there arises no difficulty of the missing returns non-response generally remains very low.
3. The interviewer can usually control which person will answer the questions. This is not possible in mailed questionnaire approach.
4. The interviewer can collect supplementary information about the respondents.

There are also certain weaknesses of interview method which are given below:

1. It is very expansive method, especially when large and widely spread geographical sample is taken.
2. There remain possibilities of the bias of interviewer as well as that of the respondent, there also remains the headache of supervision and control the interviewers.
3. This method is relatively more time consuming specially when sample is large and recall upon the respondents are necessary.

Telephonic interviews: This method of collecting information consists in contacting respondents on telephone itself. It is not very widely used method. However, this method plays important part in industrial surveys, particularly in developed regions. There are some advantages of this method which are given below:

1. It is more flexible in comparison to mailing method.
2. It is faster than other methods.

3. It is cheaper than personal interview methods.
4. Recall is easy, callbacks are simple and economical.
5. Replies can be recorded without causing embarrassment to respondents.
6. Interviewer can explain requirements more easily.
7. No field staff is required.

Some of the disadvantages of this method are given below:

1. Little time is given to respondents.
2. Surveys are restricted to respondents who have telephone.
3. It is not suitable for intensive surveys where comprehensive answers are required to various questions.
4. Possibility of the bias of the interviewer is relatively more.

Collection of Data through Questionnaire: This method of data collection is quite popular, particularly in case of big enquiries. It is being adopted by private individuals, research scholars, private and public organizations, NGOs and even by governments. In this method a questionnaire is sent to the person concerned with are quiet to answer the questions and return the questionnaire. A questionnaire consists a number of questions printed or typed in a definite order on a form or set of forms. The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own. The method of collecting data by mail mailing the questionnaires to respondents is most extensively employed in various economic and business surveys. The advantages claimed on behalf of this method are given below:

1. There is low cast even when the universe is large and is widely spread geographically.
2. It is free from the bias of the interviewer; answers are in respondents own words.
3. Respondents have adequate time to give well thought out answers.
4. Respondents, who are not easily approachable, can also be reached conveniently.
5. Large samples can be made use of and thus can be made more dependable and reliable.

The main disadvantages of this method are given below:

1. Low rate of return of the duly filled in questionnaires, bias due to no response is often indeterminate.
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may lost once it is sent.
4. There is inbuilt flexibility because of the difficulty of amending the approach once questionnaire have been dispatched.
5. There is also be possibility of ambiguous replies or omission of replies altogether to certain questions, interpretation of omissions is difficult task.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

Collection of Data through Schedules: This method of data collection is very much like the collection of data through questionnaire with little difference which lies in fact that schedules (Performa containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose. These enumerators along with schedules go to respondents, put to them the questions from the Proforma in the order the questions are listed and record the replies in the space meant for the same in proforma. In certain situations, schedules may be handed over to respondents and enumerators may help them in recording their answers to various questions in the said schedules. Enumerators explain the aims and objectives of the investigation and also remove the difficulties which any respondents may feel in understanding the implications of a particular question or the definition or concept of difficult terms.

This method requires the selection of enumerators for filling up schedules or assisting respondents to fill up schedules and as such enumerator should be very carefully selected. The enumerators should be trained to perform their job well and the nature and scope of the investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule. Enumerator should be intelligent and must possess the capacity of cross-examination in order to find out the truth. Above all they should be honest, sincere, hardworking and should have patience and perseverance.

This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results. However, it is very expansive and is usually adopted in investigations

conducted by governmental agencies or by some big reputed organizations. Population census all over the world is conducted through this method.

Some other method of Data collection: There are some other important methods of data collection which are discussed below:

Warranty Card: Warranty cards are mostly postal sized cards. These warranty cards are used by the dealers of consumers to collect the data related to their products. Warranty cards may contain various columns such as durability, performance, prices etc. The information sought is printed in the form of questions on the warranty cards which is packed inside the envelope along with the purchased product. With a request to consumer to fill in the card and post it back to the dealers.

Distributor or store audits: Distributor or store audits are performed by distributors as well as manufactures through salesman at continuous basis. Distributors get the retail stores audited through salesman and use such data to predict market size, market share, seasonal marketing etc. The data is collected in cases not collected by questioning but through observations. The main advantage of this method of data collection is that it offers the most efficient method of evaluating the effect on sales of variations of different techniques of in store promotion.

Pantry Audits: This technique is used to predict consumption of the basket of goods at the consumer level. In this type of data collection, the investigator collects a inventory of types, quantities and prices of commodities consumed. Therefore, pantry audit data are recorded from the examination of consumer's pantry. The general objective in a pantry audit is to find out what types of consumers buy products and certain brands, the assumptions being that the content of the pantry accurately portray consumer's preferences. Quite often, pantry audits are supplemented by direct questioning relating to reasons and circumstances under which particular products were purchased in an attempt to relate these factors to purchasing habits. A pantry audit may or may not be setup as panel operation, since a single visit is often considered sufficient to yield an accurate picture of consumers' preferences.

Consumer panel: It is an extension of the pantry audit approach on regular basis. In this method a set of consumers are arranged to come to an understanding to maintain detailed daily records of their consumption and the same is made available to investigator on demand. We can also say that, a consumer panel is essentially a sample of consumers who

are interviewed repeatedly over a period of time. Mostly consumer panel are of two types viz. the transitory consumer panel and the continuing consumer panel. The former is set of measures the effect the particular phenomena usually such a panel is conducted on a before and after bases. Initial interviewers are conducted before the phenomena takes place to record the attitude of the consumer. A second set of interviews is carried out after the phenomena has taken place to find out the consequent changes that might have occurred in the consumer's attitude. In Later consumer often set up for an indefinite period with a view to collect data on a particular aspect of consumer behaviour over time, generally at periodic intervals or may be meant to serve as a general; purpose panel for researchers on variety of subjects. Such types of panels have been used in the area of consumer expenditure.

Use of mechanical device: It has been widely made to collect information by way of indirect means. Eye camera, pupillometric camera, psychogalvanometer, motion picture camera and audiometer are the principal devices so far developed and mostly used by modern big business houses, mostly in the developed world for the purpose of collecting the required information.

Projective techniques: These techniques have been developed by psychologist to use projection of respondents for inferring about underlying motives urges or intentions which are such that the respondent either resists to reveal them or unable to figure out himself. In these techniques, the respondent in supplying information tend unconsciously to project his one attitude or failing on the subject under study. These techniques play an important role in motivational researchers or in attitude surveys.

Depth Interviews: These interviews are design to discover underlying motives and desires and are generally used in motivational researches. These interviews are conducted to explore needs, desires and feelings of respondents. In other words, the objective to elicit unconscious as also other type of material relating specially to personality dynamics and motivations. Therefore, these interviews required great skills on the part of the interviews and at the same time involve considerable time. Unless the researchers have specialized training, depth interview should not be attempted.

Content analysis: It consists of analyzing the contents of documentary materials such as book, magazine, newspapers and the content of all of other verbal material which can be either spoken or printed. It is prior to 1940s was mostly quantitative analysis of documentary

materials concerning certain characteristics that can be identified and counted. But since 1950s it is mostly qualitative analysis concerning the general import or message of the existing documents.

6.4. Collection of Secondary Data

There are various methods and sources of secondary data collection which are summarized in fig-2 and also discussed below:

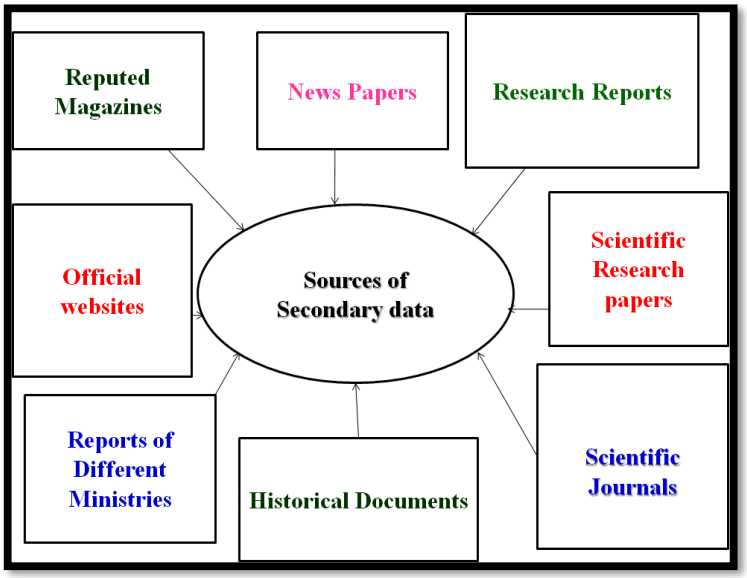


Fig-2: Showing different sources of secondary data collection

Secondary data is a type of data that has already been published in newspapers, magazines, journals, research reports, reputed websites etc. Whenever, researchers utilize this type of data then he/she seeks in to various sources from where he/she may collect the data. Secondary data usually publishes data and available in various publications of central, state and local governments, foreign researches, national and international research papers, reports prepared by research scholars, universities, ministries etc. Researcher should be very careful in using secondary data. He must make a minute observation because it is just possible that the SD may be inappropriate or may be insufficient in context of the research problem. By way of caution, the research scholar must see various characteristics of data such as reliability of data, suitability of the data and adequacy of the data. The reliability of data can be observed by the following important things.

1. Who collected the data?

2. What is the source of data?
3. Which method was applied during collection of data?
4. When they collected the data?
5. What was the purpose of collection of data?

After gone through these questions researcher should be satisfied on every aspect before using the data. Second important characteristic is suitability of data. As you know data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. If the available data found to be inappropriate, they should not be used by the research scholar. In the same way the object, scope and nature of original enquiry must also be studied. A third important characteristic is adequacy of data. If the level of accuracy achieved in data is found in insufficient for the aim of the present investigation, they will be considered insufficient and must not be utilized by the research scholar.

6.5. Selection of appropriate method for data collection

Researchers have to select appropriate method for data collection. Followings are the important factor should be considered during the selection of appropriate method.

Factor influencing the Methods for data collection: There are various important factors to consider when selecting data collection methods. Following factors should be kept in mind when researcher selects method for data collection. The first factor is nature, Scope and object of enquiry. Second factor is availability of funds. Third factor is time factor and fourth factor is precision required. The first factor i.e., nature, Scope and object of enquiry constitutes the most important factor affecting the choice of a particular method. The method selected should be such that it suits the type of enquiry that is to be conducted by the researcher. This factor is also important in determining whether the secondary data are to be used or the primary data are to be collected. The second factor is availability of funds for the research project. It determines to a large extent the method to be used for the collection of data. When funds at the disposal of the researcher are very limited, he will have to select a comparatively cheaper method which may not be as resourceful and effective as some other costly method. Finance, in fact, is a big constraint in practice and the researcher has to act within this limitation. The third factor is time. It has also to be taken into account in deciding a particular method of data collection. Some methods take comparatively more

time, whereas with others the data can be collected in a comparatively shorter time. The time at the disposal of the researcher, thus, affects the selection of the method by which the data are to be collected. Precision required is yet another important factor to be considered at the time of selecting the method of collection of data.

Case study method: The case study method is an accepted form of qualitative investigation and entails a careful and whole observation of a social unit, be that unit a person, a family, an institution, a cultural group etc. Case study is a method of study in depth relatively than span. The case study is more emphasis on the full investigation of a restricted number of events and their interrelationships. The case study is essentially an intensive investigation of the particular unit under consideration. As per H. Odum, "The case study method of data collection is a technique by which individual factor whether it is an institution or just an episode in the life of an individual or a group is analyzed in its relationship to any other in the group." Burgess has called the case study as "the social microscope"."

According to Pauline V. Young case study is "a comprehensive study of a social unit be that unit a person, a group, a social institution, a district or a community." In short, we can state that case study method is a form of qualitative analysis. In case study, where careful and complete observation of an individual or an event or an institution is completed. There are various important characteristics of case study method which are summarized in Fig-3 and also pointed below:

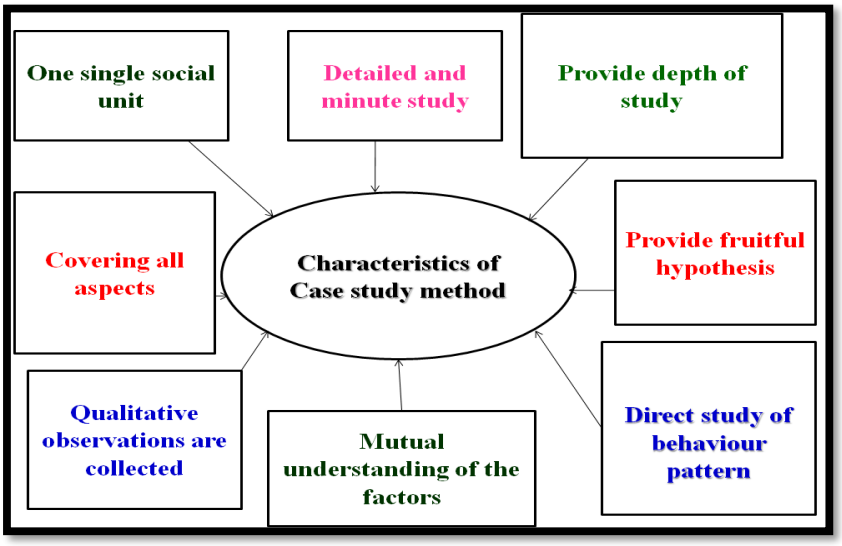


Fig-3: Showing Characteristics of Case study method

1. Under case study, the researcher may take single social unit or more of such units for his study purpose; he can even take a condition to study the same broadly.
2. In the case study method, the selected unit is studied in minute details. Usually, the study expands over a long period to determine the natural history of the unit therefore as to obtain enough information for drawing correct supposition.
3. In the case study method, researchers make complete study of the social unit covering all the aspects. With the help of this method, researchers try to know the complex of factors that are effective within a social unit as an integrated totality.
4. Under case study method the approach happens to be qualitative and not quantitative. Simple quantitative information is not collected in case study method. Every possible effort is made to collect data concerning all aspects. As such, case study deepens perception and gives researcher a clear insight. For example, under this method researcher not only study how many crimes a man has done but will peep into the reasons that forced man to commit crimes when researchers are making a case study of "a man as a criminal".
5. In case study method an effort is made to understand the mutual inter-relationship of fundamental factors.
6. In "case study method" the behavior pattern of the concerning unit is also studied directly.
7. Case study method consequences in productive hypotheses along with the data which may be helpful in testing them, and thus it allows the generalized understanding. Without it, generalized sociology may get handicapped.

Assumptions: The case study method is based on numerous assumptions. The important assumptions are given below:

1. The assumption of uniformity in the basic human nature in spite of the fact that human behavior may vary according to circumstances.
2. The assumption of studying the natural history of the unit concerned.
3. The assumption of comprehensive study of the unit concerned.

Advantages of case study method: There are various advantages of the case study method some of the important advantages of case study methods are given below:

1. Being an exhaustive study of a social unit, the case study method enables us to understand fully the behavior pattern of the concerned unit. In the words of Charles Horton Cooley, “case study deepens our perception and gives us a clearer insight into life”.
2. Through case study a researcher can obtain a real and enlightened record of personal experiences which would reveal man’s inner strivings, tensions and motivations that drive him to action along with the forces that direct him to adopt a certain pattern of behavior.
3. This method enables the researcher to trace out the natural history of the social unit and its relationship with the social factors and the forces involved in its surrounding environment.
4. It helps in formulating relevant hypotheses along with the data which may be helpful in testing them. Case studies, thus, enable the generalized knowledge to get richer and richer.
5. The method facilitates intensive study of social units which is generally not possible if we use either the observation method or the method of collecting information through schedules. This is the reason why case study method is being frequently used, particularly in social researches.
6. Information collected under the case study method helps a lot to the researcher in the task of constructing the appropriate questionnaire or schedule for the said task requires thorough knowledge of the concerning universe.
7. The researcher can use one or more of the several research methods under the case study method depending upon the prevalent circumstances. In other words, the use of different methods such as depth interviews, questionnaires, documents, study reports of individuals, letters, and the like is possible under case study method.
8. Case study method has proved beneficial in determining the nature of units to be studied along with the nature of the universe. This is the reason why at times the case study method is alternatively known as “mode of organizing data”.
9. This method is a means to well understand the past of a social unit because of its emphasis of historical analysis. Besides, it is also a technique to suggest measures

for improvement in the context of the present environment of the concerned social units.

10. Case studies constitute the perfect type of sociological material as they represent a real record of personal experiences which very often escape the attention of most of the skilled researchers using other techniques.
11. Case study method enhances the experience of the researcher and this in turn increases his analyzing ability and skill.
12. This method makes possible the study of social modifications. On account of the minute study of the different facets of a social unit, the researcher can well understand the social change then and now. This also facilitates the drawing of inferences and helps in maintaining the continuity of the research process. In fact, it may be considered the gateway to and at the same time the final destination of abstract knowledge.
13. Case study techniques are indispensable for therapeutic and administrative purposes. They are also of immense value in taking decisions regarding several management problems. Case data are quite useful for diagnosis, therapy and other practical case problems.

Disadvantages of case study methods: There are few disadvantages of case study method which are given below:

1. Case situations are rarely comparable and as such the information gathered in case studies is often not comparable. Since the subject under case study tells history in his own words, logical concepts and units of scientific classification have to be read into it or out of it by the investigator.
2. The danger of false generalization is always there in view of the fact that no set rules are followed in collection of the information and only few units are studied.
3. It consumes more time and requires lot of expenditure. More time is needed under case study method since one study the natural history cycles of social units and that too minutely.
4. The case data are often vitiated because the subject, according to Read Bain, may write what he thinks the investigator wants; and the greater the rapport, the more subjective the whole process is.

5. Case study method is based on several assumptions which may not be very realistic at times, and as such the usefulness of case data is always subject to doubt.
6. Case study method can be used only in a limited sphere. It is not possible to use it in case of a big society.
7. Response of the researcher is an important limitation of the case study method. He often thinks that he has full knowledge of the unit and can himself answer about it. In case the same is not true, then consequences follow. In fact, this is more the fault of the researcher rather than that of the case method.

6.6. Summary

In this unit we have discussed various aspects of data collection. So far you have learnt that:

- Data collection is a process of collecting information from all the relevant sources to find answers to the research problem. It is one of the most important steps of research. Data collection methods can be divided into two categories: Primary methods of data collection and secondary methods of data collection.
- The task of data collection begins after a research problem has been defined and research design /plan chalked out.
- There are several methods of collecting primary data, particularly in surveys and descriptive researches. Some important methods of primary data collection are Observation method, interview method, through questionnaire, through schedules etc.
- Warranty card, distributor or store audits, pantry Audits, Consumer panel, use of mechanical device, projective techniques, depth interviews and content analysis are also other types of methods of primary data collection.
- Secondary data is a type of data that has already been published in newspapers, magazines, journals, research reports, reputed websites etc. Whenever, researchers utilize this type of data then he/she seeks in to various sources from where he/she may collect the data. Secondary data usually publishes data and available in various publications of central, state and local governments, foreign

researches, national and international research papers, reports prepared by research scholars, universities, ministries etc.

- There are various important factors to consider when selecting data collection methods. Following factors should be kept in mind when researcher selects method for data collection.
- The first factor i.e., nature, Scope and object of enquiry constitutes the most important factor affecting the choice of a particular method. The method selected should be such that it suits the type of enquiry that is to be conducted by the researcher.
- The second factor is availability of funds for the research project. It determines to a large extent the method to be used for the collection of data. When funds at the disposal of the researcher are very limited, he will have to select a comparatively cheaper method which may not be as resourceful and effective as some other costly method. Finance, in fact, is a big constraint in practice and the researcher has to act within this limitation.
- The third factor is time. It has also to be taken into account in deciding a particular method of data collection. Some methods take comparatively more time, whereas with others the data can be collected in a comparatively shorter time.
- The case study method is an accepted form of qualitative investigation and entails a careful and whole observation of a social unit, be that unit a person, a family, an institution, a cultural group etc.
- There are various advantages of the case study method some of the important advantages of case study methods are as: case study method enables us to understand fully the behavior pattern of the concerned unit, researcher can obtain a real and enlightened record of personal experiences, method enables the researcher to trace out the natural history of the social unit, Case studies, enable the generalized knowledge to get richer and richer.
- There are few disadvantages of case study method which are as: research rarely comparable and as such the information gathered in case studies is often not comparable. The danger of false generalization is always there in view of the fact

that no set rules are followed in collection of the information and only few units are studied.

TERMINAL QUESTIONS

1. (a) Fill the blank spaces with appropriate words.

Personal interview method requires a person known as the interviewer asking …….. generally in a …….to face contact to the other person or persons. This sort of ……. may be in the form of direct personal investigation or it may be indirect oral……..... In case of direct personal investigation, the interviewer has to collect information personally from the sources concerned. He has to be on the spot and has to meet people from whom data have to be……. This method is particularly suitable for …….. investigations. But in certain cases it may not be possible or worthwhile to contact …….. the person concerned or on account of the extensive scope of enquiry, the direct personal investigation technique may not be used. The method of collecting information through personal interview is generally carried out in a structured way. This type of interview is called as ……interview. Such interviews involve the use of set of …….. questions and of highly standardized techniques of……. As against it …….... interviews are characterized by flexibility of approach to questioning. …… interviews do not follow a system of ……questions and standardized techniques of recording information. ……. interview is other type of interview in which focus attention on the given experience of the respondents and its effects.

2. (a) What do you understand by data collection?

(b) Write about primary and secondary data

3. (a) Write about Primary data collection.

(b) Give a note on secondary data collection

4. (a) Write about advantages and disadvantages of interview method of data collection.

5. (a) What do you understand by Warranty Card method of data collection?

6. (a) Fill the blank spaces with appropriate words

The ……. method is the most commonly used method specially in studies relating around us, but this sort of observation is not scientific observation. Observation becomes a ……. tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and …….and is

subjected to checks and controls on validity and…….. Under the observation methods, the information is sought by way of investigator's own direct …….. without asking from the respondents. For example, in study related to consumer behaviour, the investigator instead of asking the brand of wrist watch used by respondent, may himself look at the watch. The main advantages of this method are that subjective bias is ……. if observation is done accurately. Secondly, the information obtained under this method relates to what is currently happening it is not complicated by either the past behaviour or future ……or attitudes. Thirdly, this method is ……of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happen to be the case in the interview or the questionnaire method. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feeling for one reason or the other.

(b) The data collected from the research papers, newspapers, reports etc. is known as (Primary data/Secondary data)

(c) When researchers conducted interview, it means he/she is collecting the …….. data (Primary/secondary)

(d) Which is characteristic feature of case study method? (Depth study/Qualitative observation/covering almost all aspects/all of the above)

7. (a) what do you understand by distributor or store audits of data collection?

(b) What do you understand by Pantry audits method of data collection?

(c) What do you understand by consumer panel method of data collection?

(d) Write about the selection of appropriate method of data collection.

8. (a) Write an essay on case study method.

ANSWERS

1. (a) questions, face, interview, investigation, collected, intensive, directly, structured, predetermined, recording, unstructured, Unstructured, pre-determined, Focused

2. (a) see the section 6.2

(b) See the section 6.2

3. (a) See the section 6.3

(b) See the section 6.4

4. (a) See the section 6.3 under heading interview method

5. (a) See the section 6.3 under heading warranty card

6. (a) observation, scientific, recorded, reliability, observation, eliminated, intentions, independent
(b) Secondary Data

(c) Primary

(d) All of the above

7. (a) See the section 6.3 under heading distributor or store audit

(b) See the section 6.3 under heading pantry audit

(c) See the section 6.3 under heading consumer panel

(d) See the section 6.5

8. (a) See the section 6.6

Unit 7: Data preparation: Process and problems in preparation process

Unit Structure

7.0. Learning Objectives

After studying this unit, you are able to answer the following questions:

- What is data preparation?
- About importance of data preparation.
- What is the process of data preparation?
- What are the problems in data preparation?

7.1. Introduction

The data after collection has to be prepared for analysis. The collected research data is raw and it must be converted to the form that is suitable for the required analysis. Data preparation is must to get reliable observations. The research data may be collected by mainly two ways viz. primary data and secondary data. You have learnt about these data in unit-6. The primary data is collected by the researchers at very first time and it is original. Primary data is generally collected when researchers do empirical research. When primary data is collected then it is arranged in table according to their sequence. Data preparation is one of the most important aspects of the research and must be carefully followed by the researchers. To express the research explicitly, data must be prepared through various

steps. However, data preparation is time consuming process. In data preparation researchers have to follow various steps such as questionnaire checking, editing, coding, classification, tabulation, graph preparation, data cleaning and data adjusting. After the completion of these steps' researchers can publish their research papers or thesis in reputed journals. In this unit you will learn about the process and problems in preparation process.

7.2. Importance of data preparation

Data preparation is an integral step of research methodology. It is decisive processes in research. Data preparation is the method of consolidating the data for use in analysis. It improves the data, transforms it and improves the accuracy of the outcome. There are some challenges to researchers in data preparation and these challenges are as multiple data formats, data inconsistency, limited/large access to data and lack of data integration infrastructure. It is mostly done through analytical or traditional extract. Data preparation has its own importance which has given below:

It crosses out the manual work of searching, cleansing and transforming the data for analysis. Moreover, the self-service data preparation tools reduce the dependence on IT support and decrease the time to prepare data. Data cleansing and manipulation tools improve the integrity and quality of data and could be easily connected to multiple sources. The tools save a lot of time, correct and improve the quality of data, and help analysts uncover business insights that are useful for business decision-making. Growing complexity in data necessitates embracing analytical techniques during the ETL process. With these techniques, analysts and data scientists would be able to identify outliers and missing data; know the distribution and variance in data; and use machine learning in reducing, and classifying the data for better analysis. Metadata describes the data and helps in labeling the data variables. A common metadata drives collaboration across the data management and analytical domains. It provides lineage information on the data preparation process and results in the accuracy of business models. A lot of organizations base their analysis on historical data. Although this significantly helps in predicting the future, but they miss out a lot of what is currently happening in the real world. Embedded analytics offer real-time information for timely decision-making, and also help in accessing a variety of data sources. Once the data is ready, analysts and data scientists would be able to build different models

and derive useful insights from the data. All these practices will prove helpful to organizations in realizing the true value of data in the form of accurate insights.

7.3. Process in data preparation

After data collection, the researcher must prepare the data to be analyzed. Organizing the data correctly can save a lot of time and avoid error. Most of the researcher scholars select to use a database e.g., Microsoft Excel, SPSS) that they can format to fit their requirements and organize their data efficiently. The plan of data analysis is decided in advance before colleting the data. The process of data preparation is guided by the plan of data analysis. There are various important steps of data preparation are as: Questionnaire checking, editing, coding, classification, tabulation, graphical representation, data cleaning and data adjusting.

7.3.1. Questionnaire Checking

When is data is collected through the help of questionnaires, first step of data preparation process is the check/monitored the questionnaire? In this step we check questionnaires are acceptable or not. This constitutes the analysis of all the questionnaires for their completeness and interviewing quality. Generally, this step is taken after the time of the data collection. If this checking is not completed at the time of collection, it should be done after sometimes. A questionnaire should not be acceptable if it is incomplete or it is filled by the person who has insufficient knowledge regarding questionnaire. It should also unbiased questionnaire.

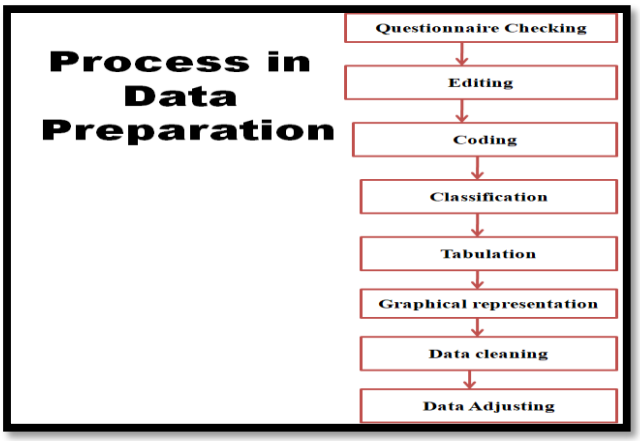


Fig-1: Showing steps in Data preparation

7.3.2. Editing

Editing is the process of data preparation in which examining the collected raw data to detect mistakes and omissions. In this step we can correct these errors or omissions. Editing involves a careful scrutiny of the questionnaires. Editing is completed a assured that the data are correct, consistent with other information gathered informally entered as completed as possible have been well arranged to facilitate coding and tabulation. With regards to points or stages at which editing should be done, one can talk a field editing and central editing. Field editing consists in the review of the reporting forms by the investigators for completing what the latter has written in abbreviated form at the time of recording to respondent's responses. Field editing is necessary in view of the facts that individual's writing styles often can be difficult for other types to decode. This sort of editing should be done as soon as possible after interview, preferably on the very day or on the next day. While doing field editing, the investigator must restrain himself and must not correct errors of omission by simply guessing what the informant would have said if the question had been asked.

Central editing should take place when all forms and schedules have been completed and returned to the office, this type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a huge enquiry. Editors can correct the obvious errors such as an entry in the wrong place, entry recorded in inappropriate time etc. Editors must keep in mind several points while performing editing. These points may as:

1. They should be familiar with instructions given to the interviewers and coders as well as with the editor's instructions supplied to them for the objective.
2. While crossing out an original entry for one reason or another they should just draw a single line on it. Therefore, same may remain readable.
3. They should make entries initial all answers which they change.
4. Editor's initials and the data of editing should be placed on each completed form.

Coding:

It refers to the process of assigning numerals or other symbols to answer so that responses can be put into a limited number of categories. Such categories should be appropriate to the research problem under analyzed. They must also contain the characters of exhaustiveness and also that of mutual exclusively which means that a specific answer can be placed one and only one cell in a given category set. Another rule to be found is that of unidimensional by which it meant that every class is defined in term of only one concept.

Coding is important for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis. Coding decisions should generally be taken at the designing stage of the questionnaire. This makes it possible to precede the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch from the original questionnaire.

7.3.3. Classification

Most research studies result in a large volume of raw data which must be reduced into homogenous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characters. Data having common characters are placed in one class and in this method the whole data get classified into two number of groups. On the basis of nature of phenomena involved, classification may be of following types.

Classification according to attributes: As said that above, data are classified on the basis common characters which can either be descriptive such as literacy, sex, honesty etc or numerical such as weight, height, income etc. Descriptive characters may include to qualitative phenomenon which cannot analyzed numerically only their presence and absence in an individual items can be observed. Data gathered this way on the basis of certain characters are known as statistics of characters and their classification is called to be Classification according to attributes. This type of classification is may be simple or manifold. In simple classification we consider only one character and divide the universe into two classes viz. One class consisting of items possessing the given attribute and the other class consisting of item which do not possess the given attribute. But In case of manifold

classification, we consider two or more attribute simultaneously and divide that data into a number of classes (Total number of classes to final order is given by 2^n) where n=number of attributes considered. Whenever, data are classified according to attribute the researcher must see that the attributes are defined in such a manner that there is least possibilities of any doubt concerning the said attributes.

Classification according class intervals: Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to income, production, age, weight etc. come under this type of classification. Such data is called as statistics of variables and are classified on the basis of class intervals. For example, person whose weights say are within 40-50 kg can form one group, those whose weight are within 51 to 70 kg can form another group and so on. In this way the entire data may be divided into a number of groups or classes which is generally called as class intervals. Each class intervals have lower limit and upper limit which is collectively called as class limits. The difference between two class limits is known as class magnitude. We may have classes with equal class magnitude or with unequal class magnitudes. The number of items which fall in a given class is known as frequency of the given class.

Exclusive type class intervals: They are as 10-20, 20-30, 30-40 and 40-50. These intervals should be read as 10 and under 20, 20 and under 30, 30 and under 40 and 40 and under 50.

Inclusive type of class intervals: They are as 11-20, 21-30, 31-40 and 41-50. In this type of class interval, the upper limit of class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11-20 class interval. In above stated upper limit of class interval 11-20 is 20 but the real limit 20.99 and as such 11-20 class interval really means 11 and under 21.

Researchers can categorize biodiversity according to their phylum, classes, orders, families etc. For example, research can categorize data on biodiversity as follows:

Animals			Plants		
Phylum/Class	Class	Number of Species	Division/Class	Family/Order	Number of Species
			Algae	Chlorophyceae	12
Protozoa	Rhizopoda	6		Bacillariophyceae	14
	Flagellata	4		Cyanophyceae	5
	Ciliata	5		Rhodophyceae	3
Arthropoda	Insecta	23		Xanthophyceae	7
	Crustacea	9	Bryophytes	Riccia	2
	Millipoda	2		Marchantia	1
Annelida	Oligochaeta	2	Pteridophytes		5
	Polychaeta	2	Gymnosperm		2
	Hirudinia	7	Angiosperm		34

7.3.4. Tabulation

When a mass of data has been assembled, it become necessary to arrange the data in some type of concise and logical order. This process is known as tabulation. Tabulation is the process in which researcher summarizing raw data and displaying the same in compact form or in the form of statistical table for further analysis. In broader sense tabulation is an orderly arrangement of data in columns and rows. Tabulation is important because of the following reasons.

1. It conserves space and reduces explanatory and descriptive statement to a minimum.
2. It facilitates the process of comparison.
3. It facilitates the summation of items and the detection of errors and omissions.
4. It provides a basis for various statistical computations.

Tabulation can be done by hand or by the help of computers. The choice depends on the size and kind of study, time pressure and availability of computers. In relatively large inquiries, we may use computer tabulation if other factors are favorable and necessary facilities are available. Tabulation may be classified in to two categories viz. simple and complex. Simple tabulation gives information about one or more groups of independent questions, whereas the complex type of tabulation shows the division of data in two or more categories and such as the deigned to give information concerning one or more set of

interrelated questions. Simple tabulation generally results in one way table which supply answer to questions about one characterizes of data only. As against this complex tabulation usually result in two-way table, three-way table or still higher order tables also called as manifold tables, which supply information about several interrelated characteristics of data. Two-way table, three-way table or manifolds tables are examples of cross tabulations.

Tabulation is very important aspect of data preparation and the tables must have following characters.

1. Each table should have clear, concise and satisfactory title so as to make the table clear without reference to the text and title should always be placed just above the table.
2. Each table should be given distinct number of facilitate essay reference.
3. The column headings and the row headings of the table should be clear, concise and brief.
4. The units of measurements under each heading or sub –heading must always be indicated.
5. Explanatory foot notes if any concerning the table should be place directly beneath the table, along with the reference symbol used in the table.
6. Sources from where the data in the table have been collected must be showed just below the table.
7. Generally, the columns are separated from one another by lines which make the table clearer.
8. There should be thick likes to segregate the data under one class from the data under another class and the lines separating the sub-division of the classes should be comparatively thin lines.
9. The column may be number to facilitate references.
10. Those columns whose data are to be compared should be kept side by side. Similarly, percentages and averages must also be kept close to the data.
11. Table should be aligned perfectly by which any one can read the number easily.

12. Miscellaneous items should be kept in last of the table.

Example: Table showing Air Quality Index of different cities of India (Mean Value)

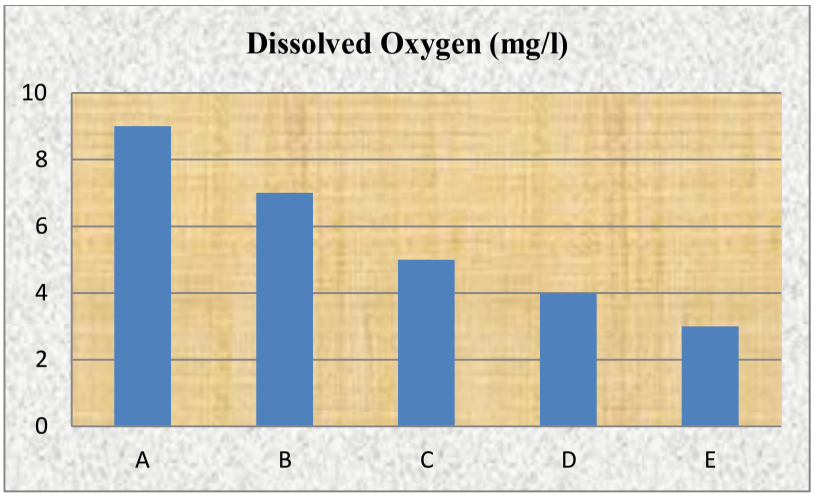
S.N.	Cities	Year		
		2018	2019	2020
1.	Delhi	125	120	130
2.	Chennai	120	115	125
3.	Haldwani	100	100	105
4.	Haridwar	105	95	100
5.	Jaipur	130	135	140
6.	Jodhpur	137	140	135
7.	Kanpur	97	100	110
8.	Kolkata	95	90	100
9.	Mumbai	100	110	105
10.	Srinagar,	130	120	135

(Source: Report of CPCB, 2021)

7.3.5. Graphical representation:

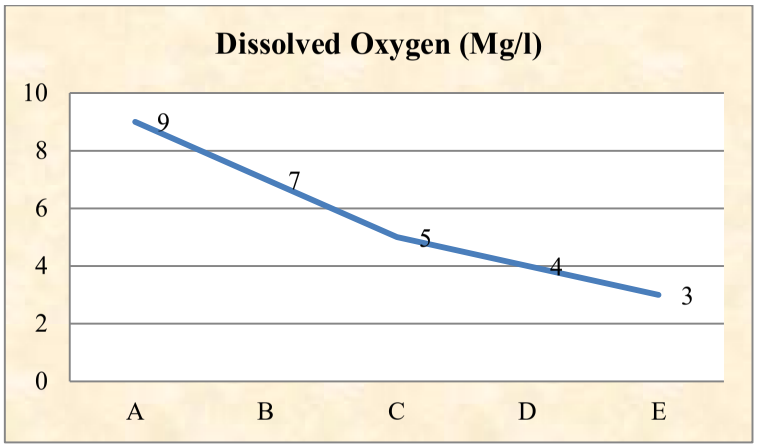
Graphs are very important in understanding the data in easiest way. Graphs are easy to understand as compared to table. All statistical packages, MS excel and open office.org offer wide range of graphs. In case of qualitative data most common graphs are as follows.

1. Bar Chart: It consists of a series of rectangles (or bars). The height of each bar is determined by the frequency of that category. Suppose that the dissolved oxygen level in Ganga River in year 2011-2012 in five different regions, denoted as ABCD and E are 9, 7, 5, 4 and 3, respectively measures in mg/l. The bar chart of this data is as below:



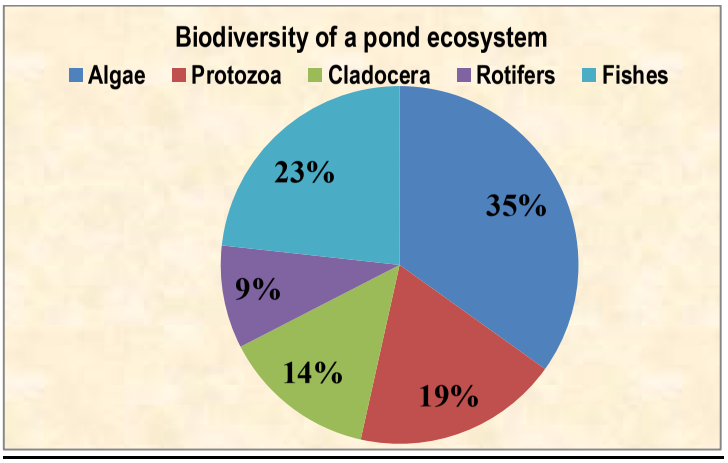
Bar Chart showing dissolved oxygen in five geographical regions

2. Line graphs: Line graphs or charts are useful when we wish to compare to data as we can overlap to line charts. For example, the dissolved oxygen in Ganga River in year 2011-2012 in five different regions, denoted as ABCD and E are 9, 7, 5, 4 and 3, respectively. The line chart of the data is given below:



Line Chart showing dissolved oxygen in five geographical regions

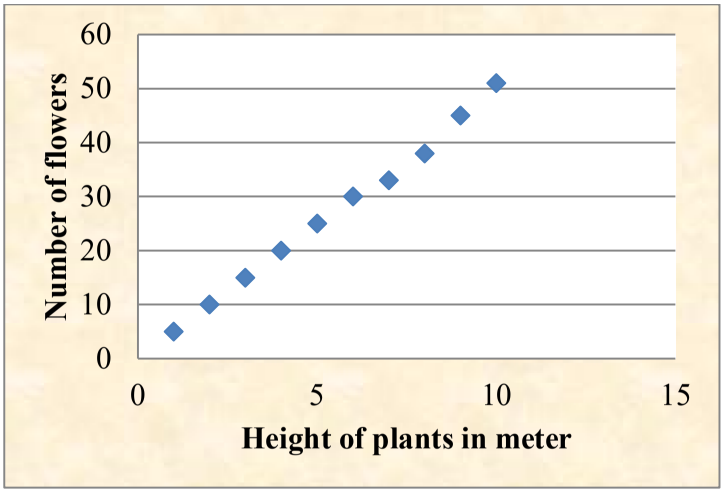
3. **Pie chart:** A pie chart is used to focus relative proportion or shares of each category. It's a circular chart divided into sectors, illustrating relative frequencies. The relative frequencies in each category or sector are proportional to the arch length of that sector or the area of that sector or the central angle of the sector. Suppose in pond ecosystem you have find 15 species of algae, 8 species of protozoa, 6 species Cladocera, 4 species of rotifers and 10 species of fishes. The pie chart of this data is given below:



Pie chart showing diversity of pond ecosystem

4. Scattered Chart: This type of graph or chart also used in presentation of data. Suppose we have data in which we have counted number of flowers in different plants. We have also recorded height of the plants then we can represent this data in scattered chart as follows:

Height of plant in meter	1	2	3	4	5	6	7	8	9	10
No. of flowers in number	5	10	15	20	25	30	33	38	45	51



Scattered chart showing number of flowers and height of plants

7.3.6. Data cleaning:

This includes checking the data for consistency and treatment for missing value preliminary consistency check are made in editing._Here we check the consistency in an extensive manner. Consistency check looks for the data which are not consistent or outlines. Such data may either be discarded or replaced by the mean value. However, the researcher should be careful while doing this. Extreme values of outlines are not always mistaken. Missing value is the value which is unknown or not answered by respondents. In place of such missing value some neutral value may be used. This neutral value may be the mean of available values. The other option could be using the pattern of responses to other questions to calculate a suitable substitute to the missing values.

7.3.7. Data Adjusting:

Data adjusting is not always necessary but it may improve the quality of analysis sometimes. this consists of following methods.

1. Weight-assigning: Each respondent or case is assigned a weight to reflects its importance related other respondents or cases. In this method, the collected sample can be made a stronger representative of a target population on particular character. For instance, the case of literate people could be assigned higher weights and of illiterate people could be assigned lower weight in some survey. The value 1.0 means unweightage case.
2. Variable re-specification: This includes creating new variables or modifying variables. For instance, if the usefulness of a certain product is measured on 10-point scale, it may be recorded on 4-point scale –very useful, useful, neutral, not useful. Ratio of two variables may also be taken to create a new variable. Method of dummy variables for re-specifying categorical variable is also very popular. Dummy variable which generally takes numerical values based on the corresponding category in the original variable. For instance, a group of people is divided in to smoker and non-smokers. We can define dummy variable taking a value 1 for smokers and 0 for non-smokers.
3. Scale transformation: It is done to ensure the comparability with other scale or to make the data suitable for analysis. Different types of characters are measured on different scales. For instance, attitude variables are measured on continuous scale, life style variables are generally measured on 5-point Likert scale. Therefore, the variables which are measured on different scales cannot be compared common transformation is subtracting all the value of characters by corresponding mean and dividing by corresponding standard deviation.

7.4 Problems in data preparation

Researchers may face various problems in data preparation. some of them are given below:

1. Spending too much time preparing data: Data preparation required much time as compared to other steps of research. Sometimes researchers can not publish their research papers within time due to lengthy process of data preparation.

2. Data preparation heavily depend on IT departments: Many data and analytics teams rely heavily on their IT department to source the data they need to run their projects.

3. Preparing data manually: Manual data preparation tools like Microsoft Excel can hinder collaboration and efficiency but remain popular among researchers. This reliance on manually driven data preparation tools will continue to delay data initiatives and deter new insights.

4. Not spotting data quality issues: Remediating issues of data quality can significantly impact on the end analysis. For example, marketing lead data is far more valuable when it has been enriched with external data to complete missing values. Or consider the difference between outdated versus up-to-date data when predicting sales and calculating margins. In both instances, improperly prepared data can have a huge impact.

5. The problem concerning "Don't know" responses: When the DK response group is small, it is of little significance. But when it is relatively big, it becomes a matter of major concern in which case the question arises: Is the question which elicited DK response useless? The answer depends on two points viz., the respondent actually may not know the answer or the researcher may fail in obtaining the appropriate information. In the first case the concerned question is said to be alright and DK response is taken as legitimate DK response. But in the second case, DK response is more likely to be a failure of the questioning process.

Use or percentages: Percentages are often used in data presentation for they simplify numbers, reducing all of them to a 0 to 100 range. Through the use of percentages, the data are reduced in the standard form with base equal to 100 which fact facilitates relative comparisons. While using percentages, the following rules should be kept in mind:

1. Two or more percentages must not be averaged unless each is weighted by the group size from which it has been derived.
2. Use of too large percentages should be avoided, since a large percentage is difficult to understand and tends to confuse, defeating the very purpose for which percentages are used.

3. Percentages hide the base from which they have been computed. If this is not kept in view, the real differences may not be correctly read.
4. Percentage decreases can never exceed 100 per cent and as such for calculating the percentage of decrease, the higher figure should invariably be taken as the base.
5. Percentages should generally be worked out in the direction of the causal-factor in case of two-dimension tables and for this purpose we must select the more significant factor out of the two given factors as the causal factor.

7.5 Summary

In this unit we have discussed various aspects of data preparation. So far you have learnt that:

- After data collection, the researcher must prepare the data to be analyzed. Organizing the data correctly can save a lot of time and avoid error. Most of the researcher scholars select to use a database e.g., Microsoft Excel, SPSS) that they can format to fit their requirements and organize their data efficiently.
- The plan of data analysis is decided in advance before colleting the data. The process of data preparation is guided by the plan of data analysis. There are various important steps of data preparation are as: Questionnaire checking, editing, coding, classification, tabulation, graphical representation, data cleaning and data adjusting.
- When is data is collected through the help of questionnaires, first step of data preparation process is the check/monitored the questionnaire. In this step we check questionnaires are acceptable or not. This constitutes the analysis of all the questionnaires for their completeness and interviewing quality.
- Editing is the process of data preparation in which examining the collected raw data to detect mistakes and omissions. In this step we can correct these errors or omissions.
- Coding is referring to the process of assigning numerals or other symbols to answer so that responses can be put into a limited number of categories. Such categories should be appropriate to the research problem under analyzed. They must also contain the characters of exhaustiveness and also that of mutual exclusively which

means that a specific answer can be placed one and only one cell in a given category set. Another rule to be found is that of one-dimensionality by which it meant that every class is defined in term of only one concept.

- Most research studies result in a large volume of raw data which must be reduced into homogenous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characters. Data having common characters are placed in one class and in this method the whole data get classified into two number of groups. On the basis of nature of phenomena involved, classification may be of following types.
- Tabulation is the process in which researcher summarizing raw data and displaying the same in compact form or in the form of statistical table for further analysis. In broader sense tabulation is an orderly arrangement of data in columns and rows. Tabulation is important because of the following reasons.
- Graphical representation: Graphs are very important in understanding the data in easiest way. Graphs are easy to understand as compared to table. All statistical packages, MS excel and open office.org offer wide range of graphs.
- Data cleaning includes checking the data for consistency and treatment for missing value preliminary consistency check are made in editing.
- While processing the data, the researcher often comes across some responses that are difficult to handle. One category of such responses may be 'Don't Know Response'. Two or more percentages must not be averaged unless each is weighted by the group size from which it has been derived.

TERMINAL QUESTIONS

1. (a) Fill the blank spaces with appropriate words.

……..is the process in which researcher summarizing ……data and displaying the same in compact form or in the form of statistical ……for further analysis. In broader sense ……..is an orderly arrangement of data in columns and **rows**. Tabulation is important because of the following reasons. **……**can be done by

hand or by the help of……. The choice depends on the size and kind of study, time pressure and availability of computers. In relatively large inquiries, we may use computer tabulation if other factors are favourable and necessary facilities are available. ……tabulation gives information about one or more groups of independent questions, whereas the complex type of tabulation shows the division of data in two or more categories and such as the deigned to give information concerning one or more set of ……questions. Simple tabulation generally result in one way table which supply answer to questions about one characterizes of data only. As against this ……..tabulation usually result in two way table, three way table or still higher order tables also called as …….tables.

2. (a) What do you understand by data preparation?

(b) Write about process in data preparation.

3. (a) What do you understand by editing in data preparation? Explain

(b) Write about step "Classification" in data preparation

4. (a) Write the characteristics of "Tabulation".

5. (a) What do you understand by Graphical representation of data?

6. (a) Fill the blank spaces with appropriate words:

……..is the process of data preparation in which examining the collected ……data to detect ……..and omissions. In this ……we can ……..these errors or omissions. ……..involves a careful scrutiny of the questionnaires. Editing is completed a assured that the data are correct, consistent with other information gathered informally entered as completed as possible have been well arranged to facilitate coding and tabulation. With regards to points or stages at which editing should be done, one can talk a field editing and ……..editing. Field **……..**consists in the review of the reporting forms by the investigators for completing what the latter has written in abbreviated form at the time of recording to respondent's responses. Field editing is necessary in view of the facts that individual's writing styles often can be difficult for other types to decode. This sort of editing should be done as soon as possible after interview, preferably on

the very day or on the next day. While doing……. ……., the investigator must restrain himself and must not correct errors of omission by simply guessing what the informant would have said if the question had been asked.

7. (a) Describe the problems in data preparation.

ANSWERS

1. **(**a) Tabulation, raw, table, tabulation, Tabulation, computers, Simple, interrelated, complex, manifold

2. (a) see the section 7.1
   (b) See the section 7.2
3. (a) See the section 7.2.2
   (b) See the section 7.2.4
4. (a) See the section 7.2.5
5. (a) See the section 7.2.6
6. (a) Editing, raw, mistakes, step, correct, Editing, central, editing, field editing
7. (a) See the section 7.3

Unit 8: Descriptive analysis: Measures of central tendency (Mean, median, mode, other averages) Measures of dispersion (range, mean deviation and standard deviation; Measures of skewness and relationship, Association in case of attributes and other measures (index numbers and time series)

Unit Structure

8.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What are descriptive statistics?
- What is measure of central tendency?
- What is mean, mode and median?
- What is measure of dispersion?
- What is range?
- What is mean deviation and how it calculated?
- What is standard deviation and how it calculated?

8.1. Introduction

Statistics is a science that uses different tools and techniques to organize the data in the descriptive form (in the form of table, graphs and pictorial presentation) and extract information from the data that helps in making decisions. Statistics is considered as a study of collecting, analyzing, interpreting, presenting and organizing the data.

Statistics deals with collection of data related with an objective, its analysis and finally its interpretation that is understandable by all the concerning persons involved with the objective in a direct or indirect manner.

Classification of Statistics

Statistical methods have different branches and each branch has its great importance in the literature as well as for practical point of view. The following model explains the relation between these methods.

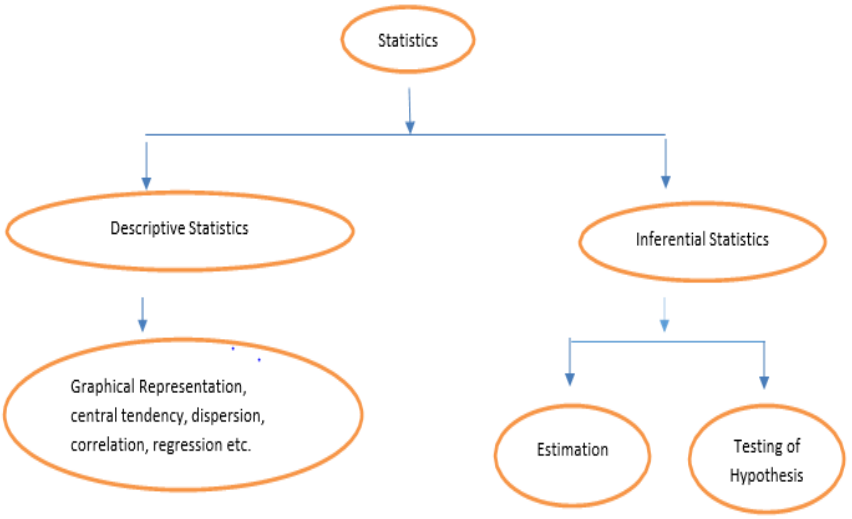


Figure 1: Classification of Statistics

From the above model, one can see that for statistical methods are segmented into two methods. One is descriptive statistics and second is inferential statistics.

(a) Descriptive Statistics

In Descriptive Statistics, graphical representation methods like histogram, bar chart, pie chart, ogive curves, box plot, line chart etc. are used to present data in graphical

form. Descriptive methods also deal with the summary of the data. This includes these following measures:

- Central Tendency Measures: mean, median, mode.
- Measure of Dispersion: Variance, standard deviation, range.

There are other measures like skewness, kurtosis, quartiles, quantile, percentile etc. These measures are used to study other characteristics of the data. In descriptive statistics, one can study the scatterplot, correlation between variables, regression analysis. Basically, descriptive statistics deals with quantitatively describing the features of a data available from the sample/population.

(b) Inferential Statistics

Descriptive statistics is used just to represent the information available in the data. But sometimes it is difficult to collect all the observation in the study for analysis. In this situation, sampling technique is used to collect element of the population in a non-random/random manner. Then inferential statistics is used to extract information from the sample as well to test the significance of an objective based on observations using testing of hypotheses. It is further segmented into two parts: These are

- Estimation: Estimation techniques is to extract information about the population from the sample.
- Testing of Hypotheses: This method is used to test the significance of a statement about a problem.

By using statistical methods, tools and techniques, one can find a numerical value that can help a person to find out solution.

8.2. Measure of Central Tendency

In real life, we collect data from population that has same characteristics for a particular objective. The data that are collected contain elements may have different information. In this scenario, we use a measure that provides an idea about the data. This measure is called 'central tendency measure'. Central tendency is a measure that provides a single value that represents a group of values. However, it should satisfy some certain conditions.

There are some properties that are expected from a central tendency measure:

- It should be defined in a rigid manner such that its meaning should be unambiguous.
- It should be calculated on the entire observations in the data.
- It should be calculated in a little time frame and in an easy manner.
- It should possess mathematical properties so that we can further use it.
- It should not be influenced by the extreme values of the data.
- It must be sensitive to the small changes in the data values.

In Statistics, three measures are considered generally that is mean, median and mode.

8.2.1 Mean

Mean is among the most widely used measure of central tendency for a single representation of observations. The formula for the evaluation of the mean is given as:

$$\text{Mean} = \frac{\text{Sum of all observations}}{\text{Number of observations}}$$

Let x_1, x_2, …, x_n be n observation is a data set. The mean of the data is denoted $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For example, let 23, 24, 25, 35, 32 be 5 class size of different sections of 9th class in a private school. We want to know about the average class size in 9th class. To calculate this, we first calculate the sum of all class sizes and divide it by no of sections. The mean is $27.8 \approx 28$. It means that there are approximately 28 students on average basis in each class.

This method cannot be applied if the no. of the elements in the data set are very large. For the convenience purpose frequency tabulation method is used.

Let x_1, x_2, …, x_n have frequencies f_1, f_2, …, f_n respectively that means that in the complete data set x_1 appears f_1 times and x_2 appears f_2 times and so on. Hence the formula for mean in this case will be

$$\bar{x} = \frac{(f_1x_1 + f_2x_2 + \ldots\ldots\ldots + f_nx_n)}{f_1 + f_2 + \ldots\ldots + f_n}$$

This method of calculating Mean in term of frequencies is known as Weighted Mean. As we are multiplying the x_i's with their corresponding frequencies (f_i).

Table 1: An example consists of no. of child in 50 families in a locality is presented below will give you a better understanding of the weighted mean.

No. of Child	0	1	2	3 or above
Families	10	15	20	5

Now to calculate the average number of children in a family. One can use weighted mean and the value is 1.4 ≈ 1. Hence on average there is a single child in each family.

Table 2: Also in real life, we are interested in the marks of the students spending hours on daily basis to learn through online program. An artificial data was available of 1000 students and information is given in the following table as:

Time Spent (in hours)	No. of Students
0-4	690
4-8	285
8-12	20
12-16	5

Now if one is interested to find the average no of hours spend on average basis using A.M. then the formula is:

$$\bar{x} = \frac{(f_1x_1 + f_2x_2 + \ldots\ldots\ldots + f_nx_n)}{f_1 + f_2 + \ldots\ldots + f_n}$$

but here $x_1, x_2, \ldots, x_n$ are the midpoint of class interval. So, the mean of the data can be found as

$$= \frac{2 * 690 + 6 * 285 + 10 * 20 + 14 * 5}{690 + 285 + 20 + 5}$$

$$= \frac{3360}{1000} = 3.360$$

Hence from the above example, we can conclude that 3.36 hours or 3 hours and 6 minutes on average are spent by the students on daily basis.

8.2.2 Median

When the data set has outliers, mean becomes flowed as a representative of the data set. In such a case, median is used as a measure of central tendency. median divides the data set into two equal parts. half of the items are less than the median and remaining half of the items larger than the median. In order to obtain the median, we first arrange the data set into ascending or descending order. if number of observations in the data set is n, then:

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{the observation, when n is odd}$$

$$= \frac{1}{2}\left[\left(\frac{n}{2}\right)\text{the observation} + \left(\frac{n}{2}+1\right)\text{th observation}\right], \text{when n is even}$$

For example, median of the data set 2, 5, 6, 11, 56 is 6, while median of the data set 5, 6, 11, 59, or 8.5.

Median is a positional average and is used only in the context of qualitative phenomenon, for example, in estimating intelligence, etc.., which are often encountered in sociological fields. median is not useful where items need to be assigned relative importance and weights .it is not frequently used in sampling statistics.

Table 3: Find out the median of following dataset-

N	1	2	3	4	5	6	7	8	9
X	25	36	21	56	43	53	63	49	71

Ascending ordered series is: 21, 25, 36, 43, **49**, 53, 56, 63, 71.

N is odd here, so the median is 49.

Table 4: Find out the median of following dataset-

N		1	2	3	4	5	6	7	8	9	10
X		25	36	21	56	43	53	63	49	71	67

Here, N is an even value for finding out the median value. Again, arrange the observations in ascending order: 21, 25, 36, 43, **49**, **53**, 56, 63, 67, 71.

Here, the median value lies between the 5th and 6th value in the ordered data set. After taking mean of these two values the median value is calculated as 51.

8.2.3 Mode

The most frequently occurring observation in the data set is mode. Mode is a French word having the meaning fashion. It is particularly useful in the study of popular size. For example, a manufacture of shoes is usually interested in finding out the size most in demand so that he may manufacture a larger quantity of that size like median mode is also a positional average and is not affected by extreme values. Mode is not amenable to algebraic. A data set may not have any mode or there may be more than one mode in a data set.

Relationship between Mean, Median and Mode:

There is a relationship between mean, median and mode depending upon the symmetry of the data. If data are symmetric then,

Mean = median = mode

If data are not symmetric i.e., asymmetric then either

Mean > median > mode

Or

Mode > median > mean

Hence, in literature another important exist between three measures. It is called an empirical relation between these three measures.

Mean – Mode = 3 (Mean – Median)

8.3 Measure of dispersion

The basic purpose of central tendency measures is to find out a single value that represent the whole dataset. Also, concentration of observations about the central part of the data were observed. But this measure will not take into account of the fact that whether two or more different dataset have same mean values but it does not mean

that the observations are same. Hence, one should not rely just on the central tendency measure to take any opinion about the observations but also one should also think about the dispersion or variation among the observation.

Some authors have defined the measures of dispersion as:

"Dispersion is the measure of variation of the items" by A.L. Bowley.

"Dispersion or spread is the degree of the scatter or the variation of the variable about a central value" by B.C. Brooks and W. F. L. Dicks.

8.3.1. Types of Measures of Dispersion

The measures of dispersion are further categorized into two types. These are shown in the following flowchart:

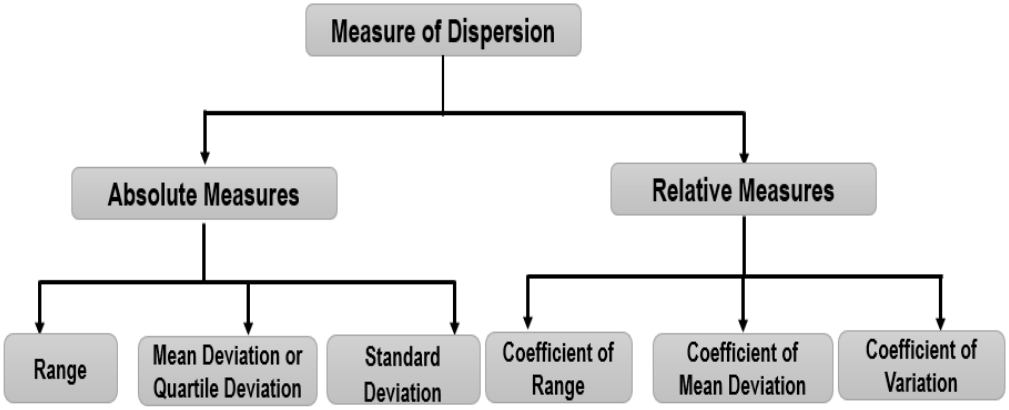


Figure 2: Categorization of Measure of Dispersion

8.3.1.1 Absolute Measure of Dispersion: As dispersion measure is used to detect the deviation of the observations from the central tendency. If the measures of dispersion express the dispersion of the observations in the original units, then the measures are called absolute measures of dispersion.

(a) Range: Range is considered as the simplest absolute measure of dispersion. It is evaluated by just taking the difference between the maximum value and minimum value in the data set.

Range = Maximum value- Minimum value

Ques: Calculate the range of the following dataset:

23, 45, 56, 52, 64, 35, 42, 51, 76, 65.

Ans: Maximum value is 76 and minimum value is 23.

Hence Range = 76-23 = 53.

(b) Quartile Deviation: This measure of dispersion is related with range and it removes the drawbacks of range up to some extent. It is defined as the difference between third quartile and first quartile divided by 2. In other words,

$$\text{Quartile Deviation} = \frac{\text{Q}_3 - \text{Q}_1}{2}$$

where Q_1 is the first quartile;

Q_3 is the third quartile.

Also, $Q_3 - Q_1$ is known as the interquartile range.

Ques. Find out the quartile deviation from the following data set related to the marks obtained by 15 students in statistics.

60, 67, 56, 78, 92, 55, 72, 54, 49, 59, 37, 84, 83, 69, 62

Ans: Arrange the observations in the ascending order 37, 49, 54, 55, 56, 59, 60, 62, 67,69, 72,78, 83, 84, 92

Now compute the first quartile $\frac{(\text{N}+1)^{\text{th}}}{4}$ term i.e., 55

Similarly, third quartile as 12th term i.e., 78. Quartile deviation is (78 – 53)/2 = 25/2 = 12.5.

(c) Mean Deviation: Mean deviation is computed as the absolute value of difference of observations from a central tendency i.e., mean, median and mode. Mostly, mean deviation is calculated by taking deviation of observations from mean and median.

Let x_1, x_2, …, x_n be the n observations to compute mean deviation. First assume a number A and take absolute deviation of each observations from the value A i.e., $|x_1 - A|$, $|x_2 - A|$,…,$|x_n - A|$. As

some of the observations are more than A and some of them are less so modulus is taken to accumulate the total deviation of observation from a point A.

Now the sum of deviated observations from A is $\sum|x_i - A|$ and the mean of these observations will give mean deviation about A.

$$\text{Mean deviation about A} = \frac{1}{n}\sum |x_i - A|$$

If A is chosen as mean value, then the deviation is called mean deviation about mean and if A is chosen as median then the deviation is called mean deviation about median.

Ques. Calculate the mean deviation about mean for the following observations. 30, 40, 50, 55, 60, 65, 70, 80, 90, 100

Ans: First find out the mean of the observations.

$$\text{mean} = \frac{30 + 40 + 50 + 55 + 60 + 65 + 70 + 80 + 90 + 100}{10}$$

$$\text{mean} = \frac{640}{10} = 64$$

Now take absolute deviation of each observation from 64. The values are 34, 24, 14, 9, 4, 1, 6, 16, 26. Take the sum of these observation and divide it by 10 i.e., number of observations.

The answer is 134/10=13.4.

The above formula is used for simple series data.

For ungrouped frequency distribution, the formula will be

$$\text{Mean deviation about A} = \frac{1}{N}\sum f_i|x_i - A|$$

For grouped frequency distribution, the formula will be

$$\text{Mean deviation about A} = \frac{1}{N}\sum f_i|m_i - A|$$

where N is the total frequency;

f_i is the frequency corresponding to the i^{th} observation;

m_i be the midpoint of the i[th] class interval.

Table 5: Calculate the mean deviation about mean for the following dataset.

Observations	Frequency
38	7
43	9
46	10
49	6
51	4
54	8
67	3
78	5

Ans: First compute mean of the data using the formula $\frac{\sum \text{fixi}}{N}$, where, $N=\sum f_i$. The values are shown in the Table 2. After that subtract this mean value from each observation and take modulus. Now multiply this absolute value with the corresponding frequency. Now compute the mean of these absolute observation. Hence 8.22 is the mean deviation about mean.

Table 6- Calculation Table for mean deviation measurement

Observations	Frequency	$f_i x_i$	deviation from mean $x_i - \bar{x}$	absolute deviation about mean $\lvert x_i - \bar{x}\rvert$	$f_i * \lvert x_i - \bar{x}\rvert$
38	7	266	-12.653846	12.65384615	88.57692308
43	9	387	-7.6538462	7.653846154	68.88461538
46	10	460	-4.6538462	4.653846154	46.53846154
49	6	294	-1.6538462	1.653846154	9.923076923
51	4	204	0.34615385	0.346153846	1.384615385
54	8	432	3.34615385	3.346153846	26.76923077
67	3	201	16.3461538	16.34615385	49.03846154
78	5	390	27.3461538	27.34615385	136.7307692
	N = 52	$\sum f_i x_i = 2634$			427.8461538

$$\text{Mean} = \frac{\sum \text{fixi}}{\text{N}} = \frac{2634}{52} = 50.65384615$$

$$\text{Mean deviation} = \frac{1}{\text{N}} \sum \text{f}_\text{i} |\text{x}_\text{i} - \text{A}|$$

$$= \frac{427.8461538}{52} = 8.22$$

(d) Standard Deviation: Karl Pearson in 1823 first introduced this measure and after that it is the most widely used measure of dispersion till date. Standard deviation is the measure that prevails all the features that other measures lack of. So basically, it is considered as an ideal measure of deviation. It is also known as square root of variance; root mean square deviation etc. and denoted by the Greek letter σ (sigma). Standard deviation value is high or low when there is more or less variation among observations respectively.

$Let\ x_1, x_2, \ldots, x_n$ be the n observations and $\bar{x}$ be the average value, then the standard deviation (S.D.) is evaluated as

$$\text{S. D.} = \frac{\sqrt{\sum(\text{x}_\text{i} - \bar{\text{x}})^2}}{n}$$

For frequency data, the formula is

$$\text{S. D.} = \frac{\sqrt{\sum \text{f}_\text{i}(\text{x}_\text{i} - \bar{\text{x}})^2}}{n}$$

For grouped frequency data, the formula is

$$\text{S. D.} = \frac{\sqrt{\sum(\text{m}_\text{i} - \bar{\text{x}})^2}}{n}$$

where N is the total frequency;

f_i is the frequency corresponding to the i^{th} observation;

m_i be the midpoint of the i^{th} class interval.

Ques: Calculate standard deviation of the following dataset. 30, 40, 50, 55, 60, 65, 70, 80, 90, 100

Ans: First calculate the mean of the observations i.e., 64. Now subtract each value from 64 and take square. Add the square for all observation i.e. $(30 - 64)^2 + (40 - 64)^2 + \cdots + (100 - 64)^2$ which is 4290. Divide this sum by 10 and take a square root.

Standard deviation value is $\sqrt{4290/10} = 20.712$.

As the observations are very far away from the mean value. Hence the standard deviation is coming out to be very high.

Table 7: Calculate the standard deviation for the following dataset.

Observations	Frequency
38	7
43	9
46	10
49	6
51	4
54	8
67	3
78	5

Ans: Calculate the mean of the observations as shown in Table 6. Subtract the mean from the observations as shown in 4th column. Now take square of these observations as shown in the 5th column in Table 6. Column 6 shows the product of frequencies with values from column 5. Now take sum of the observations in column 5 and divide it by N i.e., total frequency. Take square root of the value 125.22 and the require standard deviation is 11.190.

Table 8: Calculation for standard deviation

Observations (x_i)	Frequency (f_i)	$f_i\ x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i * (x_i - \bar{x})^2$
38	7	266	-12.65384	160.1198225	1120.838757
43	9	387	-7.653846	58.58136095	527.2322485
46	10	460	-4.653846	21.65828402	216.5828402
49	6	294	-1.653846	2.735207101	16.4112426
51	4	204	0.346153	0.119822485	0.479289941
54	8	432	3.346153	11.19674556	89.5739645
67	3	201	16.34615	267.1967456	801.5902367
78	5	390	27.34615	747.8121302	3739.060651

	N = 52	$\sum f_i x_i = 2634$			6511.769231

$$\text{Mean} = \frac{\sum \text{fixi}}{\text{N}} = \frac{2634}{52} = 50.65384615$$

$$\text{S. D.} = \frac{\sqrt{\sum \text{f}_\text{i}(\text{x}_\text{i}-\bar{\text{x}})^2}}{n} = \frac{\sqrt{6511.769231}}{52} = \frac{125.2263314}{52} = 11.19045716$$

8.3.1.2 Relative Measure of Dispersion: Measures that give the dispersion values in terms of ratio and percentage are called relative measures.

A relative measure of dispersion is the ratio of absolute measure of dispersion with its appropriate average. These measures are independent of units and these are termed as coefficient of dispersion. One important thing, while the calculation of relative measures, is that the units of absolute measure and the appropriate average must be same.

(a) Coefficient of Range: This measure of dispersion is evaluated from the range of the data set. First range of the data set is calculated then,

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

Where, H is the highest value in the dataset;

L is the lowest value in the dataset.

(b) Coefficient of quartile deviation: This measure is defined as the ratio of quartile deviations to its average value. It is defined as

$$\text{Coefficient of Quartile Deviation} = \frac{\text{Q}_3 - \text{Q}_1}{\text{Q}_3 + \text{Q}_!}$$

where $Q1$ is the first quartile;

$Q3$ is the third quartile.

Ques: Find out the quartile deviation from the following data set related to the marks obtained by 15 students in statistics.

60, 67, 56, 78, 92, 55, 72, 54, 49, 59, 37, 84, 83, 69, 62

Ans: Arrange the observations in the ascending order 37, 49, 54, 55, 56, 59, 60, 62, 67,69, 72,78, 83, 84, 92

Compute the range that is 92-37 = 55 Coefficient of Range is = 55/ 129 = 42.63%.

Now compute the first quartile $\frac{(N+1)^{th}}{4}$ term i.e., 55

Similarly, third quartile as 12th term i.e., 78. Quartile deviation is (78 – 53)/2 = 25/2 = 12.5. Coefficient of quartile deviation = 25/131=19.08%

Hence coefficient of range and coefficient of quartile deviation is 42.63% and 19.08% respectively.

(c) **Coefficient of Mean Deviation:** This relative measure of dispersion is derived from the mean deviation. Mean deviation is computed as the absolute value of difference of observations from a central tendency i.e., mean, median and mode. Mostly, mean deviation is calculated by taking deviation of observations from mean and median. So, in order to convert the mean deviation measure into independent of unit coefficient of mean deviation is computed.

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean Deviation}}{\text{Central Tendency Measure}}$$

Central tendency measure can be mean or median but it is divided by the measure that is used for the derivation of mean deviation or from which the mean deviation is derived.

(d) **Coefficient of Standard Deviation:** As standard deviation is evaluated in terms of the observations units and is considered as a absolute measure of dispersion. It is essential for the comparison purpose that the measure must be independent of units. The relative measure based on standard deviation that is independent of units is called coefficient of standard deviation. It is defined as,

$$\text{Coefficient of Standard Deviation} = \frac{\sigma}{\bar{x}}$$

As coefficient of standard deviation would be given in fraction. So if we want to express our coefficient value in term of percentage by multiplying the coefficient by 100. Then this relative measure is called coefficient of variation (C.V.). It is defined as,

$$\text{Coefficient of Standard Deviation} = \frac{\sigma}{\bar{x}} \times 100$$

The coefficient of variation is among the most popular relative measure of dispersion. It is basically used to compare the variability among two or more dataset. The dataset that has more value of coefficient of variation among two is said to be more variable and vice versa.

Now, if one is interested to know about the concentration of the observations around a central tendency measure then it is essential to study two more measures. These are Skewness and Kurtosis. These two measures are considered as a supportive measure for better understanding the characteristics of the data.

8.4 Skewness

A skewness is basically to see tendency of the shape of the distribution. If the frequency distribution of the data is not equally distributed about the mean i.e. the frequency distribution is not symmetric then the term that is used to refer this situation is called skewness. Skewness has many synonyms like asymmetry and lack of symmetrical. Some authors give definitions of skewness as:

"When a series is not symmetrical, it is said to be asymmetrical or skewed" by Croxton and Cowden.

"Measure of skewness tell us the direction and the extent of skewness. In symmetrical distribution, the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness" by Simpson and Kafka.

Hence skewness means that the data is not symmetrical about the mean. It is also be defined in term of normal distribution. Normal distribution is the distribution which has mean, median and mode all are equal. Hence the shape of the frequency of this distribution is like bell shape.

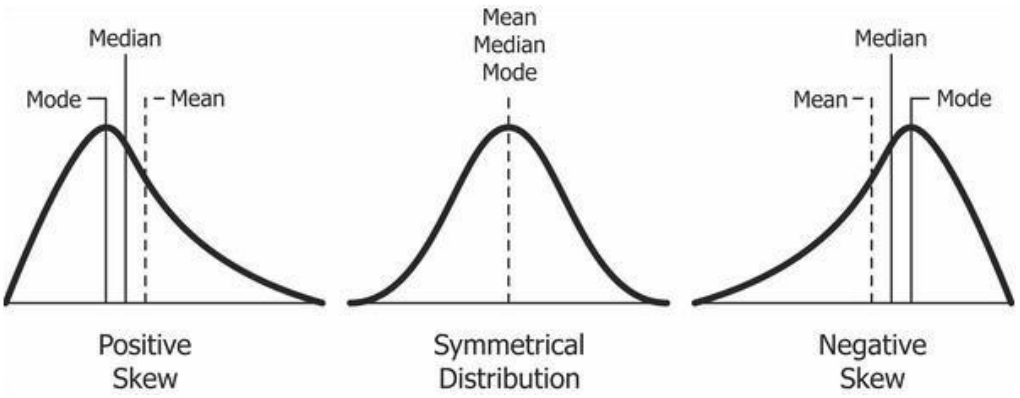


Figure 3: Frequency Distribution to understand Skewness.

- A frequency distribution is said to be positive skewed when the mean (μ) > Median > Mode. In this case, the value of mean is more than the value of median and mode. Also, median value is more than the value of mode.
- A frequency distribution is said to be symmetric distribution when the mean (μ) = Median = Mode. In this case, the values of mean, median and mode all are same.
- A frequency distribution is said to be negative skewed distribution when the mean (μ) < Median < Mode. In this case, the value of mode is more than the median and mean. Also, the median is more than the mean.

Difference Between Skewness and Measure of Dispersion: There are some important differences between measure of dispersion and skewness. These are:

(i) Skewness is basically concerned about the shape of the frequency distribution while measures of dispersion are more concerned about the amount of variations.

(ii) Skewness shows the nature of data about its central value while dispersion try to measure up to what extent the central tendency value represents the whole data set.

(iii) It is possible that the data that is more dispersed but has symmetric frequency distribution. Hence in that case one can say that symmetric does not mean that variation is less.

(iv) Measures of dispersion are based on first and second order moments while skewness is based on first, second and third order moments.

This is the reason that both skewness and measures of dispersion are studied together in literature. As both measures help in understanding the features of the frequency distribution in depth.

Measure of skewness: The formula for Karl Pearson coefficient of skewness is given as,

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

8.5 Kurtosis

Kurtosis word comes from the Greek language with a meaning curved arching. Kurtosis is basically used to measure the peakedness of the frequency distribution. It is possible that two data set have same arithmetic mean, standard deviation and coefficient of skewness but still one has different concentration of values near the mode value. So, the distribution can have more peakedness than the usual normal distribution, less peakedness than the usual normal curve and equal to the normal distribution curve. So basically, kurtosis is a measure that compare the peakedness of the curve relative to the peakedness of a normal curve. So, kurtosis is basically used to measure the extent how the distribution is more peaked or less peaked than the normal distribution curve.

Many authors give the definitions of the kurtosis as:

“A measure of kurtosis indicated the degree to which a curve of a frequency distribution is peaked or flat topped” by Croxton and Cowden.

“Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution” by Spiegel.

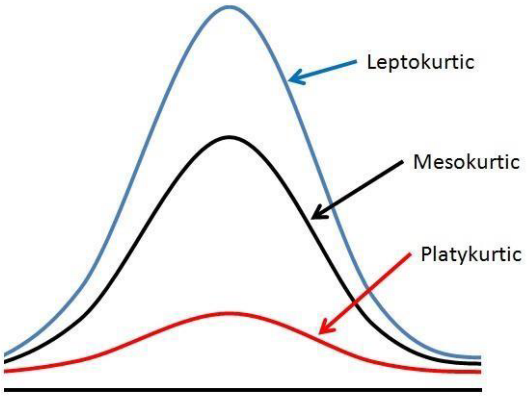


Figure 4: Kurtosis measurement through frequency distribution

So, if the distribution curve (blue curve) is more peaked than the normal distribution as shown with blue curve in Figure 4, then the distribution is called **Leptokurtic**. If the distribution curve (red curve) is flatter than the normal distribution curve then the distribution is called **Platykurtic**. Hence, the black curve represents the normal curve is also known as **Mesokurtic**.

Measure of Kurtosis: Kurtosis is defined as,

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

where, μ_4 is the 4th order moment about mean;

σ is the standard deviation.

- If $\beta_2 > 3$ then the distribution is Leptokurtic; If $\beta_2 = 3$ then the distribution is Mesokurtic; If $\beta_2 < 3$ then the distribution is Platykurtic.

Another measure that is used in literature is $\gamma_2 = \beta_2 - 3$.

Hence in this case,

- If $\gamma_2 > 0$ then the distribution is Leptokurtic; If $\gamma_2 = 0$ then the distribution is Mesokurtic; If $\gamma_2 < 0$ then the distribution is Platykurtic.

8.6 Measures of relationship

So far, we have deal with those statistical measures that we use in context of univariate population i.e., the population consisting of measurement of only one variable. But if we have the data on two variables, we are said to have a bivariate population and if the data happen to be on more than two variables, the population is known as multivariate population. Suppose we collect the data for the monthly income

and expenditure of the individuals in a group for each individual in the group we have a value for income (variable x) and expenditure (variable y) thus, we have a pair of values for each individual. this is an example of bivariate population. similarly, we can have multivariate population also. suppose, we have the data for experience (variable z) of the individuals in the same group. Now we have triplet of values for each individual in the group. corresponding population is a multivariate population.

When the population is based on two or more characteristics (variables), we may have like to measure the relationship between the variables. the measurement of relationship between two or more variables can give us some idea of the effect of one variable on the other.

A scatter plot is a useful graphical representation to have some approximate idea of relationship between two variables. scatter plot is obtained by plotting the pairs of observation taking one variable on x –axis and other on y-axis. suppose, for example we want to investigate the relationship between cigarette smoking and lung capacity. we might ask a group of people about their smoking habits, and measure their lung capacities.

Table 9: Dataset to check relationship between two variables

Cigarettes (X)	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29

The scatter plot representation of this data is presented below.

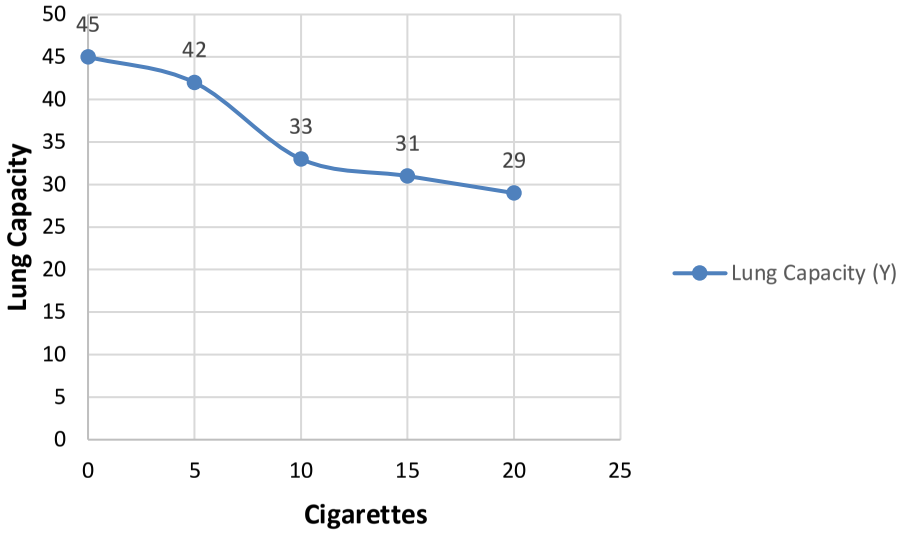


Figure 5: Scattered Plot to understand relationship between two variables

We can see that as smoking goes up, lung capacity tends to go down. the two variables change the values in the opposite direction. When the two variables change values in the opposite direction, we have negative correlation. when they change values in the same direction, they are said to be positively correlated. Some scatter plots are presented below exhibiting various types of relationship.

Association Of Case of Attributes:

When data is collected on the basis of some 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 is provides against small-pox. In other word, we may be interested in knowing whether inoculation and immunity from small-pox are associated.

Technically, we say that the two be expected 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 dissociation. 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 two attributes are positively associated; but if the class frequency of AB is less than the two attributes are said to be negatively associated. In case the class frequency pf 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:

If (AB)>$\frac{(A)}{N}\times\frac{(B)}{N}\times$N, then AB are positively related/associated.

If (AB)<$\frac{(A)}{N}\times\frac{(B)}{N}\times$, then, AB are negatively related/associated.

If (AB)=$\frac{(A)}{N}\times\frac{(B)}{N}\times$N, then AB are independent i.e., have no association.

Where, (AB)= frequency of class AN 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 so 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 +1and -1 if the attributes are completely associating with each other the coefficient will be +1 and if they are completely disassociating (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 coefficient of association are to 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 work out 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, $Q_{AB.C}$= 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 association may be 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 attributes 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 signification of association between two attributes, we make use of *Chi-square test** by finding the value of Chi-square (X^2) and using Chi-square distribution the value of (X^2) can be worked out as under:

$X^2=\sum \frac{(O_{ij}-E_{ij})2}{E_{ij}}$ i=1,2,3……and j= 1,2,3…..

Where, 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 each the two attributes are first observed and then each one is classified into two or more sub-classes, 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 of which has been further classified into sub – categories.

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(A_1) + (A_2) to form (A) and (A_3) + (A_4) to form (a) and similarly if we combine (B_1) + (B_2) to form (B) and (B_3) + (B_4) 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 8.3,

After reducing a contingency table in 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 for association in contingency tables. This can be obtained as under

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

Where, C=Coefficient of contingency

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

N= Number of items.

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

Other Measures:

(a) Index Numbers: When series are expressed in same unit we can use averages for the purpose of comparison but when in units in which two or more series are express happened to be different statistical averages cannot be used to compare them in such situation we have rely too upon some measurement which consist in reducing the figure to a common base once such method is to convert the series into a series of index number .this is done when this express the given figure as percentage of some specific figure on certain data we can thus define as index number which is used to measure the level of a given phenomenon as compare to the level of the same phenomenon at some standard date .the use of index number weights more as a special type of average meant to study the change in the effect of such factor which are incapable of measured directly ,but one must always remember that index number measured only the relative changes.

Changes in various economic and social phenomena can be measured and compared through index numbers. Different indices serve different purposes. Specific commodity indices are to serve as a measure of changes in the phenomena of that commodity only. Index number may measure cost of living of different classes of people. In economic sphere, index numbers are often termed as economic barometers measuring the economic phenomenon in all its aspects either directly by measuring the same phenomenon or indirectly by measuring something else which reflects upon the main phenomenon.

But index numbers have their own limitations with which researcher must always keep himself aware. For instance, index numbers are only approximate

indicators and as such give only a fair idea of changes but cannot give an accurate idea. Chances of error also remain at one point or the other while constructing an index number but this does not diminish the utility of index number for, they still can indicate the trend of the phenomenon being measured. However, to avoid fallacious conclusions, index numbers prepared for one purpose should not be used for other purposes or for the same purpose at other places.

(b) Time Series:

In the context of economic and business researchers, we may obtain quite often data relating to some time period concerning a given phenomenon. such data is labeled as time series, more clearly it can be state that series of successive observation of the given phenomenon over a period of time are referred to as time series. such series are usually the result of the effect of one or more of the following factors:

- Secular trend: or long-term trend that shows the direction of the series in a long period of time. The effect of trend (whether it happens to be a growth factor or a decline factor) is gradual, but extends more or less consistently throughout the entire period of time under consideration. Sometimes secular trend is simply state as trend or (T)
- short time oscillations i.e., change taking place in short period of time only and such change can be the effect of the following factors:
    (i) cyclic fluctuation: are the fluctuation as a result of business cycle and are generally referred as to long -term movement that represent consistently recurring rises and decline in an activity.
    (ii) seasonal fluctuation or (S): are the short duration occurring in a regular sequence at specific intervals of time. Such fluctuation is the result of changing seasons. Usually, these fluctuations involve patterns of change within a year that tend to be repeated from year to year. cyclical fluctuation and seasonal fluctuation taken together constitute short period regular fluctuation.

(iii) Irregular fluctuation (or) I: also known as random fluctuation, are variations which take place in a completely unpredictable fashion.

All these factors stated above are termed as component of time series and when we try to analyze time series, we try to isolate and measure the effects of various type of these factors on a series .to study the effect of one type of factor, the other type of factor is elimination from the series the given series is thus, left with the effect of one type of factor only.

For analysis time series, we usually have two model;

Multiplicative model: Multiplicative model assume that the various components interact in a multiplicative manner to produce the given values of the overall time series and can be state as under:

$$Y=T \times C \times S \times I$$

Where, y=observed value of time series

T=trend

C=cyclical fluctuations

S= seasonal fluctuations

I= irregular fluctuations.

Additive model: Additive model considers the total of various components resulting in the given values of the overall time series and can be stated as:

$$Y=T + C + S + I$$

There are various methods of isolating trend from the given series viz.., free hand method, semi –average method, method of moving average, method of least squares and similarly there are method of measuring cyclical and seasonal variation and whatever variations are left over are considered as random or irregular fluctuations.

The analysis of time series is done to understands the dynamic conditions for achieving the short –term and long –term goals of business firm (s). The past trends can be used to evaluate the success or failure of management policy or policies

practiced hitherto. On the basis of past trends, the future patterns can be predicted and policy or policies may accordingly be formulated. We can as well study properly the effect of factors causing change in the short period of time only, once we have eliminate the effects of trend .by studying cyclical variation ,we can keep in view the impact of cyclical change while formulating various policies to make them as realistic as possible .the knowledge of seasonal variation will be of great help to us in taking decisions regarding inventory ,production ,purchases and sales policies so as to optimize working result .thus analysis of time series in important in context of long term as well as short term forecasting and is considered a very powerful tool in the hands of business analysis and researchers.

8.7 Summary

As from above, the statistics has great importance in almost every field whether it is social sciences, natural sciences and industry. The reason behind is as statistics methodology is based on analysis of data (sample or whole population). Due to increase in complexity and volume of data, it is difficult to extract information easily. So, consistent research is going on to develop methods and tools to overcome these problems. These days most of the analysis is done using statistical software and programming languages.

Terminal Questions

1. **Multiple Choice Type Questions:**

(a) What's the term for the middle number when you list numbers in order?

1) Mean
2) Median
3) Mode
4) Range

(b) Which of the following is NOT an absolute measure of dispersion?

1) Range
2) Quartile Deviation

3) Coefficient of Variation

4) Mean Deviation

(c) The coefficient of range is calculated as:

1) (Q3 - Q1) / (Q3 + Q1)

2) (H - L) / (H + L)

3) (Range) / (Mean)

4) (Mean Deviation) / (Median)

(d) Which of the following best describes a positive skewed distribution?

1) Mean = Median = Mode

2) Mean < Median < Mode

3) Mean > Median > Mode

4) Mean = Mode > Median

(e) In which type of distribution is the mean, median, and mode all equal?

1) Leptokurtic

2) Platykurtic

3) Symmetric Distribution

4) Negative Skewed

(f) What does a positive value of β2 indicate about a distribution?

1) Platykurtic

2) Mesokurtic

3) Leptokurtic

4) Symmetric

2. (a) Given the dataset 10, 15, 20, 25, 30, find the range.

(b) Calculate the mean deviation about the mean for the dataset: 5, 10, 15, 20.

(c) For the dataset with frequencies provided, calculate the coefficient of variation if the standard deviation is 8 and the mean is 50.

(d) If the mean is 70, the median is 65, and the mode is 60, determine if the distribution is positively skewed, negatively skewed, or symmetric.

(e) Given $\beta 2 = 4$, what type of distribution does this represent?

(f) If the standard deviation is 10 and the mean is 80 while the mode is 70, calculate the coefficient of skewness.

3. (a) What is the formula for calculating the Karl Pearson coefficient of skewness?

(b) Define kurtosis and explain its significance in statistical analysis.

(c) What does skewness describe in the context of a frequency distribution?

(d) How do Croxton and Cowden define skewness?

(e) What is the main difference between skewness and measures of dispersion?

Answers

1. (a) 2) Median

(b) 3) Coefficient of Variation

(c) 2) (H - L) / (H + L)

(d) 3) Mean > Median > Mode

(e) 3) Symmetric Distribution

(f) 3) Leptokurtic

2. (a) 20

(b) Mean=12.5. Mean deviation = 5.0

(c) Coefficient of Variation = 16%

(d) Positively skewed (since Mean > Median > Mode)

(e) Leptokurtic (since $\beta 2 > 3$)

(f) Coefficient of Skewness = 1

3. (a) The formula is (Mean - Mode) / Standard Deviation, which can also be expressed as 3(Mean - Median) / Standard Deviation.

(b) Kurtosis measures the peakedness or flatness of a frequency distribution relative to a normal distribution. It helps in understanding how much the data clusters around the mean compared to a normal distribution.

(c) Skewness describes the tendency of the shape of the frequency distribution. It indicates whether the data is symmetrically distributed around the mean or not. If the distribution is not symmetric, it is referred to as skewed.

(d) Croxton and Cowden define skewness as a situation where a series is not symmetrical, referring to it as asymmetrical or skewed.

(e) Skewness concerns the shape of the frequency distribution, specifically its symmetry or lack thereof, while measures of dispersion focus on the amount of variation in the data.

Unit 9: Parameter, sampling and non-sampling error, Sampling distribution, degree of freedom, standard deviation and error; Correlation and regression; Statistical inference (point and internal estimation, sample size determination and hypothesis testing)

Unit Structure

9.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- About sampling and its types?
- Sampling distribution
- Degree of freedom
- Standard deviation and error
- Correlation and Regression
- Statistical inference

9.1. Introduction

Sampling error is the error that arises in a data collection process as a result of taking a sample from a population rather than using the whole population. Sampling error is one of two reasons for the difference between an estimate of a population parameter and the true, but unknown, value of the population parameter. The sampling error for a given sample is unknown but when the sampling is random, for some estimates (for example, sample mean, sample proportion) theoretical methods may be used to measure the extent of the variation caused by sampling error." Sampling error is mainly cause due to the reason that sample not whole population. Non-sampling error is the error that arises in a data collection process as a result of factors other than taking a sample. Non-sampling errors have the potential to cause bias in polls, surveys or samples. There are many different types of non-sampling errors and the names used to describe them are not consistent. This may be due to poor sampling method, measurement errors, and behavioral effect. In this unit you will learn about Parameter, sampling and non-sampling error, Sampling distribution, degree of freedom, standard deviation and error; Correlation and regression; Statistical inference (point and internal estimation, sample size determination and hypothesis testing.

9.2. Parameter and Sampling

Sampling is defined as the section of some part of an aggregates or totality on the basis of which a judgment or inference about the aggregate or totality is made. In other words, it is the process of obtained informing about an entire population by examining only a part of it .in most of the researchers work and survey, the usual approach happens to be to make generalizations or to draw inferences based on samples about the parameters of population from which the samples are taken. The researchers quite often select only a few items from the universe for his study purposes. all this is done on the assumption that the sample data will enable him to estimate the population parameters. the items so selected constitute what

is technically called a sample, there selection process or technique is called sample design and the survey conducted on the basis of sample is described as sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions. Sampling is used in practice for a variety of reasons such as:

1. Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.
2. Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.
3. Sampling remains the only choice when a test involves the destruction of the item under study.
4. Sampling remains the only way when population contains infinitely many members.
5. Sampling usually enables to estimate the sampling errors and thus, assists in obtaining information concerning some characteristic of the populations.

The collection of all the items about which the information (on one or more characteristic) is desired, is called population. For example, if we want to measure the average cell phone bill of the people in a particular city in a particular month, our population would consist of the specified month's cellphone bills of all the people in that city. The population or universe can be *finite* or *infinite. The* population is said to be finite if it consists of a fixed number of elements so that it is possible to enumerate it in its totality. For instance, the population of a city, the number of workers in a factory, the examples of finite populations. The symbol '*N*' is generally used to indicate how many elements (or items) are there in case of a finite population. An infinite population is that population in which it is theoretically impossible to observe all the elements. Thus, in an infinite population the number of items is infinite i.e., we cannot have any idea about the total number of items. The number of stars in sky,

possible rolls of a pair of dice are examples of infinite population. One should remember that no truly infinite population of physical objects does actually exist in spite of the fact that many such population appear to be very large. From a practical consideration, we then use the term infinite population.

Any characteristic or measure of population units is known as a parameter. Population mean, population standard deviation and population proportion are commonly studied parameters. These parameters are denoted by $\mu, \sigma \; and \; \pi$ respectively. Parameters are usually unknown as we do not always study the whole population. Unknown parameter is estimated by studying a subpart of the whole population which is known as a sample.

Any characteristic or measure of sample items is known as a specific. Sample mean (X), sample standard deviation (s_1) and sample proportion (p) are the examples of statistics. Obtaining estimate of an unknown parameter using a statistic and studying the properties of the obtained estimate are the prime objective of statistical inference.

9.3. Sampling and Non-Sampling Errors

Sample surveys do imply the study of a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be termed as sampling error or error variance. In other words, sampling errors are those errors which arise on account of sampling and they generally happen to be random variations.

Sampling error occurs randomly and are equally likely to be in either direction. The magnitude of the sampling error depends upon the nature of the universe: the more homogenous the universe smaller the sampling error. Sampling error is universally related to the size of the sample i.e., sampling error decrease as the sample size increase and vice versa. A measure of the random sampling error can be calculated for a given sample design

and size and this measure is often called the precision of the sampling plan. sampling error is usually worked out as the product of the critical value at a certain level of significance and the standard error.

As opposed to sampling errors, we may have non sampling errors which may creep in during the process of collecting actual information and such error occur in all surveys whether census or sample. we have no way to measure non sampling errors.

Sampling Distribution

Distribution of a statistic may not be the same as the distribution of population. We are often concerned with sampling distribution in sampling analysis. If we take certain number of sample and for each sample compute various statistical measure such as mean, standard deviation., then we can find that each sample may give its own value for the statistic under consideration. all such values of a particular statistic, say mean. together we their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure.

Here, it is important to discuss normal distribution briefly:

In many statistical analyses, it is important that the given data has normal distribution. normal distribution is denoted by N (μ, σ) where μ is mean and σ is standard deviation of the distribution. Normal distribution can be used to model any variable taking value $-\infty \text{ to } \infty$, provide the value on that variable have a bell-shaped histogram. Reader may refer to same standard text on statistics to know more about normal distribution. Some commonly used sampling distribution are given below:

Sampling distribution of mean

Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population if sample are taken from

a normal population an (μ, σ) ,the sampling distribution of mean would also be normal with mean $\mu\, x = \mu$and standard deviation =$\sigma\, \sqrt{n}$, where μ is the mean of population which is not normal (may be positive or negative skewed),even than as per the central limit theorem , the sampling distribution of mean tends quite closer to the normal distribution , provided the number of sampling items is large i.e., more than 30 .in case we want to reduce the sampling distribution of mean to unit distribution i.e., N (0,1) ,we can write the normal variate z=$\frac{x-\mu}{\sigma\sqrt{n}}$ for the sampling distribution of mean .this characteristics of the sampling distribution of mean is very useful in several decisions situations.

Sampling distribution of proportion

Proportion is the measure of the attributes. let as consider to that the population is divided into two mutuality exclusive and collecting exhaustive classes – one class possessing a particular attribute while other class not possessing that attribute. for example, people in a city could be divided into "smokers" and "Non –smokers". Let N = population size, X = number of people out of N possessing a particular attribute, $\pi = \frac{X}{N}$ = actual proportion of the people possessing the specified attributes. Let a sample is selected from this population with n= sample size, x=number of people in the sample possessing the specified particular attributes, p=x/n=sample proportion. note that, X and p is population parameters, while x and p is sample statistics. Also, p provides an estimate of π. it can be shower that the distribution of x is binomial (n,π). Using the property of binomial distribution, for sufficiently large sample size n, we have

$$Z = \frac{X-N\pi}{\sqrt{n\pi(I-n)}} = \frac{p-\pi}{\sqrt{\pi(I-\pi)/n}} \sim N(0,1)$$

Practically, this result is true for n≥ 30 or, when nπ ≥5 as well as n (1-π)≥5.

Student's t-distribution

When population standard deviation (σ) is not known and the sample is of a small size (i.e.,≤30) we use t-distribution for the sampling distribution of mean and workout t-variable as:

$t = \frac{\bar{x}-\mu}{s_1/\sqrt{n}}$, where $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is sample variance.

Student's t-distribution (or t-distribution) is also symmetrical and is very close to the distribution of standard normal variate, z, except for small values of n. the variable t differs from z. in the sense that we use sample standard deviation (s^2) in the calculation of t, whereas we use standard deviation of population (σ) in the calculation of z. there is a different t-distribution for every possible sample size i.e., for different degrees of freedom. The degrees of freedom for a sample of size n are n- 1. As the sample size gets larger, the shape of the t-distribution becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the t- distribution is so close to the normal distribution that we can use the normal to approximate the t- distribution. Like normal distribution, the range of t- distribution is also $(-\infty, \infty)$. The applications of t- distribution are discussed in further chapters.

Chi –square (χ^2) distribution

Chi-square distribution is encountered when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and thus have distribution that is related to chi-square distribution. Suppose we have a random sample $x_{1.....}x_{2...,}$ x_n from N (μ, π) population than the distribution of statistic $\sum_{i=1}^{n}\left(\frac{x_i-x}{\sigma}\right)^2$ is chi –square distribution with n degree of freedom. Alternatively, the distribution of the statistic $\sum_{i=1}^{n}\left(\frac{x_i-x}{\sigma}\right)^2 = \frac{(n-1)s^2}{\sigma^2}$ is chi-square distribution with (n-1) degree of freedom. Clearly the range of chi –square distribution includes all non –negative

values. Chi-square distribution is used in taking decision about the variance, testing independence of attributes and in testing goodness of fit these applications of chi –square distribution is discussed in further chapters.

Snedecor's *F*-Distribution

Let X and Y be two independent sample statistics such that X has chi-square distribution with d_1 degree of freedom and Y has chi-square distribution with d_2 degree of freedom. Then distribution of the statics $F=\frac{X/d_1}{Y/d_2}$ is Snedecor's *F*- distribution with d_1 and d_2 of freedom. Then range of this distribution is $(0,\infty)$.

Alternatively, suppose that we have two independent sample (x_1, x_2, ….., x_n and y_1, y_2.....y_{n2}) of sizes n_1 and n_2 respectively from normal population having the same variance. Then distribution of the statistic $F = \frac{\sum_{i=1}^{n_1}(x_i-x)^2/(n_1-1)}{\sum_{j=1}^{n_2}(y_i-y)^2/(n_2-1)} = \frac{s_x^2}{s_y^2}$ is Snedecor's F- distribution with (n_1-1) and (n_2-1) degree of freedom. Here if s_x^2 and s_y^2 are the sample variances. F-ratio is computed in a way the larger variance is always in the numerator. F- distribution is used in comparing two variances and analysis of variance technique discussed later.

9.4. Degree of Freedom

The number of independent observation which makes up a statistic is known as the degrees of freedom (d,f) associated with that statistics . Degree of freedom is the number of values in the final calculation of a statistic that are free to vary. In general, d.f. of a statistic = number of independent Observation – number of parameters estimated.

9.5. Standard Error

We have seen that different samples of the same size from the same population will yield different values of statistic under consideration, say sample. A measure of the variability in

different values of sample mean is given by the Standard Error of the sample mean. Standard error of a statistic is the standard deviation of its sampling distribution. Standard error plays an important role in statistical hypothesis testing and interval estimation. Standard errors give an idea about the reliability and precision of the estimate. Also, standard error decreases when sample size is increased. Standard errors of some important statistics are given below.

TABLE

Statistic	Standard Error
Sample mean $x=\frac{1}{n}\sum_{i=1}^{n} x_i$	$\sigma/\sqrt{n}$
Sample variance $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$	$\sigma^2\sqrt{2/n}$
Sample variance $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$	$\sigma^2\sqrt{2/(n-1)}$
Sample proportion p=x/n	$\sqrt{p(1-p)/n}$

As can be noticed that standard error depends on (usually) unknown parameters. In order to get some approximated value of standard error, unknown population can be replaced by some feasible estimates.

9.6. Correlation and regression

Charles Spearman's coefficient of correlation (*or rank correlation*) is the technique of determining the degree of correlation between two variables in case of ordinal data where ranks are given to the different values of the variables. The main objective of this coefficient

is to determine the extent to which the two sets of ranking are similar or dissimilar. This coefficient is determined as under:

$$\text{Spearman's Coefficient of Correlation (or r)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference between ranks of i^{th} pair of the two variables; n = number of pairs of observations.

As rank correlation is a non-parametric technique for measuring relationship between paired observations of two variables when data are in the ranked form, we have dealt with this technique in greater details later on in the book in chapter entitled 'Hypotheses Testing II (Non-parametric tests)'. *Karl Pearson's coefficient of correlation* (or simple correlation) is the most widely used method of measuring the degree of relationship between two variables. This coefficient assumes the following:

(i) that there is linear relationship between the two variables;

(ii) that the two variables are casually related which means that one of the variables is independent and the other one is dependent; and

(iii) a large number of independent causes are operating in both variables so as to produce a normal distribution.

Karl Pearson's coefficient of correlation can be worked out thus. Karl Pearson's coefficient of correlation is also known as the product moment correlation coefficient. The value of '*r*' lies between ±1. Positive values of *r* indicate positive correlation between the two variables (i.e., changes in both variables take place in the statement direction), whereas negative values of '*r*' indicate negative correlation i.e., changes in the two variables taking place in the opposite directions. A zero value of '*r*' indicates that there is no association between the two variables. When *r* = (+) 1, it indicates perfect positive correlation and when it is (–)1, it indicates perfect negative correlation, meaning thereby that variations in independent

variable (*X*) explain 100% of the variations in the dependent variable (*Y*). We can also say that for a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then correlation will be termed as perfect positive. But if such change occurs in the opposite direction, the correlation will be termed as perfect negative. The value of '*r*' nearer to +1 or –1 indicates high degree of correlation between the two variables.

where,

X_i = i^{th} value of X variable

$\overline{X}$ = mean of X

Y_i = i^{th} value of Y variable

$\overline{Y}$ = mean of Y

n = number of pairs of observation of X and Y

σ_X = Standard Deviation of X

σ_Y = Standard Deviation of Y

In case we use assumed means (A_x and A_y for variables X and Y respectively) in place of true means, then Karl Person's formula is reduced to:

$$\frac{\frac{\sum dx_i . dy_i}{n} - \left(\frac{\sum dx_i \, . \, \sum dy_i}{n}\right)}{\sqrt{\frac{\sum dx_i^2}{n} - \left(\frac{\sum d\,x_i}{n}\right)^2}\sqrt{\frac{\sum dy_i^2}{n} - \left(\frac{\sum d\,y_i}{n}\right)^2}}$$

Where, $\sum dx_i = \sum(X_i - A_x)$

$\sum dy_i = \sum(Y_i - A_y)$

$\sum dx_i^2 = \sum(X_i - A_x)^2$

$\sum dy_i^2 = \sum(Y_i - A_y)^2$

$\sum dx_i . dy_i = \sum(X_i - A_x)\,(Y_i - A_y)$

n = number of pairs of observation of X and Y

this is the short cut approach for finding 'r' in case of ungrouped data. If the data happens to be grouped data (i.e., the case of bivariate frequency distribution), we shall have to write Karl Person's coefficient of correlation as under:

$$\frac{\frac{\sum f_{ij} . dx_i . dy_j}{n} - \left(\frac{\sum f_i dx_i \; . \; \sum f_j dy_j}{n}\right)}{\sqrt{\frac{\sum f_i dx_i^2}{n} - \left(\frac{\sum f_i d\,x_i}{n}\right)^2}\sqrt{\frac{\sum f_i dy_i^2}{n} - \left(\frac{\sum f_j d\,y_i}{n}\right)^2}}$$

Where, f_{ij} is a frequency of a particular cell in the correlation table and all other values are defined as earlier.

Simple Regression Analysis

Regression is the determination of a statistical relationship between two or more variables. In simple regression, we have only two variables, one variable (defined as independent) is the cause of the behaviour of another one (defined as dependent variable). Regression can only interpret what exists physically i.e., there must be a physical way in which independent variable *X* can affect dependent variable *Y*.

where the symbol *Y* denotes the estimated value of *Y* for a given value of *X*. This equation is known as the regression equation of *Y* on *X* (also represents the regression line of *Y* on *X* when drawn on a graph) which means that each unit change in *X* produces a change of *b* in *Y*, which is positive for direct and negative for inverse relationships. Then generally used method to find the 'best' fit that a straight line of this kind can give is the least-square method. To use it efficiently, we first determine

$$\sum x_i^2 = \sum X_i^2 - n\overline{X}^2$$

$$\sum y_i^2 = \sum Y_i^2 - n\overline{Y}^2$$

$$\sum x_i y_i = \sum X_i Y_i - n\overline{X}.\overline{Y}$$

Then

$$b = \frac{\sum x_i y_i}{\sum x_i^2}, a = \overline{Y} - b\overline{X}$$

These measures define a and b which will give the best possible fit through the original X and Y points and the value of r can then be worked out as under:

$$r = \frac{b\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}}$$

Thus, the regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables which can be used for the purpose of prediction of the values of dependent variable, given the values of the independent variable. [Alternatively, for fitting a regression equation of the type $

Y = *a* + *bX* to the given values of *X* and *Y* variables, we can find the values of the two constants viz., *a* and *b* by using the following two normal equations:

$$\sum Y_i = na + b\sum X_i$$

$$\sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

and then solving these equations for finding *a* and *b* values. Once these values are obtained and have been put in the equation *Y* = *a* + *bX*, we say that we have fitted the regression equation of *Y* on *X* to the given data. In a similar fashion, we can develop the regression

equation of *X* and *Y* viz., $X = a + bX$, presuming *Y* as an independent variable and *X* as dependent variable.

Multiple Correlation and Regression

When there are two or more than two independent variables, the analysis concerning relationship is known as multiple correlations and the equation describing such relationship as the multiple regression equation. We here explain multiple correlation and regression taking only two independent variables and one dependent variable (Convenient computer programs exist for dealing with a great number of variables). In this situation the results are interpreted as shown below:

Multiple regression equation assumes the form

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

Where X_1 and X_2 are two independent variables and Y being the dependent variable, and the constants a, b_1 and b_2 can be solved by solving the following three normal equations:

$$\sum Y_i = na + b_1 \sum X_{1i} + b_2 \sum X_{2i}$$

$$\sum X_{1i} Y_i = a \sum X_{1i} + b_1 \sum X_{1i}^2 + b_2 \sum X_{1i} X_{2i}$$

$$\sum X_{2i} Y_i = a \sum X_{2i} + b_1 \sum X_{1i} X_{2i} + b_2 \sum X_{2i}^2$$

(It may be noted that the number of normal equations would depend upon the number of independent variables. If there are 2 independent variables, then 3 equations, if there are 3 independent variables then 4 equations and so on, are used.) In multiple regression analysis, the regression coefficients (viz., b_1 b_2) become less reliable as the degree of correlation between the independent variables (viz., X_1, X_2) increases. If there is a high degree of correlation between independent variables, we have a problem of what is commonly described as the *problem of multi collinearity*. In such a situation we should use only one set

of the independent variable to make our estimate. In fact, adding a second variable, say X_2, that is correlated with the first variable, say X_1, distorts the values of the regression coefficients. Nevertheless, the prediction for the dependent variable can be made even when multi collinearity is present, but in such a situation enough care should be taken in selecting the independent variables to estimate a dependent variable so as to ensure that multi-collinearity is reduced to the minimum.

With more than one independent variable, we may make a difference between the collective effect of the two independent variables and the individual effect of each of them taken separately. The collective effect is given by the coefficient of multiple correlation,

$R_{y.x_1x_2}$ defined as under:

$$R_{y.x_1x_2} = \sqrt{\frac{b_1 \sum Y_i X_{1i} - n\bar{Y}\,\overline{X_1} + b_2 \sum Y_i X_{2i} - n\bar{Y}\,\overline{X_2}}{\sum Y_i^2 - n\bar{Y}^2}}$$

Alternatively, we can write

$$R_{y.x_1x_2} = \sqrt{\frac{b_1 \sum x_{1i} y_i + b_2 \sum x_{2i}\, y_i}{\sum Y_i^2}}$$

where,

$$x_{1i} = (X_{1i} - \overline{X_1})$$

$$x_{2i} = (X_{2i} - \overline{X_2})$$

$$y_i = (Y_i - \bar{Y})$$

and b_1 and b_2 are regression coefficients.

Partial Correlation

Partial correlation measures separately the relationship between two variables in such a way that the effects of other related variables are eliminated. In other words, in partial correlation

analysis, we aim at measuring the relation between a dependent variable and a particular independent variable by holding all other variables constant. Thus, each partial coefficient of correlation measures the effect of its independent variable on the dependent variable. To obtain it, it is first necessary to compute the simple coefficients of correlation between each set of pairs of variables as stated earlier. In the case of two independent variables, we shall have two partial correlation coefficients denoted $r_{yx_1 \cdot x_2}$ and $r_{yx_2 \cdot x_1}$ which are worked out as under:

$$r_{yx_1 \cdot x_2} = \frac{R^2_{y \cdot x_1 x_2} - r^2_{yx_2}}{1 - r^2_{yx_2}}$$

This measures the effort of x_1 on y, more precisely, that proportion of the variation of y not explained by x_2 which is explained by x1. Also,

$$r_{yx_2 \cdot x_1} = \frac{R^2_{y \cdot x_1 x_2} - r^2_{yx_1}}{1 - r^2_{yx_1}}$$

In which x_1 and x_2 are simply interchanged, given the added effect of x_2 on y.

These formulae of the alternative approach are based on simple coefficients of correlation (also known as zero order coefficients since no variable is held constant when simple correlation coefficients are worked out). The partial correlation coefficients are called first order coefficients when one variable is held constant as shown above; they are known as second order coefficients when two variables are held constant and so on.\

9.7. Statistical Inference

A statistical population is a collection of items or individual about which we wish to draw some conclusion. As discussed, population may be finite. It may also be hypothetical or existent. For example, consider the random experiment of tossing a coin till we get the first head. Here numbers of bulbs in a box Whenever we do not study the whole population of

concrete objects, e.g., some characteristics of that population, we study a sample and using the sample we make conclusion about the unknown characteristic of that population. Statistic is helpful in studying the sample and drawing meaningful inference from the sample. This requires probabilistic sample as statistical results are based on probability theory. Drawing inference from the probabilistic sample about unknown population parameters is called as statistical inference. Statistical inference is concerned mainly with two things- Hypothesis testing and estimation.

In hypothesis testing we test the claims made about unknown population parameters using sample. These claims are made using some past experience or beliefs. This topic is discussed in detail in further chapters.

Estimation means estimating unknown population using a sample. There are two types of estimates – point estimate interval estimate.

(a) Point Estimation

Point estimation is a single valued estimate of an unknown parameters. The statistic x= $\frac{1}{n}\sum_{i=1}^{n} x_i$ $(sample\ mean)$ is used to estimate population meanμ. So, sample mean is a point estimate of the population mean. The statistic $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$ and $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$ are used to estimate unknown population variance $\sigma^{2\cdot}$. Therefore, these two statistics s^2 and s_1^2 are point estimate of $\sigma^{2\cdot}$ Conventionally, x, s^2 and s_1^2 are called estimators and specific values (such as x =20, s^2 = 4) are called estimates of the parameters. An unknown parameter may be estimators w2hich one is better? To answer this question, we must know about the properties of a good estimator as given below:

Unbiasedness

Let us the following population of size four: 18,20,22 and 24. Clearly,

Population mean = (18+20+22+24)/4=21 and

Population variance = $[(18+21)^2+(20+21)^2+(22-21)^2+(24-21)^2]/4=5$.

Now we collect all possible samples of size two from this population and calculated sample mean $x = \frac{1}{n}\sum_{i=1}^{n} x_i$, sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$ and modified sample variance $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$ for each sample. Here n is the sample size and x_i denotes the sample observation. These values are presented in the following table

Samples	x	S^2	S_1^2
18,18	18	0	0
20,18	19	1	2
22,18	20	4	8
24,18	21	9	18
18,20	19	1	2
20,20	20	0	0
22,20	21	1	2
24,20	22	4	8
18,22	20	4	8
20,22	21	1	2
22,22	22	0	0
24,22	23	1	2
18,24	21	9	18
20,24	22	4	8
22,24	23	1	2

24,24	24	0	0
Average	21	2.5	5

We can notice that average of all sample mean values is the same as population mean 21. The average of all s^2 values is not the same as population variance, while the average of all s_1^2 values is the same as population variance 5.

The property of an estimate is known as unbiasedness. An estimate, say *t*, of a parameter,θ is known as an unbiased estimate of, θ if the average of all possible values of t (obtained from all possible samples of a given size) is the same as, θ This is denoted as E(t) = θ. Equivalently, the expected value of t (long run or overall average of t values) should be the same as the value of, θ . From example we can observe that sample mean is an umbiased estimate of population mean. Sample variance s^2 is not an unbiased estimate of population variance, while sample variance s_1^2 is an unbiased estimate of population variance. This is written as E (sample mean) = population mean, E(s^2) $\neq$ population variance, and E (s_1^2) = population variance.

Consistency

An estimator should approach the value of population parameter as the sample size becomes larger and larger. The property is referred to as the property of consistency. This is the most desirable property of an estimator. When sample size becomes population mean. \Also sample variance s^2 becomes population variance σ^2. For sufficiently large sample size n the values of s^2 and s_1^2 are almost the same. Thus, sample mean is consistent estimator or population mean. Both sample variances s^2 and s_1^2 are consistent estimators of population variance σ^2.

Sufficiency

An estimator should use as much as possible the information available from the sample. This property is known as the property of sufficiency. The details on this property are beyond the scope of this book.

Efficiency

An estimators should have a relatively small variance. This means that the most efficiency estimator, among a group of unbiased estimators, is one which has the smallest variance. This property is technically described as the property of efficiency.

Keeping in view the above stated properties, the researcher must select appropriate estimators (s) for his study. Below we present the point estimate of some important parameters

(b) Interval Estimation

In interval estimation, we obtain using a sample statistic at a desired level of confidence, e.g.-99% confidence. Meaning of 99% confidence is the probability that the obtain interval will possess the true unknown parameter is 0.99. This confidence cannot be 100% as the confidence interval in that case will be infinite ((-∞, ∞) or $(0, \infty)$ In point estimation we give a single value while in interval estimation we give an interval to estimate a parameter. In interval estimation sample to sample variation is also taken into consideration through standard error. General rule of obtaining interval estimate is describe below:

Let θ be the unknown parameter and let T be the unbiased point estimate of θ, i.e., E(T)= θ . Now fix a desired level of confidence denoted as (1-α) $\times$ 100%.Here α is the probability of Type 1 error (to be discussed later). Usually confidence level, α =0.05. The confidence interval

Terminal Questions

Q.1. Describe the sampling.

Q.2. Describe the sampling error.

Q.3. Describe correlation and regression.

References:

CR Kothari: A text book on Research Methodology second revised edition. New age international limited publisher. ISBN number 978-81-224-2488-1.

Unit 10: Testing of Hypothesis: Basic concepts, Procedure and testing of hypothesis, limitations of tests of hypotheses

Unit Structure

10.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- About hypothesis and its concepts.
- About the level of significance
- About the procedure and testing of hypothesis
- About the limitations of testing of hypothesis

10.1. Introduction

According to Lykken (1968) statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published. As you know that research is all about investigate something new. There may be many research problems in different subjects. In environmental studies, the research problems may be as: Impacts of human activities on

environmental components, biodiversity of different biogeographical regions, causes and consequences of environmental pollution etc. After deciding the research problem, researcher and his/her supervisor have to define hypothesis. As you know hypothesis is one of the basic steps in research methodology. Hypothesis is usually considered as the principal instrument in research. The main function of hypothesis is to suggest new experiments and observations. Many experiments are carried out with the deliberate object of testing hypotheses. Researchers often face condition wherein they are interested in testing hypotheses on the base of information and then take decisions on the basis of such testing. The hypothesis may not be proved absolutely, but in practice it is accepted if it has withstood a critical testing. Hypothesis gives us clear idea about the significance of the research. Generally, there are two hypothesis viz. Null hypothesis and alternative hypothesis. Hypotheses are taken in to consideration before the research. Before researchers explain how hypotheses are tested through different tests meant for the purpose, it will be appropriate to explain clearly the meaning of a hypothesis and the related concepts for better understanding of the hypothesis testing techniques. In this unit you will learn about the testing of hypothesis: basic concepts, procedure and testing of hypothesis, limitations of tests of hypotheses.

10.2. Basic concept of hypothesis testing

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.

- Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.
- The test tells the analyst whether or not his primary hypothesis is true.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

10.2.1. What is hypothesis?

Normally, when one talks about hypothesis, one simply means a simple assumption or some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he aims to determine. A 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 or accepted as highly probable in the light of established facts.

Often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable.

For example, consider statements like the following ones: "There is no pollution, where anthropogenic activities are minimum". There is hypothesis capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity. The Hypothesis may be following two types:

1. Null hypothesis (H_0): There is no impact of anthropogenic activities on environment
2. Alternative hypothesis (H_a): There is impact of anthropogenic activities on environment

10.2.2. Characters of Hypothesis

There are some important characteristics of hypothesis which are summarized in Fig-1 and also discussed below:

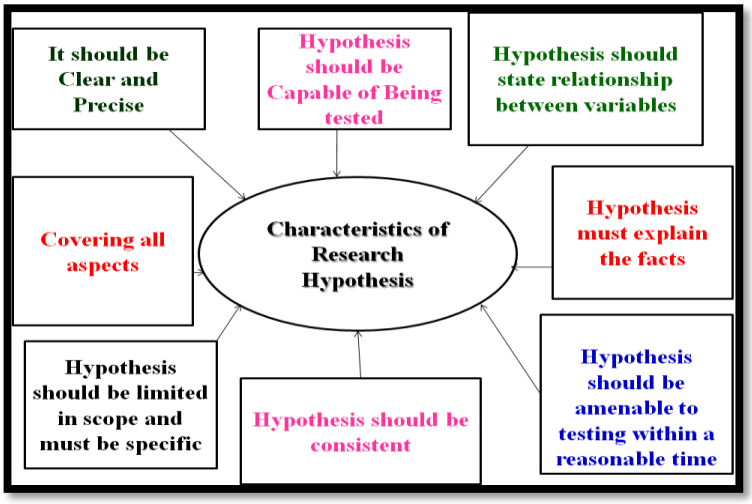


Fig-1: Showing Characteristics of Research Hypothesis

(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.

(ii) Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, many a time the research programmes have bogged down. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."

(iii) Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.

(iv) Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.

(v) Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.

(vi) Hypothesis should be consistent with most known facts i.e.; it must be consistent with a substantial body of established facts. In other words, it should be one which judge accept as being the most likely.

(vii) Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.

(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should be able to deduce the original problem condition. Thus, hypothesis must actually explain what it claims to explain; it should have empirical reference.

10.2.3. Null Hypothesis and alternative hypothesis

Definition of Null hypothesis: A null hypothesis is a statistical hypothesis in which there is no significant difference exists between the set of variables. It is the original or default statement, with no effect, often represented by H_0 (H-zero). It is always the hypothesis that is tested. It denotes the certain value of population parameter such as μ, s, p. A null hypothesis can be rejected, but it cannot be accepted just on the basis of a single test.

Definition of alternative hypothesis: A statistical hypothesis used in hypothesis testing, which states that there is a significant difference between the set of variables. It is often referred to as the hypothesis other than the null hypothesis, often denoted by H_1 (H-one). It is what the researcher seeks to prove in an indirect way, by using the test. It refers to a certain value of sample statistic, e.g., $\bar{x}$, s, p. The acceptance of alternative hypothesis depends on the rejection of the null hypothesis i.e., until and unless null hypothesis is rejected, an alternative hypothesis cannot be accepted. In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method *A* with method *B* about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method *A* is superior or the method *B* is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a.

If our sample results do not support this null hypothesis, we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as

alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H_0, then we are rejecting H_a and if we reject H_0, then we are accepting H_a. For H_0, we may consider three possible alternative hypotheses as follows:

The null hypothesis and the alternative hypothesis are chosen before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data). In the choice of null hypothesis, the following considerations are usually kept in view:

(a) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

(b) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is 0.05 (the level of significance) which is chosen very small.

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

Table:1: Difference between Null Hypothesis (H_0) and Alternative Hypothesis (H_a)

BASIS	H_0	H_a
Meaning	It is a statement, in which there is no relationship between two variables.	It is statement in which there is some statistical significance between two measured phenomena.
Represents	No observed effect	Some observed effect
Aim of Researcher	Researcher tries to disprove H_0.	Researcher tries to prove H_a.
Acceptance	No changes in opinions or actions	Changes in opinions or actions
Testing	Indirect and implicit	Direct and explicit
Observations	Result of chance	Result of real effect
Mathematical sign	Equal sign	Unequal sign

10.2.4. Type I and Type II errors

Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent. In the context of testing of hypotheses, there are basically two types of errors we can make. We may reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by α (alpha) known as α error, also called the level of significance of test; and Type II error is denoted by β (beta) known as β error. In a tabular form the said two errors can be presented as follows:

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01. But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between two types of errors which means that the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result, one must set very high level for Type I error in one's testing technique of a given hypothesis.2 Hence, in the testing of

hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

Two-tailed and One-tailed tests: In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesized value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value not equal to the specified value of the null hypothesis. Symbolically, the two tailed test is appropriate when we have,

H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$

Decision rule or test of hypothesis: Given a hypothesis H_0 and an alternative hypothesis H_a, we make a rule which is known as decision rule according to which we accept H_0 (i.e., reject H_a) or reject H_0 (i.e., accept H_a). For instance, if (H_0 is that a certain lot is good (there are very few defective items in it) against H_a) that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept $H0$ otherwise we will reject H_0 (or accept H_a). This sort of basis is known as decision rule.

10.2.5. Level of Significance

This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that $H0$ will be rejected *If a hypothesis is of the type H_0: $\mu = \mu_0$, then we call such a hypothesis as simple (or specific) hypothesis but if it is of the type H_1: $\mu \neq \mu_0$, H_1: $\mu > \mu_{0\,o}$ or H_1: $\mu < \mu_0$; then we call it a composite (or nonspecific) hypothesis. When the sampling result has a less than 0.05 probability of occurring if H_0 is true. In other words, the 5 per cent level of significance means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesis when it (H_0) happens to be true. Thus, the significance level is the maximum

value of the probability of rejecting H_0 when it is true and is usually determined in advance before testing the hypothesis.

The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as *α*. That is, *P* (Type I error) = *α*.

Confidence level: Confidence level refers to the possibility of a parameter that lies within a specified range of values, which is denoted as *c*. Moreover, the confidence level is connected with the level of significance. The relationship between level of significance and the confidence level is c=1−α. The common level of significance and the corresponding confidence level are given below:

- The level of significance 0.10 is related to the 90% confidence level.
- The level of significance 0.05 is related to the 95% confidence level.
- The level of significance 0.01 is related to the 99% confidence level.

The rejection rule is as follows:

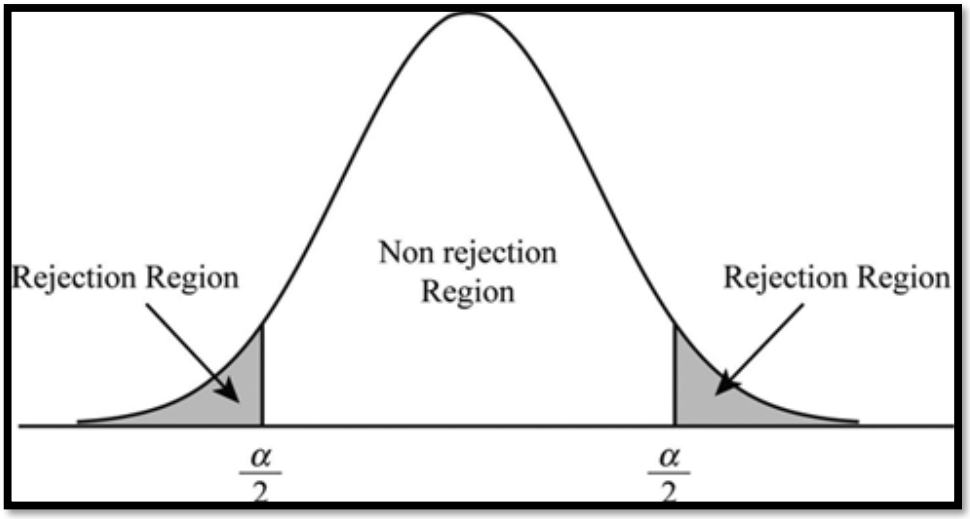
- If p-value ≤ level of significance (α, then reject the null hypothesis)
- If p-value > level of significance (α, then accept the null hypothesis)

Rejection region:

The rejection region is the values of test statistic for which the null hypothesis is rejected.

Non rejection region:

The set of all possible values for which the null hypothesis is not rejected is called the rejection region. The rejection region for two-tailed test is shown below:



10.3. Procedure and testing of hypothesis

To test a hypothesis means to tell (on the basis of the data the researcher has collected) whether or not the hypothesis seems to be valid. In hypothesis testing the main question is: whether to accept the null hypothesis or not to accept the null hypothesis? Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between the two actions i.e., rejection and acceptance of a null hypothesis. The various steps involved in hypothesis testing are stated below:

(i) Making a formal statement: The step consists in making a formal statement of the null hypothesis (H_0) and also of the alternative hypothesis (H_a). This means that hypotheses should be clearly stated, considering the nature of the research problem. For instance, Mr. Mohan of the Civil Engineering Department wants to test the load bearing capacity of an old bridge which must be more than 10 tons, in that case he can state his hypotheses as under: Null hypothesis H_0: $\mu=10$ tons Alternative Hypothesis H_a: $\mu\neq10$ tons

Take another example. The average score in an aptitude test administered at the national level is 80. To evaluate a state's education system, the average score of 100 of the state's students selected on random basis was 75. The state wants to know if there is a significant difference between the local scores and the national scores. In such a situation the hypotheses may be stated as under:

Null hypothesis H0: μ=80

Alternative Hypothesis Ha: μ≠80

The formulation of hypotheses is an important step which must be accomplished with due care in accordance with the object and nature of the problem under consideration. It also indicates whether we should use a one-tailed test or a two-tailed test. If *Ha* is of the type greater than (or of the type lesser than), we use a one-tailed test, but when *Ha* is of the type "whether greater or smaller" then we use a two-tailed test.

(ii) Selecting a significance level: The hypotheses are tested on a pre-determined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% level is adopted for the purpose. The factors that affect the level of significance are: (a) the magnitude of the difference between sample means; (b) the size of the samples; (c) the variability of measurements within samples; and (d) whether the hypothesis is directional or non-directional (A directional hypothesis is one which predicts the direction of the difference between, say, means). In brief, the level of significance must be adequate in the context of the purpose and nature of enquiry.

(iii) Deciding the distribution to use: After deciding the level of significance, the next step in hypothesis testing is to determine the appropriate sampling distribution. The choice generally remains between normal distribution and the *t*-distribution. The rules for selecting the correct distribution are similar to those which we have stated earlier in the context of estimation.

(iv) Selecting a random sample and computing an appropriate value: Another step is to select a random sample(s) and compute an appropriate value from the sample data concerning the test statistic utilizing the relevant distribution. In other words, draw a sample to furnish empirical data.

(v) Calculation of the probability: One has then to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.

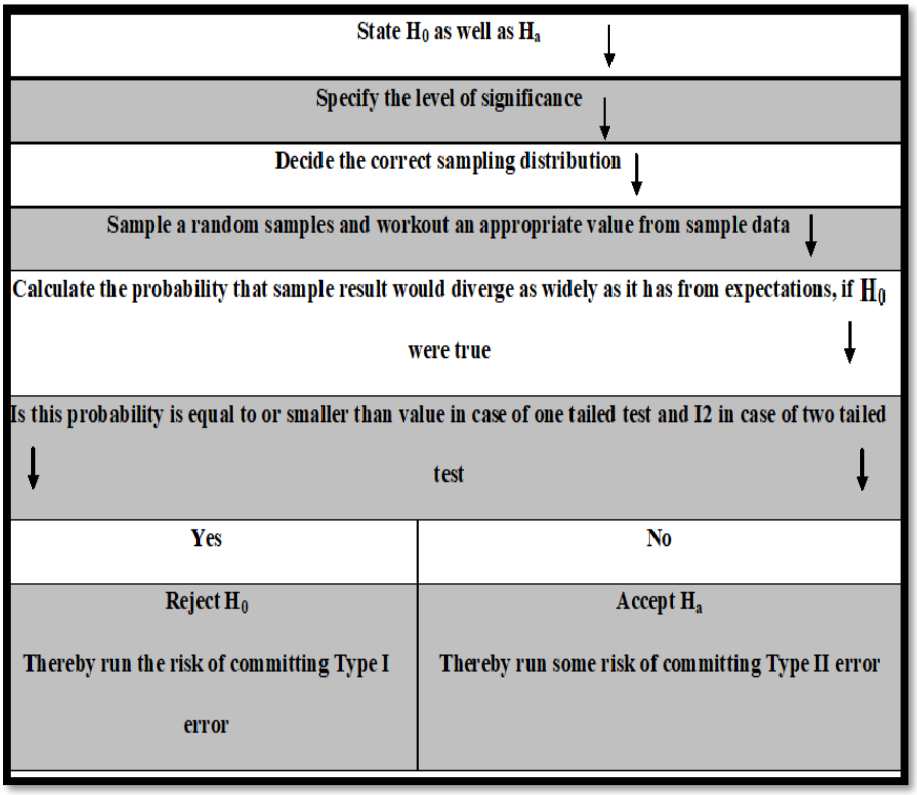


Fig-2: Showing Flow diagram for hypothesis testing

(vi) Comparing the probability: Yet another step consists in comparing the probability thus calculated with the specified value for $\propto$, the significance level. If the calculated probability is equal to or smaller than the $\propto$ value in case of one-tailed test (and $\propto/2$ in case of two-tailed test), then reject the null hypothesis (i.e., accept the alternative hypothesis), but if the calculated probability is greater, then accept the null hypothesis. In case we reject H_0, we run a risk of (at most the level of significance) committing an error of Type I, but if we accept H_0, then we run some risk (the size of which cannot be specified as long as the H_0 happens to be vague rather than specific) of committing an error of Type II. The above stated general procedure for hypothesis testing can also be depicted in the form of a flowchart for better understanding as shown in Fig-2

10.4. Limitations of tests of hypothesis

We have described above some important test often used for testing hypotheses on the basis of which important decisions may be based. But there are several limitations of the

said tests which should always be borne in mind by a researcher. Important limitations are as follows:

- The tests should not be used in a mechanical fashion. It should be kept in view that testing is not decision-making itself; the tests are only useful aids for decision-making. Hence "proper interpretation of statistical evidence is important to intelligent decisions."
- Test does not explain the reasons as to why do the difference exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.
- Results of significance tests are based on probabilities and as such cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not due to chance.
- Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypotheses. This is specially so in case of small samples where the probability of drawing erring inferences happens to be generally higher. For greater reliability, the size of samples be sufficiently enlarged. All these limitations suggest that in problems of statistical significance, the inference techniques (or the tests) must be combined with adequate knowledge of the subject-matter along with the ability of good judgment.

10.5. Summary

- There may be many research problems in different subjects. In environmental studies, the research problems may be as: Impacts of human activities on environmental components, biodiversity of different biogeographical regions, causes and consequences of environmental pollution etc. After deciding the research problem, research and his/her supervisor have to define hypothesis.
- Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst

depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.

- A 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 or accepted as highly probable in the light of established facts.
- Quite often a research hypothesis is a predictive statement, capable of being tested y scientific methods, that relates an independent variable to some dependent variable. For example, consider statements like the following ones: "Students who receive counselling will show a greater increase in creativity than students not receiving counselling" Or "the automobile A is performing as well as automobile B." These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.
- Hypothesis must possess certain characteristics such as : it hould be clear and precise, it should be capable of being tested, it should state relationship between variables, it should be limited in scope and must be specific, it should be consistent, it should be amenable to testing within a reasonable time.
- There are two types of hypotheses viz. Null Hypothesis and alternative hypothesis. If we are to compare method A with method B about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H0 and the alternative hypothesis as Ha.
- Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative

hypothesis. Hence the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

- The various steps involved in hypothesis testing are Making a formal statement, selecting a significance level, deciding the distribution to use, selecting a random sample and computing an appropriate value, Calculation of the probability and comparing the probability
- The tests should not be used in a mechanical fashion. It should be kept in view that testing is not decision-making itself; the tests are only useful aids for decision-making. Hence "proper interpretation of statistical evidence is important to intelligent decisions."
- Test does not explain the reasons as to why do the difference exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.

Terminal questions

1. (a) Fill the blank spaces with appropriate words

Normally, when one talks about……., one simply means a simple ……..or some ……to be proved or disproved. But for a researcher hypothesis is a formal question that he …….to determine. A hypothesis may be defined as "A …….or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some ……or accepted as highly probable in the light of established facts.

2. (a) What do you understand by hypothesis?

(b) Write about null hypothesis and alternative hypothesis.

3. (a) Write about level of significance.

(b) Give a note on Type I and type II errors

4. (a) Write about procedure of tests of hypothesis
5. (a) Differentiate between null and alternative hypothesis.

6. (a) Rejection of null hypothesis indicates that (Research is Significant/Research is not significant)

   (b) Which is/are character/s of a hypothesis? (Precise/Clear/understandable/All of the above)

7. (a) Describe the limitations of tests of hypothesis.

ANSWERS

1. (a) hypothesis, assumption, supposition, aims, proposition, investigation
2. (a) see the section 10.2.1

   (b) See the section 10.2.3
3. (a) See the section 10.2.5

   (b) See the section 10.2.4
4. (a) See the section 10.3.
5. (a) See the section 10.2.3 under Table-1
6. (a) Research is Significant

   (b) All of the above
7. (a) See the section 10.4.

UNIT-11: Chi-square, t, F and z tests, Turkey's Q test; ANOVA and ANOCOVA

Unit Structure

11.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is Chi Square test?
- What is T-test?
- What is F-test?
- What is Z-test?
- What is Turkey's Q test?
- What is ANOVA?
- What is ANOCOVA?

11.1. Chi Square test

The chi-square (χ^2) statistic that is a test that measures how a model compares to actual observed data. This data is to be used in calculating a chi-square statistic must be random, series, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a fair coin meet these criteria.

Chi-square tests are also to be used in the measure hypothesis testing. The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.

- A chi-square (χ^2) statistic is a measure of the difference between the observed and expected frequencies of the results of a set of events or variables.
- Chi-square is useful for analyzing such differences in categorical variables, especially those nominals in nature.
- χ^2 depends on the basis of size of the difference between actual and outcome results, the degrees of freedom, and the samples size.
- χ^2 can be used to test whether two variables are related or independent from one another.
- It can also be used to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies.

The Formula for Chi-Square Is:

$$\chi^2=\sum\frac{(O_i-E_i)2}{Ei}$$

where: χ^2 = chi-square,

c = Degrees of freedom

O_i = Observed value(s)

E_i = Expected value(s)

A chi-square test is used to help determine if observed results are in line with expected results, and to rule out that observations are due to chance. A chi-square test is appropriate for this when the data being analysed is from a random sample, and when the variable in question is a categorical variable. A categorical variable is one that consists of selections

such as type of car, race, educational attainment, male vs. female, how much somebody likes a political candidate (from very much to very little), etc.

These types of data are often collected via survey responses or questionnaires. Therefore, chi-square analysis is often most useful in analyzing this type of data.

Chi-square is a statistical test used to examine the differences between categorical variables from a random sample in order to judge goodness of fit between expected and observed results.

Test of Goodness of Fit:

Test of goodness of fit is of a statistical test that is used to typically summarize the discrepancy between observed value and the expected value these types of measure can be used in testing of hypothesis.

Caution in Using Chi-Square Test:

This test is used in more frequently it should be borne in mind that the test to be applied for when the individual observation of sample is independent which means that the occurrence of one individual observation has no effect upon the occurrence of any other observations the sample under considered.

The other possible reason concerning the improper applications or misused of this test can be:

1. Neglect of frequencies of non –occurrence.
2. Failure to equalize the sum of observed and the sum of the expected frequencies.
3. wrong determination of the degree of freedom.
4. Wrong computation, and the like.

11.2. T-test

A t-test is a type of inferential statistic used to find out if there is a significant difference between the means of two groups, which may be related in certain characteristics. It is mostly used when the data sets, like the data set recorded as the results from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. it is used

as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to find out the statistical significance. To conduct a test with three or more means, one must use an analysis of variance.

Essentially, a t-test allows us to compare the average values of the two data sets and determine if they came from the same population. In the above examples, if we were to take a sample of students from class A and another sample of students from class B, we would not expect them to have exactly the same mean and standard deviation. Similarly, samples taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and standard deviation. Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement by assuming a null hypothesis that the two means are equal. Based on the applicable formulas, certain values are calculated and compared against the standard values, and the assumed null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance. The t-test is just one of many tests used for this purpose. Statisticians must additionally use tests other than the t-test to examine more variables and tests with larger sample sizes. For a large sample size, statisticians use a z-test. Other testing options include the chi-square test and the f-test. There are three types of t-tests, and they are categorized as dependent and independent t-tests.

Ambiguous Test Results

Consider that a drug manufacturer wants to the test a newly invented medicine. It follows the standard procedure of trying the drug on one group of patients and giving a place to another group, called the control group. The placebo given to the control group is a substance of no intended therapeutic value and serves as a benchmark to measure how the other group, which is given the actual drug, responds.

After the drug trial, the members of the placebo-fed control group reported an increase in average life expectancy of three years, while the members of the group who are prescribed the new drug report an increase in average life expectancy of four years. Instant observation

may indicate that the drug is indeed working as the results are better for the group using the drug. However, it is also possible that the observation may be due to a chance occurrence, especially a surprising piece of luck. A t-test is useful to conclude if the results are actually correct and applicable to the entire population.

T-Test Assumptions

1. The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.
2. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

Calculating T-Tests

Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.

The outcome of the t-test produces the t-value. This calculated t-value is then compared against a value obtained from a critical value table (called the T-Distribution Table). This comparison helps to determine the effect of chance alone on the difference, and whether the difference is outside that chance range. The t-test questions whether the difference between the groups represents a true difference in the study or if it is possibly a meaningless random difference.

T-Values and Degrees of Freedom

The t-test produces two values as its output: t-value and degrees of freedom. The t-value is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets. While the numerator value (the difference between the mean of the two sample sets) is straight forward to calculate, the denominator (the variation that

exists within the sample sets) can become a bit complicated depending upon the type of data values involved. The denominator of the ratio is a measurement of the dispersion or variability. Higher values of the t-value, also called t-score, indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets.

- A large t-score indicates that the groups are different.
- A small t-score indicates that the groups are similar.

Degrees of freedom refers to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

Correlated (or Paired) T-Test

The correlated t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.

This method also applies to cases where the samples are related in some manner or have matching characteristics, like a comparative analysis involving children, parents or siblings. Correlated or paired t-tests are of a dependent type, as these involve cases where the two sets of samples are related.

The formula for computing the t-value and degrees of freedom for a paired t-test is:

$$t= \frac{\overline{x_1}-\overline{x_2}}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

where:

mean1 and mean2=The average values of each of the sample sets

n=The sample size (the number of paired differences)

n−1=The degrees of freedom

The remaining two types belong to the independent t-tests. The samples of these types are selected independent of each other—that is, the data sets in the two groups don't refer to the same values. They include cases like a group of 100 patients being split into two sets of 50 patients each. One of the groups becomes the control group and is given a placebo, while the other group receives the prescribed treatment. This constitutes two independent sample groups which are unpaired with each other.

Equal Variance (or Pooled) T-Test

The equal variance t-test is used when the number of samples in each group is the same, or the variance of the two data sets is similar. The following formula is used for calculating t-value and degrees of freedom for equal variance t-test:

$$\text{T-value} = s_p^2 \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

where:

$mean_1$ and $mean_2$=Average values of each of the sample sets

var_1 and var_2=Variance of each of the sample sets

n_1 and n_2=Number of records in each sample set

and, Degrees of Freedom=n_1+n_2-2

where, n_1 and n_2=Number of records in each sample set

Unequal Variance T-Test

The unequal variance t-test is used when the number of samples in each group is different, and the variance of the two data sets is also different. This test is also called the Welch's t-test

The following formula is used for calculating t-value and degrees of freedom for an unequal variance t-test:

$$t= \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{s_1^2}{N1}+\frac{s_2^2}{N2}}}$$

where:

$mean_1$ and $mean_2$ = Average values of each of sample sets

var_1 and var_2=Variance of each of the sample sets

n_1 and n_2=Number of records in each sample set and,

11.3. F-test

F-test also given by Fisher.

F-test is used to the two-independent estimation of population variance.

Two sample have same variance (S_1 and S_2).

F-test is small sample test.

$$F=\frac{\text{Larger estimate of population variance.}}{\text{Smaller estimate of population variance.}}$$

$$\text{The variance ratio} =\frac{s_1^2}{s_2^2}$$

Degree of freedom for larger population variance is V_1 and smaller is V_2.

The null hypothesis of two population variance is equal that areH_0:$S_1^2 = s_2^2$

V_1(larger)n-1

V_2(smaller)n-2

F-test are design to test if two population variances are equal.

F-test is use by comparing the ratio of the two variance S_1 & S_2.

The samples must be independent.

F-test never be negative because upper value is greater than lower. $\frac{S_2^1}{S_2^2}$

Testing of overall significance of regression by – F-test. Test for the significance of the adjusted coefficient of multiple determination is –F-Test.

11.4. Z-test

A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.

The test statistic is assumed to have a normal distribution, and nuisance parameters such as standard deviation should be known in order for an accurate z-test to be performed.

T-test is used when correlation coefficient of population is zero, but of population coefficient, correlation is not zero than z test is used

Formula Used for z Test

$$z=\frac{zr-zp}{SEz}$$

z test is used to determine whether two population means are different when population variance is known. This is also called large sample test. This is also conducting many types of tests, these are

1. one sample test
2. Two sample test
3. Location test
4. paired difference test

characteristics of z test:

- A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.
- A z-test is a hypothesis test in which the z-statistic follows a normal distribution.
- A z-statistic, or z-score, is a number representing the result from the z-test.
- Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.
- Z-tests assume the standard deviation is known, while t-tests assume it is unknown.

Understanding Z-Tests

The z-test is also a hypothesis test in which the z-statistic follows a normal distribution. The z-test is best used for greater-than-30 samples because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed.

When conducting a z-test, the null and alternative hypotheses, alpha and z-score should be stated. Next, the test statistic should be calculated, and the results and conclusion stated. A z-statistic, or z-score, is a number representing how many standard deviations above or below the mean population a score derived from a z-test is.

Examples of tests that can be conducted as z-tests include a one-sample location test, a two-sample location test, a paired difference test, and a maximum likelihood estimate. Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size. Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known. If the standard deviation of the population is unknown, the assumption of the sample variance equaling the population variance is made.

11.5. Turkey's Q test

Turkey's range test also called Turkey's method, Turkey's honest significance test or Turkey's HSD (honestly significance difference). This test is based on a formula very similar to the test of t test:

Formula Used for Turkey's Test

$$qs=\frac{YA-YB}{SE}$$

where, YA= The larger of the two means being compared

YB= the smaller of the two means being compared

SE= the standard error

The Turkey HSD ("honestly significant difference" or "honest significant difference") test is a statistical tool used to find out if the relationship between two sets of data is statistically significant – that is, whether there's a major chance that an observed numerical change in one value is causally related to an observed change in another value.

In other words, the Turkey test is a way to test an experimental hypothesis.

The Turkey test is involved when you need to determine if the interaction among three or more variables is mutually statistically significant, which unfortunately is not simply a sum or product of the individual levels of significance.

Simple statistics problems involve looking at the effects of one (independent) variable, like the number of hours studied by each student in a class for a particular test, on a second (dependent) variable, like the student's scores on the test. In such cases, you usually set your cut-off for statistical significance at $P < 0.05$, wherein the experiment reveals a greater than 95 percent chance that the variables in question truly related. Then you refer to a t-table that takes into account the number of data pairs in your experiment to see if your hypothesis was correct.

Sometimes, however, the experiment may look at multiple independent or dependent variables simultaneously. For example, in the above example, the hours of sleep each student got the night before the test and his or her class grade going in might be included. Such multivariate problems require something other than a t-test owing to the sheer number if independently varying relationships.

11.6. ANOVA

ANOVA stands for "analysis of variance" and addresses precisely the problem just described. It accounts for the rapidly expanding degrees of freedom in a sample as variables are added. For example, looking at hours vs. scores is one pairing, sleep vs. scores is another, grades vs scores are a third and meanwhile, all of those independent variables interact with one another, too.

In an ANOVA test, the variable of interest after calculations have been run is F, which is the found variation of the averages of all of the pairs, or groups, divided by the expected variation of these averages. The higher this number, the stronger the relationship, and "significance" is usually set at 0.95. Reporting ANOVA results usually requires the use of a built-in calculator such as those found in Microsoft Excel as well as dedicated statistical programs such as SPSS.

"This test does is compare the differences between means of values rather than comparing pairs of values. The value of the Turkey test is given by taking the absolute value of the difference between pairs of means and dividing it by the standard error of the mean (SE) as determined by a one-way ANOVA test. The SE is in turn the square root of (variance divided by sample size). An example of an online calculator is listed in the Resources section.

The Turkey test is a post hoc test in that the comparisons between variables are made after the data has already been collected. This differs from an a priori test, in which these comparisons are made in advance. In the former case, you might look at the mile run times of students in three different classes one year. In the latter case, you might assign students to one of three teachers and then have them run a timed mile.

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method. ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers". It was employed in experimental psychology and later expanded to subjects that were more complex:

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to

$$F=\frac{MST}{MSE}$$

$$MST = \frac{\sum_{i-1}^{k}(T_i^2 n_i)-G^2\ n}{K-1}$$

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

The ANOVA test is the first step in analysing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

Example of How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer types I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

One-Way ANOVA and Two-Way ANOVA

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

11.7. ANOCOVA

Analysis of Covariance (ANOCOVA) Definition:

The statistical technique which reduces the experimental error by eliminating the effect of variation in the ancillary or concomitant variety, resulting in increase in the precision of the estimates of treatment means in known as Analysis of Co-variance.

Ancillary observations:

Ancillary observations are values of the characteristics of those extraneous sources of variation which influence the characteristic that is compared for each treatment. These sources of variation cannot be controlled in the experiment and vary from one experimental unit to another experimental unit, but can be measured numerically in each experimental unit. Such type of variable or characteristic is called ancillary variable or concomitant variant or Co-vitiate. Example:

In the experiment conducted to compare different levels of nitrogen on the yield of wheat. The ancillary variable in this experiment may be plant population, number of tillers, age of wheat crop or straw yield per plot etc., which are not practically controlled and vary randomly from plot to plot. Each of these ancillary observations influence directly the yield of wheat in each plot.

Terminal Question:

a. What is Chi Square test?

b. What is T-test?

c. What is F-test?

d. What is Z-test?

e. What is Turkey's Q test?

f. What is ANOVA?

g. What is ANOCOVA?

Answers:

a. see the section 11.1

b. see the section 11.2

c. see the section 11.3

d. see the section 11.4

e. see the section 11.5

f. see the section 11.5

g. see the section 11.5

UNIT 12. Linear Regression Analysis; Factor Analysis; Discriminate Analysis; Using SPSS

Unit Structure

12.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is regression?
- What is the method of linear regression analysis?
- What is factor analysis?
- About the discriminate analysis.
- About Using SPSS

12.1. Introduction

As we know statistic is very important tool in research and by the use of statistics, we interpret the data in various ways. Correlation and regression are commonly used when we are interested to interpret the relationship between two variables. Generally, variables are two types i.e., independent variable and dependent variable. Independent variables control over or influence the dependent variable. A linear measure of relationship between two variables is estimated by coefficient of correlation. Correlation explains about the strength of linear relationship and the direction of relationship as well. Fitting the mathematical function between two related variables using paired observations on them is dealt in regression analysis. We wish to determine a mathematical relationship between

the two variables so that we can predict the value of dependent variables based on the value the independent variable and explain the impact of changes in the value of dependent variable on the change of the value of independent variable. Factor analysis is by far the most often used multivariate technique of research studies. Discriminant analysis is also a very important tool of research. In this unit we will learn about the regression analysis, factor analysis, discriminate analysis and SPSS.

12.2. Regression

The meaning of word "regression" is "stepping back" towards average. Regression is a statistical measurement of the average relationship between one dependent variable and a series of other changing variables. Regression is used in many sciences including biological science, environmental science and other disciplines. Regression establishes the relationship between variables and thereby provide a mechanism for prediction or forecasting.

The two basic types of regression are linear regression and multiple linear regression. Linear regression uses one independent variable X to explain or predict the outcome of the dependent variable Y, while multiple regressions use two or more independent variables to predict the outcome. The general form of each type of regression expressed as follows:

- **Linear regression:** $Y = a + bX + u$
- **Multiple regression:** $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- X_1 , X_2, ... X_t are the independent variables.
- a = the intercept.
- b = the slope.
- u = the regression residual.

12.3. Linear Regression analysis

In linear regression, we calculate scores on one variable from the scores of a second variable. The variable we are predicting is called the dependent variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the predictions of Y when plotted on a plane as a function of X form a straight line. The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you wish to predict Y from X, the higher the value of X, the higher your prediction of Y.

Linear Regression Analysis

The relationship between two characteristic or variables of a population based on paired samples collected on those variables can eb estimated by regression analysis. A good measure of relationship between two variables are estimated by coefficient which tells us about the strength of relationship and the direction of relationship.

We fit a line y = a + bX. The exact relationship between X and Y is not linear, we are only approximating the relationship by a straight line. Therefore, it is not correct to write the line equation as Y = a + bX rather we write as

$$\bar{Y} = a + bX$$

where $\bar{Y}$ is the predicted or fitted or estimated value of Y. The exact relationship between X and Y can be written as

$$Y = a + bX + \text{error}$$

The error is the difference between the observed value and the predicted value of Y. Using collected observation $(x_1\ y_1)$, $(x_2\ y_2)$…….$(x_n\ ,y_n)$ these error or residuals can be written as $(y_i - a - bx_i)$ for i =1,2,….n and we wish to have such value of a and b for which the residuals should be minimum. By least squares method, we minimize the summation of squared residuals. For this, we differentiate $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ with respect to a and b separately and equate the derivation to zero. Solving those two equations, we get following estimates of a and b:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i-1}^{n}(x_i - \bar{x})^2}$$

The value of a and b obtained using least squares method are called as least squares estimates (LSA) of a and b. The correlation coefficient (r) between X and Y can be estimated as below:

$$r = \hat{b}\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \hat{b}\sqrt{\frac{SSX}{SSY}}$$

Thus, the signs of r and b are the same.

Example: Using mobiles also reduces the amount of physical exercise, causing weight gains. A sample of fifteen 10-year-old children was taken. The value of pounds each child was overweight was recorded (a negative number indicates the children is underweight). Additionally, the number of hours of mobile use per weeks was also recorded. The data are listed here. (Source: Book on Research methodology methods and techniques written by CR Kothari and Gaurav Garg)

Mobile use	42	34	25	35	37	38	31	33	19	29	38	28	29	36	18
overweight	18	6	0	-1	13	14	7	7	-9	8	8	5	3	14	-7

Fit the regression line and describe what the coefficients tell you about the relationship between the two variables.

Solution: We make the following table:

Mobile use (x_i)	**Overweight (y_i)**	**$(x_i - \bar{x})$**	**$(y_i - \bar{y})$**	**$(x_i - \bar{x})^2$**	**$(x_i - \bar{x})(y_i - \bar{y})$**
42	18	10.5333	12.2667	110.9511	129.2089
34	6	2.5333	0.2667	6.4178	0.6756
25	0	-6.7333	0.2667	41.8178	37.0756
35	-1	3.5333	-6.7333	12.4844	-23.7911
37	13	5.5333	7.2667	30.6178	40.2089
38	14	6.5333	8.2667	42.6844	54.0089
31	7	-0.4667	1.2667	0.2178	-0.5911
33	7	1.5333	1.2667	2.3511	1.9422
19	-9	-12.4667	-14.7333	155.4178	183.6756
29	8	-2.4667	2.2667	6.0884	-5.5911
38	8	6.5333	2.2667	42.6844	14.8089

28	5	-3.4667	-0.7333	12.0178	2.5422
29	3	-2.4667	-2.7333	6.0844	6.7422
36	14	4.5333	8.2667	20.5511	37.4756
18	-7	-13.4667	-12.7333	181.3511	171.4756
X=31.4667	Y=5.7333	← Mean	Total →	SSX=671.7333	SSXY=649.8667

Using the above calculation, we obtain

$$\hat{b} = \frac{SSXY}{SSX} = \frac{649.8667}{671.7333} = 0.9674$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 5.733 - 0.9674 \text{ X } 31.4667 = -24.709.$$

Therefore, the fitted simple linear regression model is

$$\hat{y} = -24.709 + 0.9674x.$$

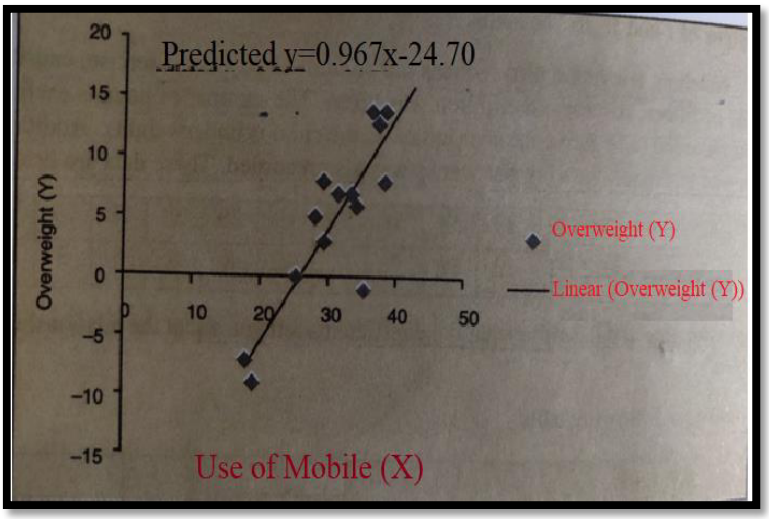


Fig-1: Showing scattered plot of data of the given example

Fig-1 shows scatter plot of the data and the fitted line as well. The value b = 0.9674 is the change in the value of y for a unit change in the value of x. The intercept is the constant or the value of y when x is zero. The fitted line can be used to predict value of y for a new (not in the collected data, however in the range) value of x. For instance, the predicted value of y when x = 30, is -24.709 + 0.967 x 30 = 4.301. Thus, weight gain for 30 hours of TV watching per week is 4.301 pounds. It is advisable to predict y values for those x values which are in the range of the value of X in the collected sample as the behavior of the data may be different beyond the range.

Coefficient of determination: When the actual relationship between observed X and Y values is almost linear, the fitted linear model will be reasonably good approximation of the actual relationship. In case, the actual relationship is not linear, the model may be misleading. As discussed, the quantities $y_i - a - bx_i$ (i=1, 2, 3…., n) are called as residuals. For a good model the magnitude of all the residual should be as small as possible. Therefore, these residuals can be utilized to tell us something about the goodness of the fitted model.

Studying Y using X means explaining the variability of Y using X. Once we get the idea of how the Y values are changing, we can predict them. The variance of Y values is $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$. We partitioned this variance into two parts as below:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i} + \widehat{y_i})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2$$

or

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \sum_{i=1}^{n}(\widehat{y} - \bar{y})^2$$

It can be shown that the product term $\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y})(\hat{y}_i - \bar{y}) = 0$. Since, $(y_i - \hat{y}_i)$ is the residual, the quantity $\sum_{i=1}^{n}(yi - \hat{y})^2$ is called as sum of squares due to error (SSE) the quantity $\sum_{i=1}^{n}(\hat{y} - \bar{y})^2$ is called as sum of square due to regression (SSR) as it is the part of the variability of Y which is explained using regression model. Thus, total variability in Y ($SST=\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is portioned into two parts viz. Explained variability (SSR) and Unexplained variability (SSE).

Clearly, SST= SSR+SSE

The fraction of SST explained by regression is given by R^2:

$$R^2 = SSR/SST = 1 - (SSE/SST)$$

It is clear that $0 \leq R^2 \leq 1$. This means that regression explains most of the variability in Y and the fitted model is good. When SSE is closed to SST, R^2 will be closed to 0. This means that regression does not explain much variability in Y and the fitted model is not good. The quantity R^2 is called as coefficient of determination and is used to evaluate the goodness of the fitted model. For simple linear regression model, coefficient of determination R^2 is the same as square of correlation coefficient between X and Y.

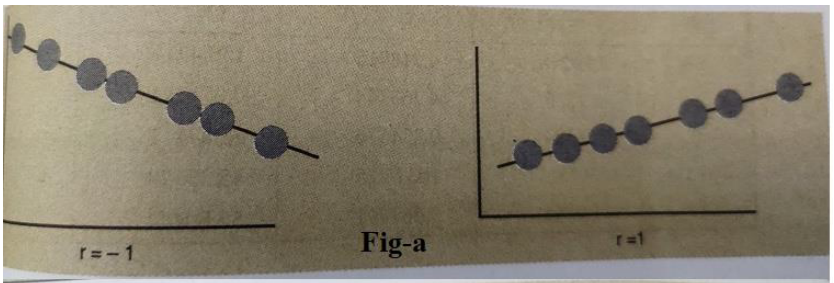


Fig-a

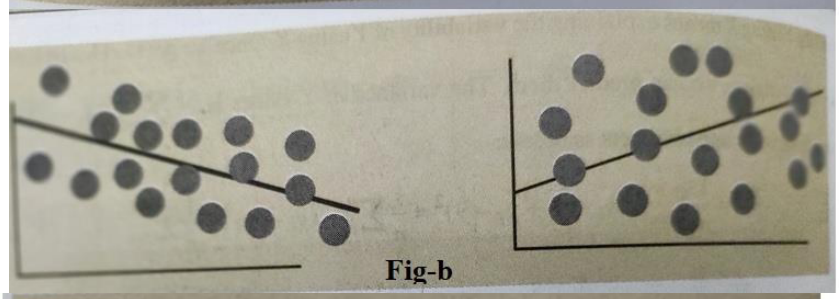


Fig-b

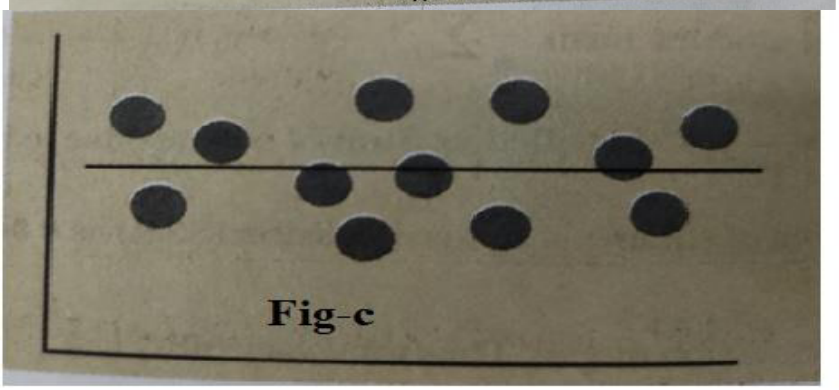


Fig-c

In fig-a, $R^2 = 1$ which means perfect linear relationship between X and Y. In these models, 100% of the variation in Y is explained by X. In fig-b $0<R^2<1$, which indicates relatively weak linear relationship. In this case, some but all of the variations in Y is explained by X. In fig-c $R^2= 0$, which means no linear relationship. In this model, none of the variation in Y is explained by X.

Example: Estimate R^2 for the model fitted in following example and comment on its value.

Solution: Using the data and the fitted model in example we make the following table.

(y_i)	$\hat{y}_i=\hat{a}+\hat{b}x_i$	$(y_i-\hat{y}_i)^2$	$(y_i-\bar{y})^2$
18	15.92	4.31	150.47
6	8.18	4.76	0.07
0	-0.52	0.27	32.87
-1	9.15	103.02	45.33

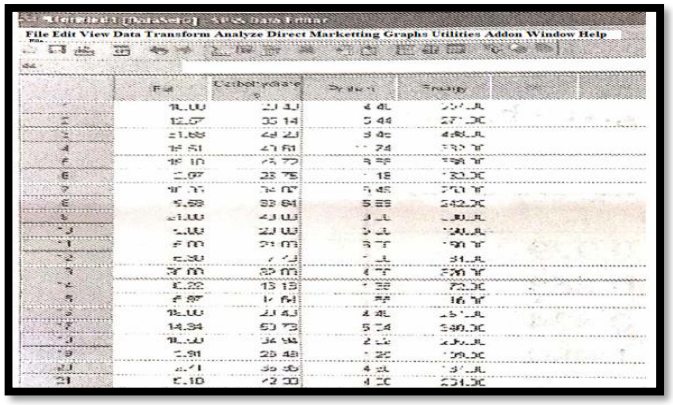
13	11.08	3.66	52.80
14	12.05	3.79	68.33
7	5.2	2.95	1.60
7	7.2	0.04	1.60
-9	-6.32	7.13	217.07
8	3.34	21.66	5.13
8	12.05	16.4	5.13
5	2.37	6.87	0.53
3	3.3	0.11	7.47
14	10.11	15.07	68.33
-7	-7.29	0.08	162.13
		SEE= 190.22	SST=818.93

Thus, R^2= SSR/SST = 1-SSE/SST= 1-190.22/818.93= 0.767. This calculated value is close to 1. Thus, the fitted model is considered to be good one.

Using SPSS in linear regression: We illustrate the use of SPSS in regression analysis through the following example.

Example: In order to fit a model to predict the calories in a cake, 59 cakes were sampled and fat, carbohydrates, proteins and energy are measures. Fit the regression model and examine the validity of the model.

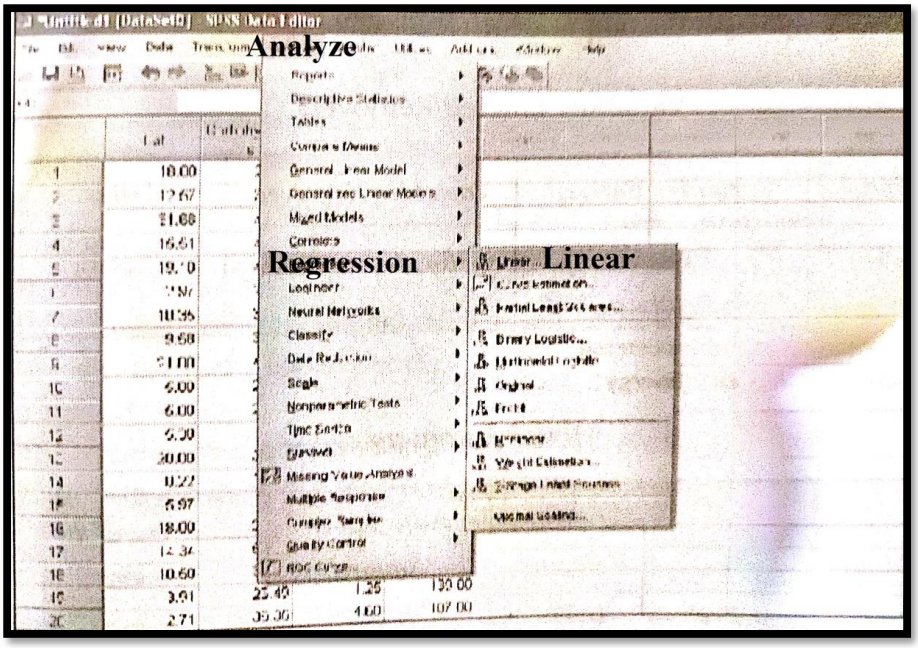
Solution: Enter the data in SPSS worksheet. You can copy the data from MS excel also. You can change the names of variables by clicking “variable view” in the left down corner of the screen. Below is the screen shot of the worksheet.



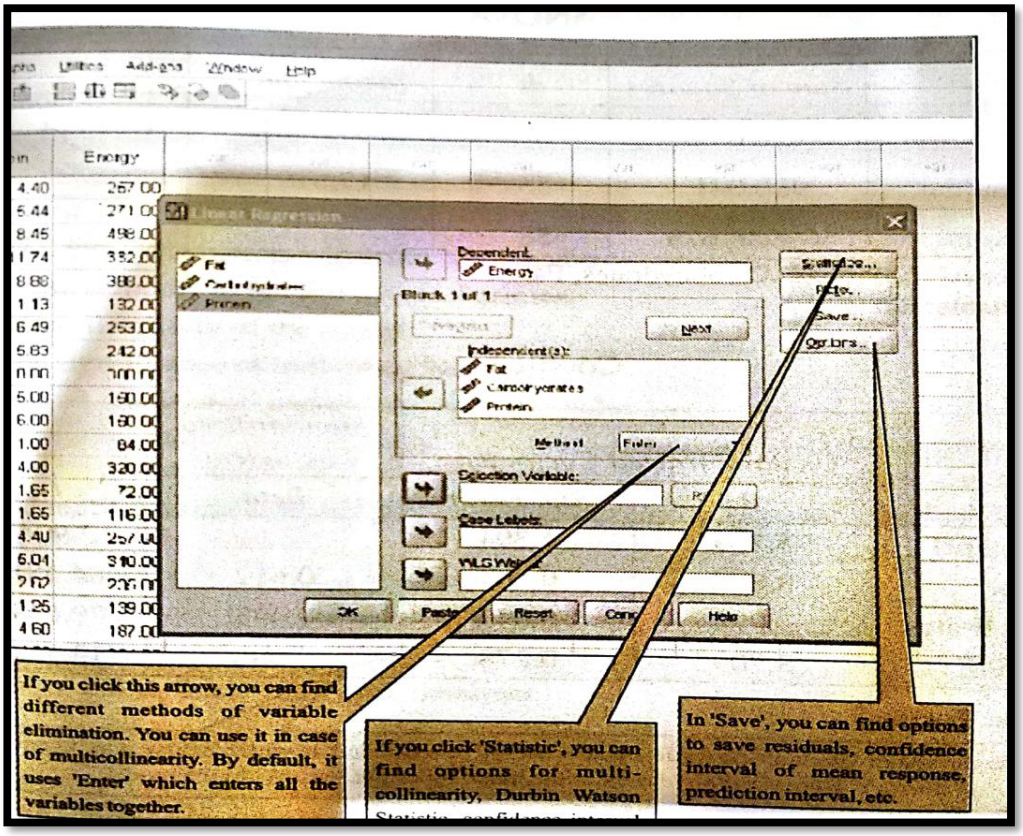
Click “Analyze”

Select “Regression” from the drop-down menu

Select “Linear” from the drop-down menu



You will get a popup window. In that window select “Energy” as the “Dependent” and remaining three as “Independent”.



When you click "OK" you get the output which is given below:

Variables Entered/Removed[b]

Model	Variables entered	Variables removed	Mothod
1	Proteins, Carbohydrates, Fats [a]		Enter

[a] All requested variables entered

[b] Dependent variable: Energy

Model Summary

Model	R	R square	Adjusted R square	Standard error of the Estimates
1	0.999[a]	0.998	0.998	4.22

[a] Predictors (Constant), protein, Carbohydrates, Fat

ANOVA[b]

Model		Sum of squares	df	Mean Square	F	Sig.
1	Regression	602976.637	3	200992.212	1.128E4	0.000[a]
	Residual	980.041	55	17.81		

	Total	603956.678	58			

[a] Predictors (Constant), protein, Carbohydrates, Fat

[b] Dependent variable: Energy

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	t	Sig
		B	Std. Error	Beta		
1	(Constant)	5.239	1.804		2.904	0.005
	Fat	8.802	0.104	0.612	84.581	0.000
	Carbohydrates	3.779	0.047	0.493	79.648	0.000
	Protein	3.709	0.278	0.088	13.320	0.000

[b] Dependent variable: Energy

If you compare the magnitude of standardized coefficient BETA you can decide which explanatory variable is more important in explaining Y. Here “Carbohydrates” is the most important variable in the model.

12.4. Factor Analysis

Factor analysis is the most often used multivariate technique of research studies, specially once experimenter is dealing many correlated variables. It is a technique applicable when a systematic interdependence among a set of observed or manifest variables and the researcher are interested in finding out something more fundamental or latent which creates this commonality.

Important method of Factor Analysis

There are several methods of factor analysis, but they do not necessarily give some result. As such factor analysis is not a single unique method but a set of techniques. Important method of factor analysis is:

- The centroid method
- the principal components method
- the maximum likelihood method

Before we describe these different methods of factor analysis, it seems appropriate that some basic terms relating to factor analysis be well understood.

I. **Factor:** A factor is an underlying dimension that account for several

a) Observed variables.

b) There can be one or more factors, depending upon the nature of the study and the number of variables in it.

II. **Factor –loadings:** Factor loading are those values which explain how closely the variables are related to each one of the latent factors. They are also known as factor –variable correlation.

III. **Communality (h_2):** Communality symbolized as h2 shows how much of each variable is accounted for the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factor is taken into consideration. It is worked out in respect of each variable as under:

h_2 of the i^{th} variable = (i^{th} factor loading of factor A)^2 + (i^{th} factor loading of factor B)^2 +……

IV. **Eigen value (or latent root):** When we take the sum of squared values of factor loading relating to a factor, the sum is referred to as Eigen value or latent root.

V. **Total sum of squares:** Sum of Eigen values of all factors are the total sum of squares. This value, when divided by the numbers of variables, result in the index that show how two particular solution accounts for what all the variables taken together represent.

VI. **Rotation:** Rotation is the context of factor analysis, and reveal different structures in the data.

VII. **Factor scores:** Factor score represent the degree to which each case gets scores on the groups of items that load high on each factor.

Centroid method: This method of factor analysis, developed by L.L Thurstone, was quite frequently used until about 1950 before the advent of large capacity high speed computers. Various steps involved in this method are as follows:

(a) This method starts with the computation of a matrix correlation, R, where in unties are place in the diagonal spaces.

(b) if the correlation matrix so obtained happens to be positive manifold the centroid method required that the weights for all variables be +1.0. in other words, the variables are not fitted they are simply summed.

(c) The first centroid factor is determined as under.

I. the sum of the coefficients (including the diagonal unity) in each column of the correlation matrix is worked out.

II. Then the sum of these column sums(T) is obtained.

III. the sum of each column obtained as per(a) above us divided by the square root of T obtained in (b) above, resulting in what are called centroid loadings. this way each centroid loading (one loading for one variable) is computed.

IV. to obtain second centroid factor (say B), one must first obtain a matrix of residual coefficients. for this purpose, the loading for the two variables on the first centroid factor are multiplied. this is done for all possible pairs of variables. the resulting matrix of factor cross products may name as Q_1. then Q_1 is subtracted clement by element from the original matrix of correlation, R and the result is the first matrix of residual coefficients, R_1.

V. One should understand the nature of the elements in R_1 matrix. Each diagonal element is a partial variance i.e., the variance that remains after the influence of the first factor is partial.

VI. For subsequent factors (C, D, etc.) the same process outlined above is repeated. After the second centroid factor is obtained, cross products are computed forming, matrix Q_2. This is then subtracted from R_1 (and not from R'_1) resulting in R_2. To obtain a third factor (C), one should operate on R_2 in the same way as on R_1. first, some of the variables would have to be reflected to maximize the sum of loadings, which would produce R'_2. Loadings would be computed from R'_2 as they were from R'_1. again, it would be necessary to give negative signs to the loadings of variables which were reflected which would result in third centroid factor (c).

Example: - Given is the following correlation matrix, R, relating to eight variables with unities in the diagonal spaces:

	1	2	3	4	5	6	7	8
1	1.000	.709	.204	.081	.626	.113	.155	.774
2	.709	1.000	.051	.089	.581	.098	.083	.652
3	.204	.051	1.000	.671	.123	.689	.582	.072
4	.081	.089	.671	1.000	.022	.798	.613	.111
5	.626	.581	.123	.022	1.000	.047	.201	.724
6	.113	.098	.689	.798	.047	1.000	.801	.120
7	.155	.083	.582	.613	.201	.801	1.000	.152
8	.774	.652	.072	.111	.724	.120	.152	1.000

Using the centroid method of factor analysis, work out the first and second centroid factors from the above information.

Solution: - given correlation matrix, R is a positive manifold and as such the weights for all variables be + 1.0 accordingly, we calculate the first centroid factor (A) as under: (Source: Book on Research methodology methods and techniques written by CR Kothari and Gaurav Garg)

	1	2	3	4	5	6	7	8
1	1.000	.709	.204	.081	.626	.113	.155	.774
2	.709	1.000	.051	.089	.581	.098	.083	.652
3	.204	.051	1.000	.671	.123	.689	.582	.072
4	.081	.089	.671	1.000	.022	.798	.613	.111
5	.626	.581	.123	.022	1.000	.047	.201	.724
6	.113	.098	.689	.798	.047	1.000	.801	.120
7	.155	.083	.582	.613	.201	.801	1.000	.152
8	.774	.652	.072	.111	.724	.120	.152	1.000
Column sum	3.662	3.263	3.392	3.385	3.324	3.666	3.587	3.605
Sum of the column sums (T) = 27.884 $\therefore \sqrt{T} = 5.281$								

3.662	3.263	3.392	3.385	3.324	3.666	3.587	3.605
5.281	5.281	5.281	5.281	5.281	5.281	5.281	5.281

First centroid factor A=
.693, .618, .642, .641, .629, .694, .679, .683

Principal components method

Principal components (PC) analysis is a procedure to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component accounts for the largest variability in the data, and each succeeding component in turn has the highest variance possible under the constraint that it is uncorrelated with the preceding components.

The aim of the principal components method is the construction out of a given set of variables X_j's (i= 1, 2, ……, k) of new variables (p_i), called principal components which are liner combinations of the X_s,

$$P_1 = a_{11} X_1 + a_{12} X_2 + \ldots + a_{1k} X_K$$

$$P_2 = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2k} X_k$$

The method is being applied mostly by using standardized variables, i.e.,

$$z_j = \frac{(X_j - \overline{X}_j)^2}{\sigma_j}$$

The a_{ij}'s are called loadings and are worked out in such a way that the extracted principal components satisfy two conditions: (i) principal components are uncorrelated (orthogonal) and (ii) the first principal component (p_1) has the maximum variance, the second principal component (p_2) has the next maximum variance and so on.

Following steps are usually involved in principal components method:

I. Estimates of a_{ij}' s is obtained with which X's is transformed into orthogonal variables i.e., the principal components. A decision is also taken with regard to the question: how many of the components to retain into the analysis?

II. We then proceed with the regression of Y on these principal components i.e.,

III. $Y = \hat{y}_1 p_1 + \hat{y}_2 p_2 + \ldots + \hat{y}_m p_m$; (m < k)

IV. From the $\hat{a}_{1j}$ and $\hat{y}_{1j}$, we may find b_{lj} of the original model, transferring back from the p's into the standardized X's.

Using SPSS in Factor Analysis

We illustrate the use of SPSS in solving factor analysis problem through following example:

Example: A market researcher wants to determine the underlying benefits consumers seek from the purchase of a car. A sample of 15 respondents was interviewed. The respondents were asked to indicate their degree of agreement with the following statement using a 10-point scale (0= strongly disagree, 9= strongly agree).

X1	I like a car that has stylish interior
X2	I prefer a car that looks great
X3	I prefer a car giving high mileage
X4	I prefer a car with low maintenance
X5	I prefer a car that provides a good value for money

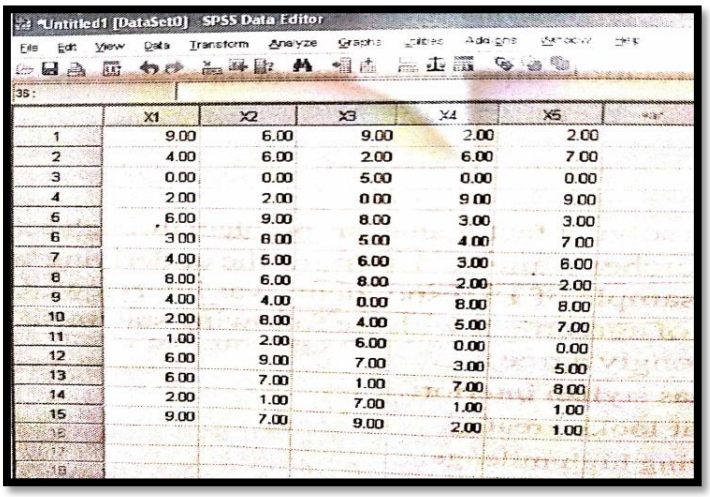
Collected data is given below:

Respondents Number	X1	X2	X3	X4	X5
1	9	6	9	2	2
2	4	6	2	6	7
3	0	0	5	0	0
4	2	2	0	9	9
5	6	9	8	3	3
6	3	8	5	4	7
7	4	5	6	3	6
8	8	6	8	2	2
9	4	4	0	8	8
10	2	8	4	5	7
11	1	2	6	0	0
12	6	9	7	3	5
13	6	7	1	7	8
14	2	1	7	1	1
15	9	7	9	2	1

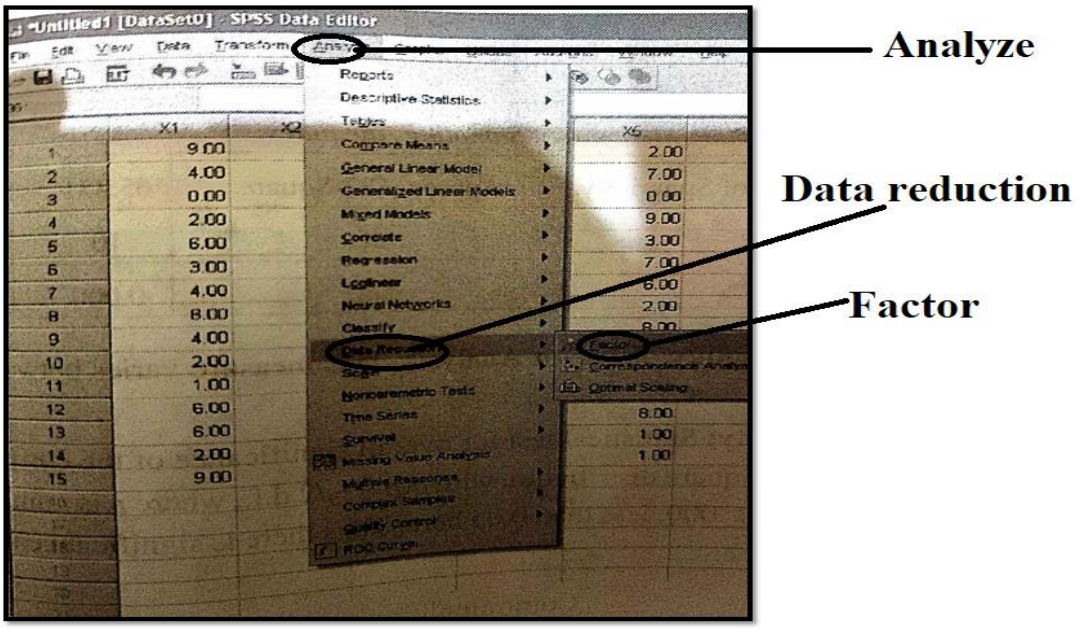
Perform a factor analysis and identify the underlying factors.

Solution: We use SPSS to perform the factor analysis

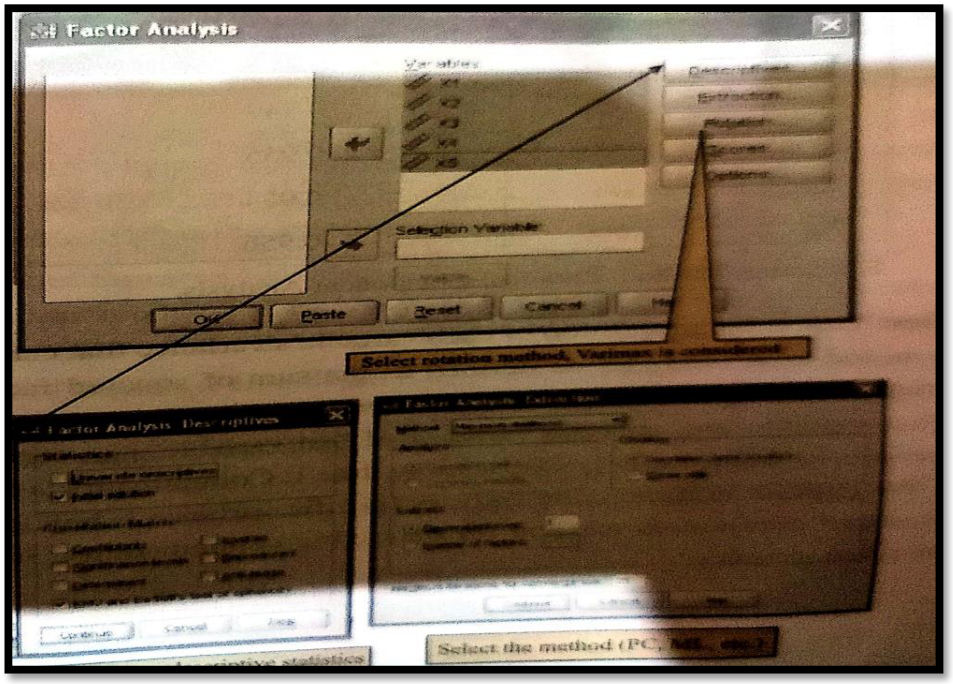
We first enter the data in the SPSS worksheet



Click “Analyze” select “Data Reduction”, Select “Factor”



Select all the variables as shown below:



Outcome is presented and explained below:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling adequacy		0.516
Bartlett's Test of Sphericity	Approx. Chi Square	65.331
	Df	10
	Sig.	0.000

KMO test: Test the suitability of factor analysis. This measure varies between 0 and 1 and values closer to 1 are better.

Bartlett's Test of Sphericity: Statistical test for overall significance of the correlations within a correlation matrix. Uses chi square distribution with p (p-1)/2 df. where p is the number of variables. Sig. gives the p-value which is 0.000, less than 0.05 here. Thus, there is significant correlation among variables.

Communalities

	Initial	Extraction
X1	1.000	0.815
X2	1.000	0.849
X3	1.000	0.956
X4	1.000	0.954
X5	1.000	0.956

Extraction method: Principal component analysis

Communalities: This is the proportion of each variable's variance that can be explained by the factors. It is also denoted as h^2 and can be defined as the sum of squared factor loadings for the variables. The variance of each variable is standardized to unity and partitioned into two parts- communality of that variable and specific variance of that variable i.e., communality + specific variance = 1. Therefore, no communality can be more than 1. Communality is due to the correlation among variables. We explain communality only as specific variance is beyond the control of our hand. We used principal component method for extracting communalities. In this method initial communalities are taken as unity.

Total variance explained (Source: Book on Research methodology methods and techniques written by CR Kothari and Gaurav Garg)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of squared Loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.755	55.09	55.09	2.75	55.09	55.09	2.73	54.71	54.71
2	1.775	35.49	90.59	1.77	35.49	90.59	1.79	35.87	90.59
3	0.377	7.54	98.13						
4	0.065	1.29	99.43						
5	0.028	0.56	100.00						

Extraction method: Principal component analysis

Total Variance explained: We used principal component (PC) method of factor analysis. In this method factor is called as component. The initial number of factors is the same as the number of variables used in the factor analysis. However, not all 5 factors will be

retained. You can choose number of factor or Eigen value method. In the later method, the number of factors = number of eigenvalues of correlation matrix with more than 1. Initial Eigen values are Eigen values of correlation matrix. We can see here, only two Eigen values are more than 1 and also two factors explain 90.59% variance, while 3 factors explain 98.131% variance.

Extraction sums of squared loading: The numbers of rows in this panel of the table correspond to the number of factors retained. The values in this panel of the table are calculated in the same ways as the value in the left panel. In some other method, these values may be smaller.

Rotation sums of squared loadings: The matrix of the factors loadings is rotated orthogonally using varimax rotation. Total amount of variance accounted for is redistributed over the two extracted factors. This helps making the factors distinct. In above example, total 90.59 % variance is redistributed over the 2 factors.

Screen Plot: This is the plot between eigen value and the factor number. From the third factor on, you can see that the line is almost flat, meaning each successive factor is accounting for smaller and smaller amounts of the total variance. This plot is called a "Scree" Plot because it is often looking like a scree slope where rocks have fallen down and accumulated on the side of mountain.

Component matrix[a]

	Components	
	1	2
X1	-0.295	0.853
X2	0.048	0.920
X3	-0.938	0.278
X4	0.950	0.228
X5	0.940	0.268

Extraction method: Principal component analysis

[a] 2 Components extracted

Component matrix (Factor matrix): This table contains the unrotated factor loadings, which are the correlations between the variable and the factor.

Reproduced correlations

		X1	X2	X3	X4	X5
Reproduced correlations	X1	0.815	0.771	0.514	-0.086	-0.049
	X2	0.771	0.849[a]	0.211	0.255	0.291
	X3	0.514	0.211	0.956[a]	-0.827	-0.807
	X4	-0.86	0.255	-0.827	0.954[a]	0.954
	X5	-0.049	0.291	-0.807	0.954	0.956[a]
Residuals[b]	X1		-0.161	-0.045	0.068	-0.047
	X2	-0.161		0.020	-0.065	0.028
	X3	-0.045	0.020		-0.005	0.033
	X4	0.068	0.065	-0.005		-0.027
	X5	-0.047	0.028	0.033	-0.027	

Extraction Method: Principal component analysis

[a] Reproduced communalities

[b] Residuals are computed between observed and reproduced correlations. There are three (30%) non-redundant residuals with absolute values greater than 0.05.

Reproduced correlations: The reproduced correlation matrix is the correlation matrix based on the extracted factors. We want the values in the reproduced matrix to be as close to the values in the original correlation matrix as possible. This means that residual matrix, which contains the difference between the original and the reproduced matrix to be close to zero. Diagonal elements of the reproduced correlation matrix are extracted communalities.

Rotated component Matrix

	Component	
	1	2
1	-0.173	0.886
2	0.175	0.904
3	-0.890	0.405
4	0.972	0.093
5	0.968	0.134

Extraction Method: Principal component analysis

Rotation method: Varimax with Kaiser Normalization

[a] Rotation converged in 3 interactions

12.5. Discriminant Analysis

Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. The term categorical variable means that the dependent variable is divided into a number of categories. For example, three brands of computers, Computer A, Computer B and Computer C can be the categorical dependent variable. The objective of discriminant analysis is to develop discriminant functions that are linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables. It also evaluates the accuracy of the classification. Discriminant analysis is described by the number of categories that is possessed by the dependent variable.

The dependent variable with two categories, is analyzed with two-group discriminant analysis. If the dependent variable has three or more than three categories, then the type used is multiple discriminant analysis. The major distinction to the types of discriminant analysis is that for a two group, it is possible to derive only one discriminant function. On the other hand, in the case of multiple discriminant analysis, more than one discriminant function can be computed.

There are many examples that can explain when discriminant analysis fits. It can be used to know whether heavy, medium and light users of soft drinks are different in terms of their consumption of frozen foods. In the field of psychology, it can be used to differentiate between the price sensitive and non-price sensitive buyers of groceries in terms of their psychological attributes or characteristics. In the field of business, it can be used to understand the characteristics or the attributes of a customer possessing store loyalty and a customer who does not have store loyalty.

For a researcher, it is important to understand the relationship of discriminant analysis with Regression and Analysis of Variance (ANOVA) which has many similarities and differences. There are some similarities and differences with discriminant analysis along

with two other procedures. The similarity is that the number of dependent variables is one in discriminant analysis and in the other two procedures, and the number of independent variables is multiple in discriminant analysis. The difference is categorical or binary in discriminant analysis, but metric in the other two procedures. The nature of the independent variables is categorical in Analysis of Variance (ANOVA), but metric in regression and discriminant analysis.

- The steps involved in conducting discriminant analysis are as follows: The problem is formulated before conducting.
- The discriminant function coefficients are estimated.
- The next step is the determination of the significance of these discriminant functions.
- One must interpret the results obtained.
- The last and the most important step is to assess the validity.

Discriminant Analysis: Discriminant analysis is technique to discriminate between two or more mutually exclusive groups on the basis of some explanatory variables. These groups are known a –priori when the criterion variable has two categories, the technique is known as two-group discriminant analysis. When three or more categories are involved, the technique is referred to as multiple discriminant analysis. Discriminant analysis helps the researchers who are interested to understand how consumers differ with respect to demographic and psychographic characteristics. Discriminant analysis is also used to predict the group membership. Banks use discriminant analysis to discriminate between the customer who default and who repay the loan in time, based on their age, income, assets, number of dependents, and previous outstanding loan, etc.

Consider the following example:

Example-1: Suppose a company wants to determine if one of its new products and improved digital camera will be commercially successful or not. The company wants to distinguish the purchaser from non-purchaser. The company make a panel of 10 potential purchaser and devised the rating scales on three characteristics viz. durability, performance and style. The panel is asked to give the digital camera ratings from 1 to 10

on these three scales. After the products was evaluated, they are asked their buying intention (would purchase or would not purchase). Five said that they would purchase the digital camera and the other five said that they would not purchase it. The ratings given by panelists are given below:

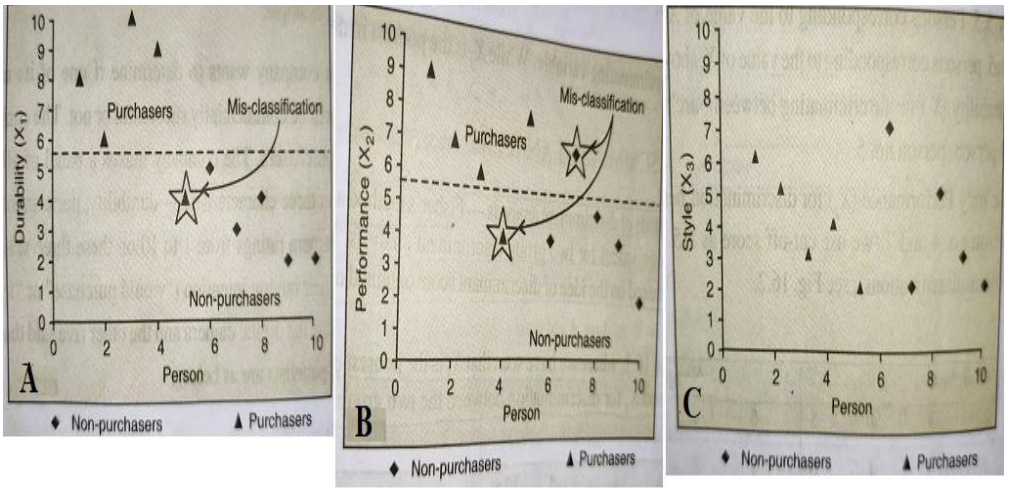
Group based on purchased intention	Durability (X1)	Performance (X2)	Style (X3)
Group A: Would purchase			
Person 1	8	9	6
Person 2	6	7	5
Person 3	10	6	3
Person 4	9	4	4
Person 5	4	8	2
Group Mean	7.4	6.8	4
Group A: Would not purchase			
Person 6	5	4	7
Person 7	3	7	2
Person 8	4	5	5
Person 9	2	4	3
Person 10	2	2	2
Group Mean	3.2	4.4	3.8
Difference between group means	4.2	2.4	0.2

Let us write the observation on the variable X1, X2 and X3 as below:

	Purchaser					Non-Purchaser				
Person	1	2	3	4	5	6	7	8	9	10
X1	8	6	10	9	4	5	3	4	2	2

	Purchaser					Non-Purchaser				
Person	1	2	3	4	5	6	7	8	9	10
X2	9	7	6	4	6	4	7	5	4	2

	Purchaser					Non-Purchaser				
Person	1	2	3	4	5	6	7	8	9	10
X3	6	5	3	4	2	7	2	5	3	2



Graph (A, B and C) showing Durability, performance and style

(Source: Book on Research methodology methods and techniques written by CR Kothari and Gaurav Garg)

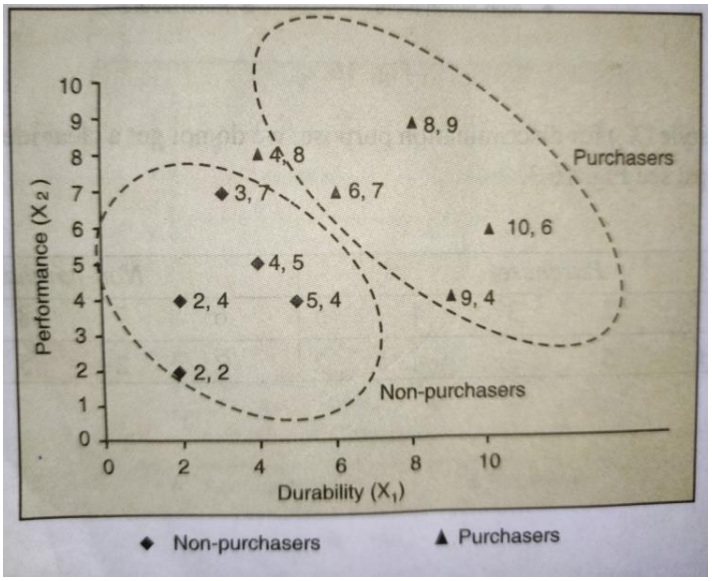
Fix a cut off score, say 5.5. Persons corresponding to the value of X_1 below 5.5 belong to non-purchaser. On the other hand, persons corresponding to the value of X_1 above 5.5 belong to purchaser. Thus while, using only durability (X_1) for discriminating between purchaser and non-purchaser, we have one misclassified person, person number 5. Similarly, when we use only performance (X_2) for discrimination between two groups, we get two misclassification, person number 4 and person number 7 (we use cut of score as 5.5, there is no better cut off score giving a smaller number of misclassifications). When we use style (X_3) for discrimination purpose we do not get a clear idea of discriminating between two groups.

Thus X_1 (Durability) is the best discriminating variable while X_3 (Style) is the poorest in discriminating.

Two Group Discriminant Analyses

There are two popular methods of discriminant analysis – fisher's method and Mahalanobis' method. Both of the methods are equivalent for two group discriminant analysis. We here present only the fisher' method which is based on the idea of discriminant score obtained using a linear combination of explanatory variables.

Let us take example-1, where we have observed that X_3 is the poorest in discriminating variable. We have X_3 and use X_1 and X_2 for discrimination between the two groups.



The figure suggests us to use a linear combination of X_1 and X_2 for discrimination

(Source: Book on Research methodology methods and techniques written by CR Kothari and Gaurav Garg)

For example, if we take (X_1+X_2) for discrimination with a cut off score 11, i.e., the individuals, having (X_1+X_2) <11, belong to the group of non-purchasers and the others belong to the group of purchasers. Tilting the axes graphically to separate the two groups may also help. How to obtain a linear combination of explanatory variables which discriminates best between the groups is the important question which we need to answer now.

Methodology: We wish to have a discriminant function, a linear combination of discrimination or explanatory variables, which maximize the difference between the two groups. We fit a discriminant function of the from

$$Z = b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Where Z is discriminant score, X_1+ X_2 +.... + X_k are k explanatory or predictor variable and b_j = discriminant weight corresponding to variable X_j (j= 1, 2…, k).

We collect n observations on each variable. So, we can write

$$Z_i = b_1 x_{1i} + b_2 x_{2i} + b_3 x_3 + \ldots + b_k x_{ki}$$

where,

Z_i= Discriminate score (i=1, 2…n) and

x_{ij}= i^{th} observation or value belonging predictor or explanatory variables x_1.

in a particular group, each individual has a discriminant score (Z). Discriminant weight b_j are obtained such that the discriminant function discriminates maximum between two groups.

Let b the column vector of discriminant weights b_j and let d= ($x_{(2)}$ × $x_{(1)}$) be the column vector of the difference between the two groups means (means vectors of groups 1 and 2 are denoted by $x_{(1)}$ and $x_{(2)}$ respectively . Also let c be the pooled within –groups covariance matrix square & cross products terms of mean corrected value of predictor variable belonging to group i (i=1,2). We obtained weights b_j so as to maximize the ratio of between groups variability to within group variability We obtain **b** such that

$$\frac{\mathrm{b'dd'b}}{\mathrm{b'Cd}}$$

is maximized. The numerator $\mathrm{b'dd'b}$ capture the between group variability, while the denominator $\mathrm{b'Cd}$ capture within group variability.

Centroid of a group is the mean of all objects within that group. A two-group discriminant analysis has two centroids. It indicates most typical location of an individual from a particular group:

Centroids of group 1 and 2 are given by $\bar{Z}_{(1)} = \bar{x}_{(1)}$ 'b and $\bar{Z}_{(2)} = \bar{x}_{(2)}$ 'b, respectively. The greater the difference between the two centroids, larger the objective function $\mathrm{b'dd'b}$ /$\mathrm{b'Cd}$ value. It can be shown that such obtained b is proportional to C^{-1}d. Because the scale of discriminant weight is not known, we must standardize the discriminant weight such that the length of vector b is 1.

for this, we divide each bj by, $\sqrt{\sum_{j=1}^{k} \mathrm{b_j}}$

After obtaining discriminant weights, we can obtain the discriminant scores for n individuals in the sample. suppose, n_1= Number of persons or individuals in first group and n_2 = number of persons and individuals in second group. Clearly, $n_1+n_2 = n$. The cut off score is obtained as $(n_1\bar{Z}(1) + n_2\bar{Z}(2)$ /$n_1+n_2)$. When the discriminant score of an individual is below the cut off score, the individual is assigned to group 1, otherwise to group 2.

Assumptions:

Results of discriminant analysis are valid only if:

1) Cases or individuals should be independent.
2) Predictor variable should have a multivariate normal distribution.
3) Within – group variance-covariance matrices should be equal across groups.
4) Group membership is assumed to be mutually exclusive, i.e., no case belongs to more than one groups.
5) Group membership should be collectively exhaustive, i.e., all cases are members of a groups.

Using SPSS in Discriminant Analysis

We use example 1 of discriminant analysis to illustrate the usage of SPSS. We first enter the data on three variables along with another variable for group membership. This new variable, denoted as Y, takes the value 1 for group 1, and the value 2 for group 2.

We click on "Analyze" then select "Classify" from the menu and then select "Discriminant".

In the pop-up window, select Y as the "Grouping variable". Define the range: Minimum =1, Maximum=2, for 2 groups, Select X_1 and X_2 as "Independents".

The output is presented and explained as below:

Eigenvalues

Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	3.315[a]	100.0	100.0	0.877

[a] First 1 canonical discriminant function were used in the analysis.

Eigenvalue: These eigenvalues describe how much discriminating ability a function possesses. The magnitudes of the eigenvalues are indicative of the function's discriminating abilities. In case of two group discriminant analysis, you have only one eigenvalue.

% of variance: This is the proportion of discriminating ability of predicting variables found in a given function. This proportional is calculated as the proportion of function's

eigenvalue to the sum of all the eigenvalues. In case of two group discriminant analysis, you have this value as 100%.

Cumulative %: This is the cumulative proportion of discriminating ability. In case of two group discriminant analysis, this value is also 100%.

Canonical correlation: These are canonical correlations of our predictor variables and the grouping variable.

Wilks' Lambda

Test of function (s)	Wilks' Lambda	Chi Square	d.f	Sig.
1	0.232	9.504	3	0.023

Test of function (s): These are functions included in the given test with the null hypothesis that canonical correlations associated with the functions are all equal to zero. In two group discriminant analysis, we have only one function.

Wilks' Lambda: In two group discriminant analysis, it is given by (1-canonical correlation^2). In multi-group discriminant analysis, it is given by product of all such values. Large values of Wilks' Lambda indicate that group means are not different, while the small values indicate that group means are different.

Chi Square: This is the Chi Square statistic calculated using Wilks' Lambda. The null hypothesis is that the function has no discriminating ability.

d.f.: This is the degree of freedom for Chi square test.

Sig.: This is p value associated with the chi square test. Here the p value is 0.023 which is smaller than 0.05. Therefore, we reject the null hypothesis at 5% level of significance and conclude that the function has significant discriminating ability.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X_1	1.110
X_2	0.709
X_3	-0.564

Standardized Canonical Discriminant Function Coefficients: There are standardized Discriminant function coefficients or discriminant weights, b_1', b_2' and b_3'.

Thus, the Discriminant function can be written as

$Zi=1.110X_{1i} + 0.709X_{2i}-0.564X_{3i}$

Structure Matrix

	Function
	1
X_1	0.666
X_2	0.394
X_3	0.032

Polled within groups correlations between discriminating variables and standardized canonical discriminant function. Variables order by absolute size of correlations within function.

Structure Matrix: This gives the correlation of predictor variables with the discriminant function. Clearly, X_1 is the best discriminating variable having highest correlation with discriminant function.

Canonical Discriminant Function Coefficients

	Function
	1
X_1	0.573
X_2	0.379
X_3	-0.297
(Constant)	-4.002

Unstandardized coefficients

Canonical Discriminant Function Coefficients: As earlier mentioned that, you get here Unstandardized discriminant weights with a constant term. SPSS obtains discriminant scores using unstandardized coefficients as

$Zi= 0.573X_{1i} + 0.379x_{2}i-0.297x_{3i}-4.002$

Adding a constant term (-4.002) makes the cut off score as zero.

Functions at Group Centroids

Y	Function
	1
1	1.629
2	-1.629

Unstandardized canonical discriminant functions evaluated at group means.

12.6. SUMMARY

In this unit we have discussed various aspects of regression analysis. So far you have learnt that:

- The meaning of word “regression” is “stepping back”. Regression is a statistical measurement of the relationship between one dependent variable and a series of other independent variables. Regression is used in many sciences including biological science, environmental science and other disciplines.
- Regression is also attempted to establish the nature of the relationship between variables and thereby provide a mechanism for prediction or forecasting.
- In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the predictions of Y when plotted as a function of X form a straight line.
- Factor analysis is most often used multivariate technique of research studies.
- It is a technique applicable when is a systematic interdependence among a set of observed or manifest variables and the researcher are interested in finding out something more fundamental or latent which creates this commonality.
- There are several methods of factor analysis, but they do not necessarily give some result. As such factor analysis is not a single unique method but a set of techniques.
- Important method of factor analysis is: the centroid method, Principal components method and the maximum likelihood method
- Discriminant analysis is technique to discriminate between two or more mutually exclusive and groups on the basis of some explanatory variables.
- These groups are known a –priori when the criterion variable has two categories, the technique is known as two-group discriminant analysis. When three or more

categories are involved, the technique is referred to as multiple discriminant analysis.

- Discriminant analysis helps the researchers who are interested to understand how consumers differ with respect to demographic and psychographic characteristics.
- Discriminant analysis is also used to predict the group membership. Banks use discriminant analysis to discriminate between the customer who default and who repay the loan in time, based on their age, income, assets, number of dependents, and previous outstanding loan, etc.

TERMINAL QUESTIONS

1. (a) Calculate regression from the following data:

Using mobiles also reduces the amount of physical exercise, causing weight gains. A sample of fifteen 10-year-old children was taken. the number of pounds each child was overweight was recorded (a negative number indicates the children is underweight). Additionally, the number of hours of mobile use per weeks was also recorded. these data are listed here.

Mobile use	42	34	25	35	37	38	31	33	19	29	38	28	29	36	18
overweight	18	6	0	-1	13	14	7	7	-9	8	8	5	3	14	-7

2. (a) What do you understand by regression?

(b) How you can analyze linear regression?

3. (a) Write about factor analysis.

(b) Give a note on discriminate analysis with suitable example

4. (a) Write a note on SPSS

5. (a) Discuss the uses of SPSS

6. (a) Suppose a company wants to determine if one of its new products a new and improved digital camera will be commercially successful or not. The company wants to distinguish the purchaser from non-purchaser. The company makes a panel of 10 potential purchasers and devised the rating scales on three characteristics viz. durability,

performance and style. The panel is asked to give the digital camera ratings from 1 to 10 on these three scales. After the products was evaluated, they are asked their buying intention (would purchase or would not purchase). Five said that they would purchase the digital camera and the other five said that they would not purchase it. Determine best discriminating variable and the poorest determining variable.

The ratings given by panelists are given below:

Group based on purchased intention	Durability (X1)	Performance (X2)	Style (X3)
Group A: Would purchase			
Person 1	8	9	6
Person 2	6	7	5
Person 3	10	6	3
Person 4	9	4	4
Person 5	4	8	2
Group Mean	7.4	6.8	4
Group A: Would purchase			
Person 6	5	4	7
Person 7	3	7	2
Person 8	4	5	5
Person 9	2	4	3
Person 10	2	2	2
Group Mean	3.2	4.4	3.8
Difference between group means	4.2	2.4	0.2

7. (a) Give the Using SPSS in regression analysis.

(b) Give the Using SPSS in factor analysis.

(c) Give the Using SPSS in discriminant analysis.

ANSWERS

1. (a) See the regression analysis in section
2. (a) see the section 12.2

   (b) See the section 12.3
3. (a) See the section 12.4

   (b) See the section 12.5
4. (a) See the section 12.6

5. (a) See the section 12.6

6. (a) See the example of discriminant analysis in section 12.5

7. (a) See the section 12.3.1

   (b) See the section 12.4.1

   (c) See the section 12.5.1

UNIT 13: Cluster analysis and Multivariate analysis

Unit Structure

Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is Cluster analysis?
- About uses and examples of cluster analysis
- What is multivariate analysis?
- About uses of and examples of multivariate analysis

13.1 INTRODUCTION

Cluster analysis is the task of grouping a group of objects in such a manner that objects within the same group are more similar one than to those in other groups (clusters). Cluster analysis is often achieved by various algorithms depending upon the understanding of what constitutes a cluster and the way to efficiently find the clusters. Popular ideas of clusters include groups with small distances between cluster members, dense areas of the info space, intervals or particular statistical distributions. Therefore, cluster analysis can be formulated as a multi-objective optimization problem. Cluster analysis is used in various sciences. A cluster analysis is employed to form spatial and temporal comparisons of communities of organisms in heterogeneous surroundings. It is also used in plant taxonomy to generate artificial phylogenies or clusters of organisms at

the species, genus or higher level that share a number of attributes. In this unit you will learn about the cluster and multivariate analysis.

13.2. Cluster analysis

The purpose of cluster analysis is to divide large group of objects or observations into smaller groups such that the observations within each group are similar or close (or homogeneous) and the observation in different groups are dissimilar or far away. These similar groups are called clusters. Thus, resulting cluster exhibit high internal (within cluster) homogeneity and high external (between clusters) heterogeneity.

Cluster analysis is an inter dependence technique and in general based on correlation or covariance between the variables. Note that, we do not combine the variables in cluster analysis as in factor analysis. In factor analysis, we reduce number of observations by grouping them into smaller set of factors, while in cluster analysis we reduce number of observations by grouping them into smaller set of clusters. Number of groups is not known in advance; they are natural cluster and derived by the data. Some applications of cluster analysis are illustrated below:

1. Identifying people with similar patterns of past purchases so that you can tailor your marketing strategies.
2. Clubbing television show into homogeneous categories based on viewer characteristics.
3. Examining patients with a diagnosis of swine flu to determine if distinct subgroups can be identified based on a symptom checklist and results from pathological test.

In cluster analysis, we group the observations based on a particular set of variables. These variables determine the cluster obtained using cluster analysis. There is no general rule of selecting the variables, however, the variables should be selected based on the objective of the study. After selecting the variables, we obtained sample observations on those variables.

Distance measures: In cluster analysis, we group the observation on the basis of proximity or similarity. These are several different measures of distance or proximity. Some important measures are discussed below:

1. Euclidean distance: suppose, there are k variables denoted as x_1, x_2, …, x_k. The Euclidean distance between two observation (x_{ij}, x_{i2}, ……, x_{ik}) and (x_{j1}, x_{j2}, ……, x_{jk}) is given by

$$[d_{ij} = \sum_{h=1}^{k} (x_{ih} - x_{jh})^2]^{1/2}$$

2. Manhattan distance: between two observation (x_{i1}, x_{i2}, …, x_{ik}) and (x_{ji}, $x_{j2, \ldots,}$ x_{jk}) is given by

$$\sum_{h=1}^{k} |x_{ih} - x_{jh}|$$

3. Chebyshev distance: Between two observation (x_{i1}, x_{i2}, …, x_{ik}) and (x_{ji}, $x_{j2 \ldots,}$ x_{jk}) is given by

$$\max|x_{ih} - x_{jh}|$$

4. Mahalanobis or correlation distance: Between two observations x_i = (x_{i1}, x_{i2}…., x_{ik}) and x_j= (x_{j1}, x_{j2}…., x_{jk}) is

$$d_{ij} = (x_i - x_j)^T \sum^{-1} (x_i - x_j)$$

Where, $\sum$ is population covariance matrix of the variables (X_1, X_2, ……, X_k)

5. Correlation: Between two observation (x_{ij}, x_{i2}, …, x_{ik}) and (x_{ji}, $x_{j2.,\ldots,}$ x_{jk}) is given by

Where, $\bar{x}_i$ and $\bar{x}_j$ are corresponding means.

6. Matching measure: It is used when the observations are categorical. It is calculated by number of matches between two categories in all the corresponding variables and divided by number of variables. For example, consider the following example.

Person	Smoker (Yes=1, No=0)	Gender (Female=1, Male=0)	Adult (Yes=1, No=0)	Graduate (Yes=1, No=0)
1	0	1	1	1
2	1	0	0	0
3	1	1	0	0
4	1	1	1	0
5	1	1	1	0

In this data, distance between person 1 and person 2 is given by number of matches divided by number of variables, which is 0/4 or 0. we can write all such distances as below:

Person	1	2	3	4	5
1	1				
2	0/4	1			
3	1/4	3/4	1		
4	2/4	2/4	3/4	1	
5	2/4	2/4	3/4	4/4	1

Euclidean, Manhattan and Chebyshev distances depend upon units of measurement. In case units of measurement differ, it is advisable to standardize each variable. Standardization is used to avoid too much weight to variables whose scale is larger. For standardization, we subtract each observation by the mean and then divide by the standard deviation. For example, to standardize the observation on X_1, given by (x_{11}, x_{12}, ……x_{1n}), we obtain mean $\bar{x}_1$ and standard deviation s_1 of this observation and transform each observation using $(\bar{x}_{1i}-\bar{x}_1)/s_1$ for i=1,2, ……n. These standardized observations are called as Z-scores.

13.3. Clustering Algorithm

There are several clustering algorithms. The popular algorithms are divided into two types i.e., non-hierarchical clustering and hierarchical clustering. These methods are explained below.

Non – hierarchical clustering: The method is also known as k-mean clustering. The steps involved in this method are described below:

a. Fix k, number of clusters in advance.

b. Take the initial guess of cluster centroids or means. The popular approach of this is to take the observations (or cases) having highest distances.

c. Proceed through the list of cases assigning a case to the cluster whose centroid is the nearest. Calculate the centroid of new clusters receiving the new case and the cluster losing an old case.

d. Repeat step (c) until the reassignment of cases occurs.

The final assignment of cases to clusters depends on the initial guess of k means. In order to check the stability of the clustering, it is desirable to run the algorithm again with a new initial guess. How to select the number of clusters k is critical and decided by the researcher what value of k yields the best solution. When the number of clusters is smaller, the solution looks simpler. However, the adequacy of the solution may increase with the number of clusters. K-means clustering is severely affected by outliers, since they will usually be selected as in initial cluster centers. This will result in outliers forming clusters with small numbers of cases. Therefore, researcher must screen the data for outliers before starting a cluster analysis.

Example-1: Following data shows the preference of 25 students of 5 varieties of carbonated soft drink:

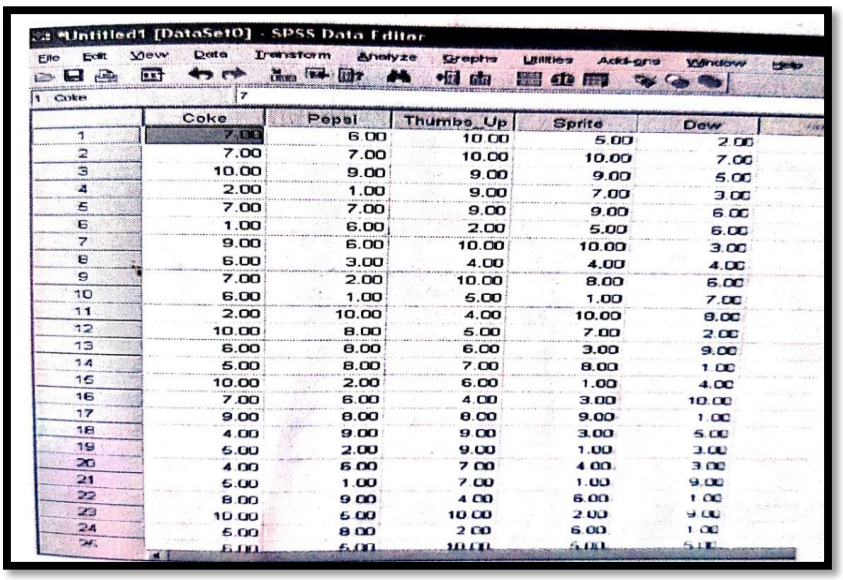
Student	Coke	Pepsi	Thumbs Up	Sprite	Dew
1	7	6	10	5	2
2	7	7	10	10	7
3	10	9	9	9	5
4	2	1	9	7	3
5	7	7	9	9	6
6	1	6	2	5	6
7	9	6	10	10	3
8	6	3	4	4	4
9	7	2	10	8	6
10	6	1	5	1	7
11	2	10	4	10	8
12	10	8	5	7	2
13	6	8	6	3	9
14	5	8	7	8	1
15	10	2	6	1	4
16	7	6	4	3	10
17	9	8	8	9	1
18	4	9	9	3	5
19	5	2	9	1	3
20	4	6	7	4	3

21	5	1	7	1	9
22	8	9	4	6	1
23	10	5	10	2	9
24	5	8	2	6	1
25	6	5	10	5	5

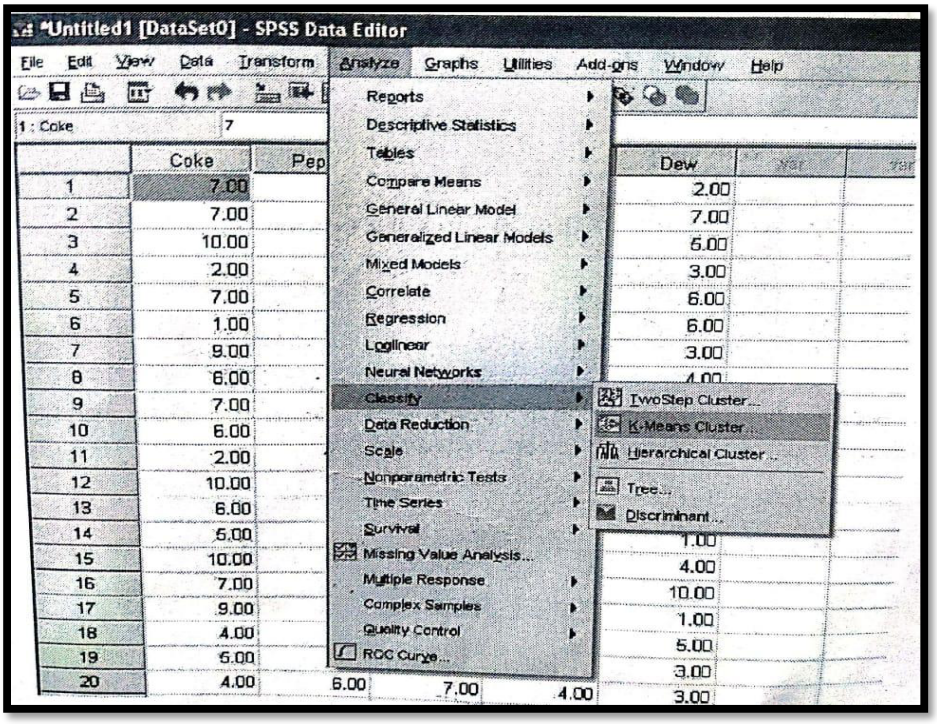
Conduct k-mean clustering with some suitable value of k and obtain homogenous clusters.

Solution: We use SPSS as below:

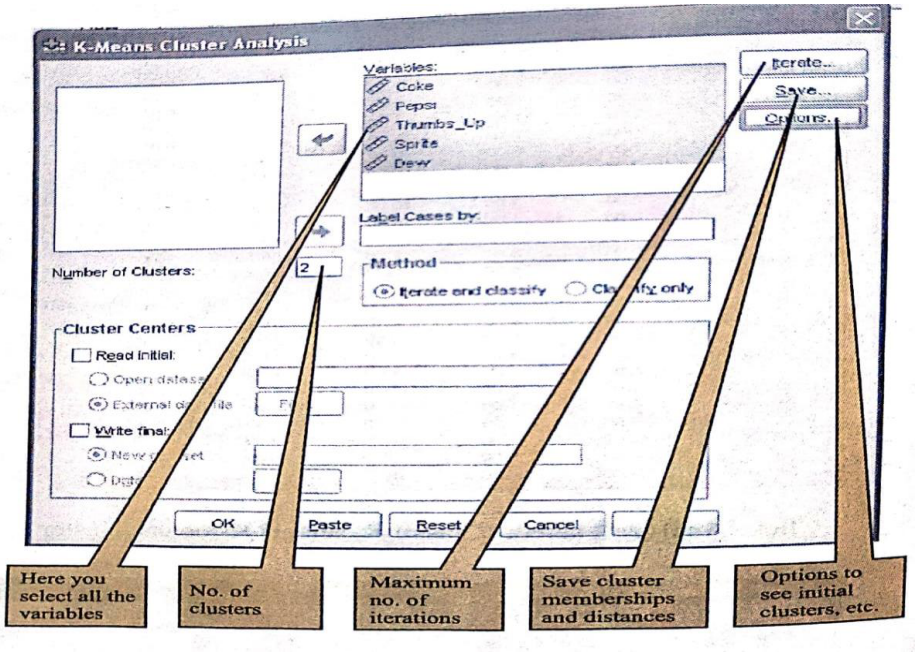
We first enter the data in SPSS worksheet as below.



Click “Analyze”, select “Classify”, select “K-means cluster”



A window pops up



We get the following outcome:

Initial Cluster Centers

	Cluster	
	1	2
Coke	5.00	8.00
Pepsi	1.00	9.00
Thumbs Up	7.00	4.00
Sprite	1.00	6.00
Dew	9.00	1.00

Interaction History[a]

Interaction	Change in Cluster Centers	
	1	2
1	4.439	4.626
2	0.000	0.000

[a]Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current interaction is 2. The maximum distance between initial centers is 13.077.

Final Cluster centers

	Cluster	
	1	2
Coke	6.36	6.29
Pepsi	3.27	7.64
Thumbs Up	7.27	6.86
Sprite	3.27	7.21
Dew	6.27	3.64

Number of cases in each cluster

	Cluster	
	1	2
Valid	11.000	14.000
Missing	25.000	0.000

Outcomes are easy to interpret. From "Final cluster centers", it is clear that the preference of Coke and Thumbs Up are almost the same in both the clusters.

Hierarchical clustering: There are two methods of hierarchical clustering i.e., divisive and agglomerative.

In divisive method, all the cases or observations are first kept in one cluster. At each successive stage, the cluster is divided into more clusters. It ends up having as many clusters as the number of cases was considered. We start with having one cluster and ends with n clusters, where n is the number of cases or observations. In this method, once the cluster is divided into two or more clusters. They are not be merged.

In agglomerative method, we first consider each case as cluster. Then cluster are combined. It ends up having only one cluster of all the cases. Thus, we start with n cluster and ends with one cluster. In this method, once the cluster are merged, they cannot be split.

In these methods, first and last iterations are not the best solutions. We need to know the number of clusters we require. There is no right or wrong answer as to how many clusters we should have. To find a good number of clusters, we should study the solution at each step and decide a reasonable number of homogeneous clusters representing the data.

We also require a suitable distance measure and a criterion to decide which cluster are divided or merged at successive steps.

13.4. Agglomerative clustering

Agglomerative clustering: In this section, we describe agglomerative clustering procedure in detail. For agglomerative clustering we need

1. A suitable distance measures.
2. A criterion to combine the cluster.
3. Number of clusters.

We have already discussed about some distance measure. It is advisable to standardize the data so that no variable is under/over represented. In subsequent section, we discuss about various criteria of combining the clusters.

Combining clusters: In Agglomerative hierarchical clustering, we start with each case being a cluster. At the next step, the two cases having the smallest distance are joined into a single cluster. At every step, we combine

1. Two individual cases or

2. An individual cases to an existing cluster, or
3. Two existing clusters.

Two distances between individual cases in unambiguous. It is the same as the value of distance measure. However, when we need to combine two clusters or a cluster and a case, we need to define the distance between two clusters. Some popular criteria of combining cluster or measure of distance between two clusters are given below:

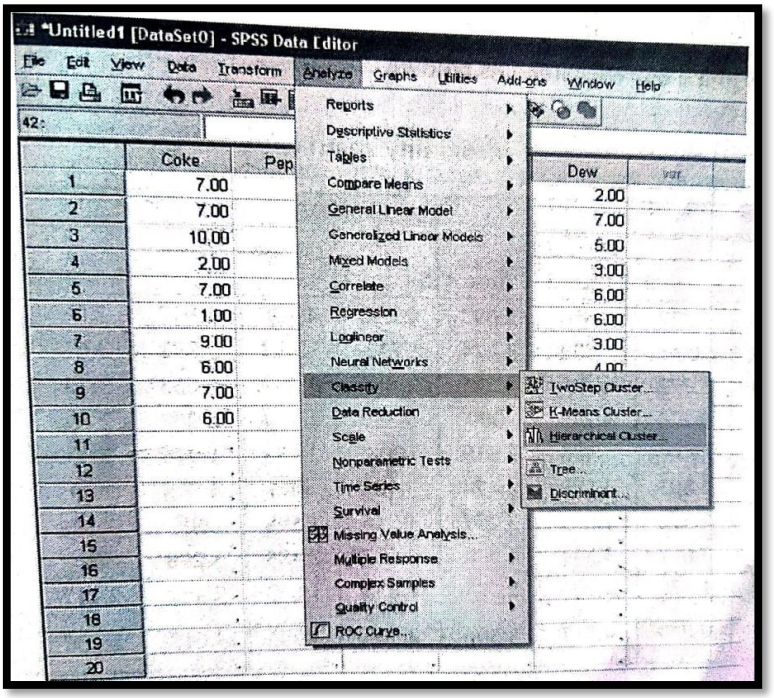
1. Nearest neighbor: It is defined as the smallest distance between two cases in the different clusters.
2. Furthest neighbor (complete linkage): The distance between two clusters is defined as the distance between the two further points.
3. UPGMA (unweighted pair–group method using arithmetic averages): This is defined as the average of the distance between all pairs of cases in which one members of the pair is form each of the cluster. Since, it uses information about all pairs of distance, and it is preferred to the single linkage and complete linkage.
4. Wards method: First we calculate the mean of all variables for each cluster. Then, for each case, the squared Euclidean distance to the cluster means is calculated. These distances are summed for all of the cases. At each step, those two clusters are combined which result in the smallest increase in the overall sum of the squared within –cluster distances.
5. Centroid method: First we calculate the mean of all variables for each cluster. The distance between two clusters is given by the sum of distance between cluster means.

Using SPSS: We illustrate the usage of SPSS and explain the outcomes through the following example:

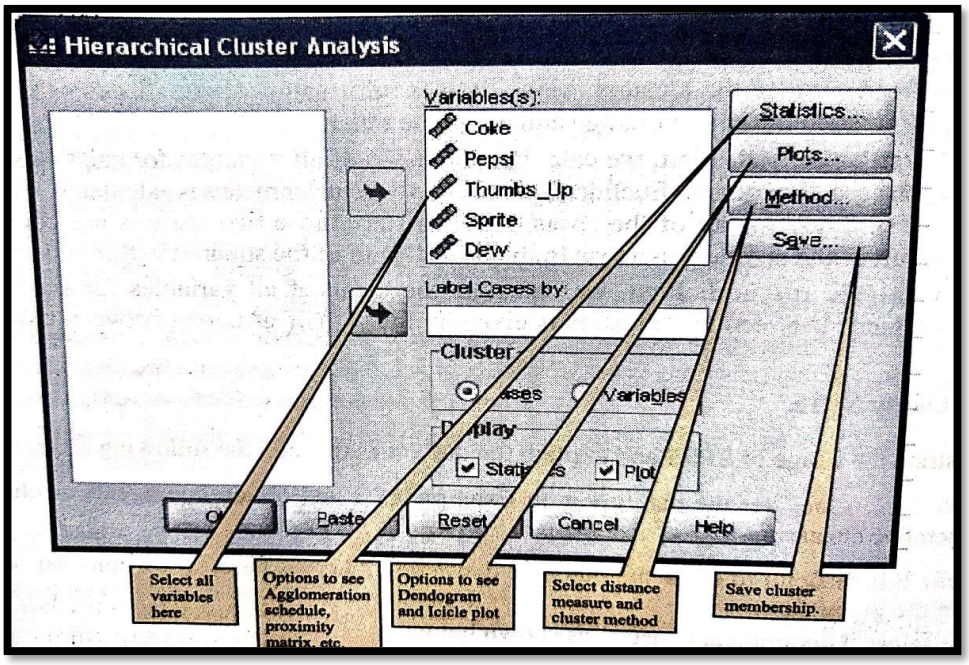
Example- Consider the data of example -1 and obtain homogenous clusters using agglomerative clustering.

Solution: It is difficult to show the outcome consisting 25 cases here, so we consider only first 10 cases in the given data.

We select “Hierarchical Cluster’ as shown below:



Following window Pop up:



The outcome is given and explained as below:

Proximity Matrix

Case	Euclidean Distance									
	1: Case1	2: Case2	3: Case3	4: Case4	5: Case5	6: Case6	7: Case7	8: Case8	9: Case9	10: Case10
1:Case1	.000	3.277	2.650	2.693	2.652	4.108	1.912	2.586	2.830	3.961
2:Case2	3.277	.000	1.772	3.724	.733	3.893	2.372	3.617	1.976	4.079
3:Case3	2.650	1.772	.000	4.245	1.406	4.389	1.653	3.544	2.804	4.517
4:Case4	2.693	3.724	4.245	.000	3.319	3.474	3.257	2.594	2.527	3.596
5:Case5	2.652	.733	1.406	3.319	.000	3.483	1.916	3.007	1.835	3.743
6:Case6	4.108	3.893	4.389	3.474	3.483	.000	4.598	2.485	3.863	3.084
7:Case7	1.912	2.372	1.653	3.257	1.916	4.598	.000	3.278	2.405	4.612
8:Case8	2.586	3.617	3.544	2.594	3.007	2.485	3.278	.000	2.717	2.106
9:Case9	2.830	1.976	2.804	2.527	1.835	3.863	2.405	2.717	.000	2.989
10:Case10	3.961	4.079	4.517	3.596	3.743	3.084	4.612	2.106	2.989	0.000

Proximity Matrix: The proximity matrix gives the distances between each pair of cases. Note that the diagonal elements are zero, as a case has zero distance from itself.

Agglomeration Schedule

Stage	Cluster combined		Coefficients	Stage cluster first appear		Next stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	5	0.367	0	0	4
2	3	7	1.193	0	0	5
3	8	10	2.246	0	0	6
4	2	9	3.395	1	0	8
5	1	3	4.640	0	2	8
6	6	8	6.145	0	3	7
7	4	6	7.921	0	6	9
8	1	2	9.700	5	4	9
9	1	4	13.628	8	7	0

Agglomeration Schedule: In agglomeration schedule, we get the schedule of cluster agglomeration. In first step cluster 2 (case 2) and cluster 5 (Case 5) are merged. The new cluster made by combining these two cases will be called as cluster 2 (smallest number). This cluster 2 is next time used in stage 4. In coefficients, we get the distance or measure obtained using cluster combining method (here we use Ward's method). These coefficients are used to decide how many clusters are needed to represent the data. We want to stop cluster formation when the increase in the coefficients column between two adjacent steps is large. In the given example we can stop cluster formation after 8th stage. Thus, we have two clusters.

In this example, we may like to stop at the four clusters solution, after stage 3. We can write the schedule as below also:

Stage 1: (2) + (5) = (2,5)

Stage 2: (3) + (7) = (3,7)

Stage 3: (8) + (10) = (8,10)

Stage 4: (2,5) + (9) = (2,5,9)

Stage 5: (1) + (3,7) = (1,3,7)

Stage 6: (6) + (8,10) = (6,8,10)

Stage 7: (4) + (6,8,10) = (4,6,8,10)

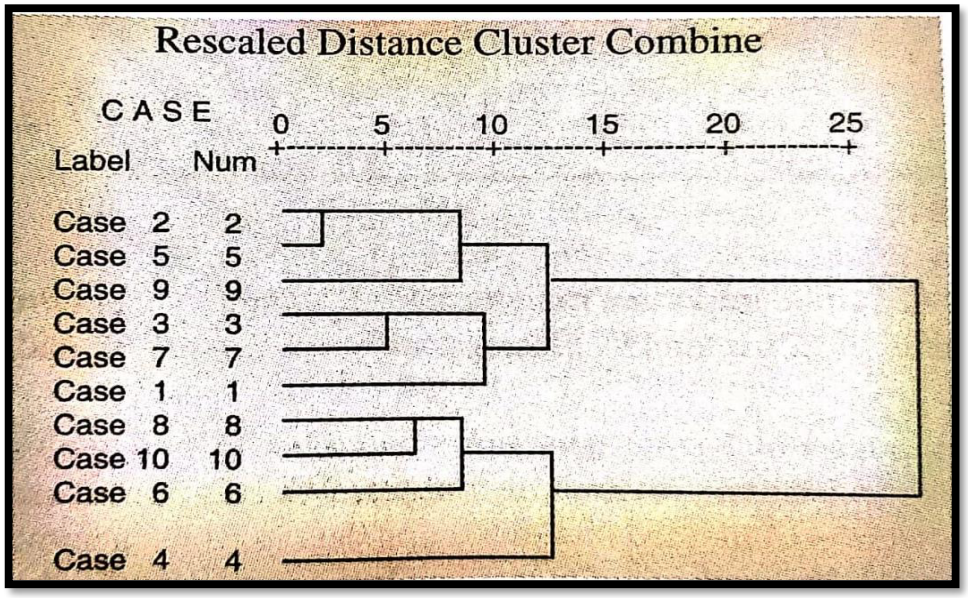
Stage 8: (1,3,7) + (2,5,9) = (1,2,3,5,7,9)

Stage 9: (1,2,3,5,7,9) + (4,6,8,10) = (1,2,3,4,5,6,7,8,9,10)

Here the bracket '( )' shows the cluster and the number inside the bracket shows cases in the cluster.

Hierarchical Cluster analysis

Dendrogram using Ward Method:



Dendrogram: The word dendrogram is made by two Greek words first is Dendron which means tree and second is gramma which means drawing. Dendrogram is used to illustrate the arrangement of the cluster produced by hierarchical clustering. It gives a visual representation of the distance at which cluster are combined. It is read from left to right. Vertical lines show joined cluster. The positions of these lines showed the distance at which cluster are joined. These distances in the plot are not the actual distances; they are proportional to the actual distances. The first vertical line corresponds to case 2 and case 5. Based on this, we can decide having 2 clusters i.e., cluster 1 having cases 2,5,9,3,7 and cluster 2 having cases 8,10,6 and 4.

13.5. Multivariate analysis

We have discussed some important multivariate viz. factor analysis, cluster analysis etc. in previous units of this course. All statistical method which concurrently analyze more than two variables on a sample of observations can be classified as multivariate techniques. We may as well use the term “Multivariate analysis” which is the collection of methods for analyzing the data in which number of observations are available for each object. For example, in the field of intelligence testing if you start with the theory that general intelligence is reflected in a variety of specific performance measures, then to study intelligence in the context of this theory one must administer many tests of mental skills,

such as vocabulary, speed of recall, mental arithmetic, verbal analogies etc. The score on each test is one variable, Xi and there are several, k, of such score for each object, represented as X_1, X_2,…,…X_k. Most of the researches include more than two variables in which situation analysis is preferred of the association between one criterion variable and several independent variables, or we may be required to study the relationship between variables having no dependency relationship. All such analysis is termed as multivariate analysis. In short, techniques that take explanation of the joint relationships among more than two variables are termed as multivariate analysis.

Multivariate techniques have emerged as a powerful tool to analyze data represented in terms of many variables. The main cause behind that a series of univariate analysis conceded out individually for each variable may, at times, lead to incorrect interpretation of the consequences. It is, because univariate analysis does not consider the 1joint relationship among the variables. As a result, during the last 50 years a number of statisticians have contributed to the development of several multivariate techniques. Today, these techniques are being applied in many fields such as sociology, agricultural science, biology, medicines, environment, forestry etc. These techniques are used in analyzing data of above-mentioned sciences. Applications of multivariate techniques in practice have been accelerated in modern times because of the advent of high-speed computers.

13.6. Characteristics and applications of Multivariate analysis

Most of the multivariate techniques are empirical and deal with the reality. Accordingly, in most of the applied and behavioural researches, we generally use multivariate analysis techniques for realistic results. Besides, being a tool for analyzing the data, multivariate techniques also help in various types of decision making. For instance, take the case of entrance examination wherein a number of entrance tests are managed to candidates, and the candidates scoring high total marks based on subjects are admitted. This system, though, apparently fair, may at time be biased in favour of some subjects with the larger standard deviation. Multivariate analysis may be suitably used in such situation for developing standards as to who should be admitted in university. We can take the example from the medical field. Many medical examinations such as blood pressure and cholesterol

test are administered to patients. Each of the result of such examinations has significance of its own, but it is also important to consider relationships between different test results or results of the same tests at different occasions in order to draw proper diagnostic conclusions and to determine an appropriate therapy. Multivariate analysis can assist in such a situation.

The basic objective underlying multivariate analysis is to present a collection of massive data in a simplified way. In other words, multivariate technique transforms a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in a raw data obtained from a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into simplified visible form.

For better appreciation and understanding of multivariate techniques, one must be familiar with fundamental concepts of linear algebra, vector spaces, orthogonal and oblique projections and univariate analysis. Even then before applying multivariate techniques for meaningful results, one must consider the nature and structure of data and the real aim of the analysis. We should also not forget that multivariate techniques do involve several complex mathematical computations and as such can be utilized largely with the availability of computing facility.

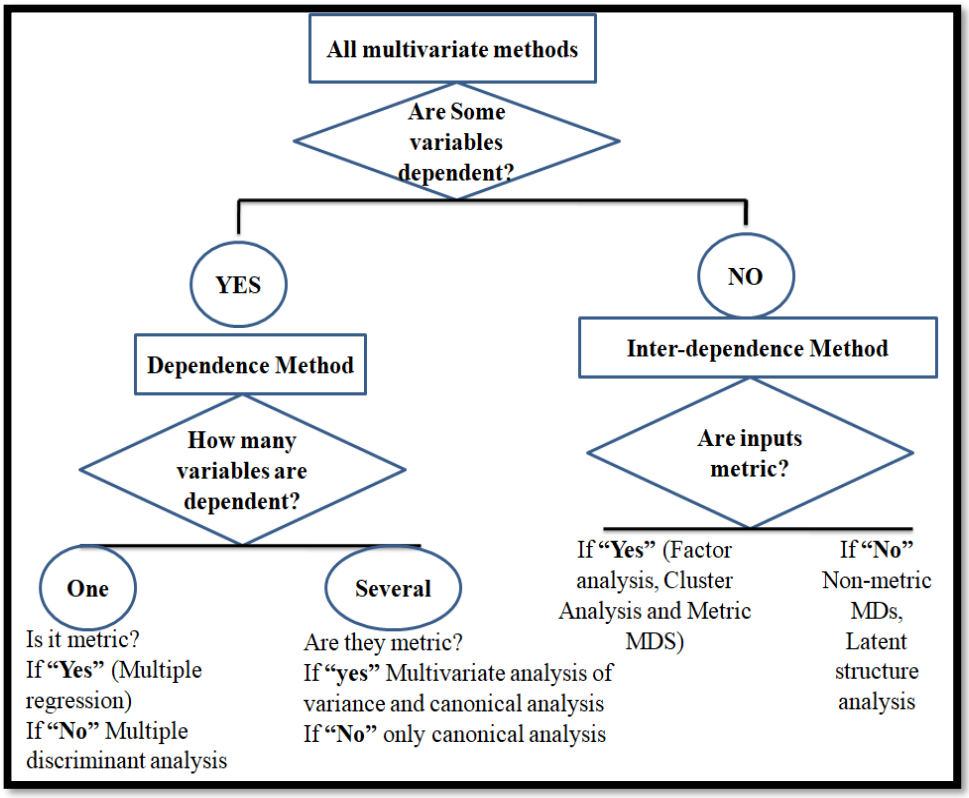
Description of the data

Let's pursue Example 1 from above. We have a hypothetical dataset with 600 observations on seven variables. The psychological variables are locus of control (**locus of control**), self-concept (**self-concept**), and motivation (**motivation**). The academic variables are standardized tests scores in reading (**read**), writing (**write**), and science (**science**), as well as a categorical variable (**prog**) giving the type of program the student is in (general, academic, or vocational).

13.7. Classification of Multivariate techniques

Today, there exit a great variety of multivariate techniques which can be conveniently classified into two broad categories viz. dependence method and interdependence methods. This sort of classification depends upon the question: Are some of the involved

variable's dependent upon other? If the answer is yes we have dependence method but in case the answer is no we have interdependence methods. Two more questions are relevant for understanding the nature of multivariate techniques. Firstly, in case some variables are dependent, the question is how many variables are dependent? The other question is, whether the data are metric or non-metric? This means whether the data are quantitative, collected on intervals or ratio scale or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for given situation depends upon the answer to all these very questions. Jadish N. Seth in his article on "The multivariate revolution in marketing research" has given the flow chart that clearly exhibits the nature of some important multivariate techniques.

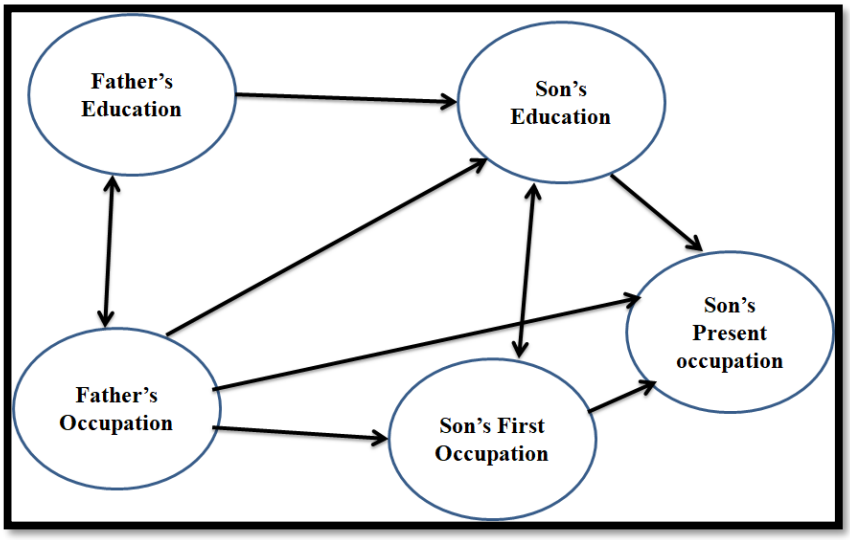


Thus, we have two types of multivariate techniques i.e., one type for data containing both dependent and independent variables and the other types for data containing several variables without dependency relationship. In the former category, the techniques are multiple regression analysis, multiple discriminant analysis, multivariate analysis of variance and canonical analysis, whereas the latter category the techniques like factor

analysis, cluster analysis, multidimensional scaling or MDS (both metric and non-metric) and the latent structure analysis.

There are various multivariate techniques such as path analysis, canonical correlation, multivariate ANOVA, multidimensional scaling and latent structure analysis.

I. **Path analysis:** Path analysis is a method to discern and assess the effects of multivariate data which facilitate for a specified outcome through multiple causal pathways. The term Path analysis was first introduced by the biologist Sewall Wright in the year 1934 in connection with decomposing the total correlation between any two variables in a casual system. The technique of path analysis is based on series of multiple regression analysis with the added assumption of casual relationship between independent and dependent variables. This technique lays relatively heavier emphasis on the heuristic use of visual diagram, technically described as path diagram. An illustrative path diagram showing interrelationship between father's education, father's occupation, Son's education, Son's first and Son's present occupation can be shown in following figure.

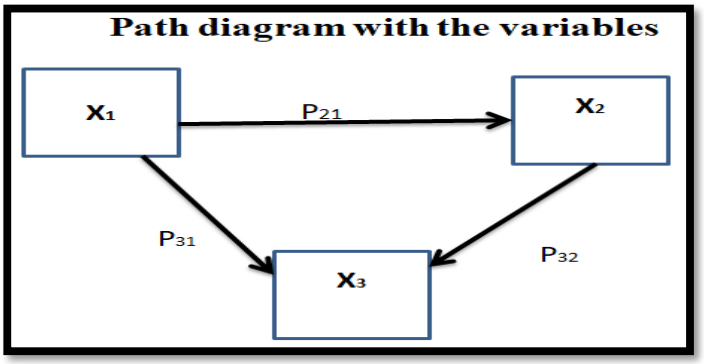


In path diagram, the pattern of relationships among the variables is represented and described by a diagram. The diagram is actually a type of directed graph. In the diagram, variables linked by straight arrows indicate the directions of the causal relationships between the variables. Straight arrows may direct towards one direction, and assumed that

the variable cannot be both a cause and an effect of another variable. Curved, double-headed arrows indicate correlation between the variables.

Path analysis makes use of standardized partial regression coefficients (known as beta weights) as effect coefficients. If linear additives effects are assumed, then through path analysis a simple set of equation can be built up showing how each variable depends on preceding variables. "The main principle of path analysis is that any correlation coefficient between two variables or a gross or over all measure of empirical relationship can be decomposed into a series of parts: separate paths of influence leading through chronologically intermediate variable to which both the correlated variables have links"

The merit of path analysis in comparison to correlational analysis is that makes possible the assessment of the relative influence of each antecedent or explanatory variable on the consequent or criterion variable by first making explicit the assumptions underlying the casual connections and then by elucidating the indirect effect of the explanatory variables. The use of the path analysis technique requires the assumption that there are linear additive a symmetric relationships among a set of variable which can be measured at least on a quasi-interval scale. Each dependent variable is regarded as determined by the variables preceding it in the path diagram, and a residual variable, defined as uncorrelated with the other variables, is postulated to account for the unexplained portion of the variance in the dependent variable. The determining variables are assumed for the analysis to be given. We may illustrate the path analysis technique in connection with a simple problem of testing a casual model with three explicit variables as shown in the following path diagram.



The structural equation for above can be written as

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} e_1 \\ P_{21}X_1+e_2 \\ P_{31}X_2+P_{32}X_2+e_3 \end{pmatrix} = pX+e$$

Where, the X variables are measured as deviations from their respective means. P_{21}may be estimated from the simple regression of X_2 on X_1 i.e. $X_2=B_{21}X_1$ and P_{32} may be estimated from the regression on X_3 on X_2 and X_1 as under:

$$\hat{X}_3= b_{31.2}\, X_1+b_{2.1}X_2$$

Where, $b_{31.2}$means the standardized partial regression coefficient for predicting variable 3 from variable 1 when the effect of variable 2 is held constant.

In path the beta coefficient indicates the direct effect of Xj (j=1,2,3,….p) on the dependent variable. squaring the direct yields the proportion of the variance in the dependent variable Y which is due to each of the p number of independent variables Xj (i=1,2,3….,p). After calculating the direct effect, one may then obtain a summary measure of the total indirect effect of Xj on the dependent variable Y by subtracting from the zero-correlation coefficient r_{yxj} the beta coefficient bj i.e.

Indirect effect of X_j on $Y=c_{jy}=r_{yxj}-b_j$

For all j=1,2,….p.

Such indirect effects include the unanalyzed effects and spurious relationship due to antecedent variables. In the end, it may again be emphasized that the main virtue of path analysis lies in making explicit the assumption underlying the casual connections and in elucidating the indirect effects due to antecedent variables of the given system.

Path analysis is theory-driven and the same data set can be described by many different causal patterns depending upon the considered theory. Therefore, it is essentially desired to have ana priori model describing the causal relationships among the variables considered for the analysis.

II. **Canonical correlation:** This technique was first developed by Hotelling wherein an effort is made to simultaneously predict a set of criterion variables from their joint co-variance with set of explanatory variables. A canonical correlation is used to identify and measure the associations among two sets of variables and is a correlation between two latent types of variables. In this, one variable is an independent variable and the other variable is a dependent variable. Both metric and non-metric data can be used in the context of canonical correlation. The analysis is suitable for the situation which has multiple intercorrelated outcome variables. The procedure followed is to obtain a set of weights for the dependent and independent variables in such a way that linear composite of the criterion variables has a maximum correlation with the linear composite of the explanatory variables. For example, if we want to relate grade school adjustment to health and physical maturity of the child, we can then use canonical correlation analysis provided, we have for each child a number of adjustment scores (such as tests, teacher's ratings, parent's rating and so on) and also, we have for each child a number of health and physical maturity scores (such as heart rate, height, weight, index of intensity of illness and so on). The main objective of canonical correlation analysis is to discover factors separately in the two sets of variables such that the multiple correlations between sets of factors will be the maximum possible. Mathematically, in canonical correlation analysis, the weights of the two sets viz. a_1, a_2, …., a_k and y_1, y_2, y_3, ….., yj are so determined that the variables $X=a_1X_1 + a_2X_2 + …. + a_kX_k + a$ and $Y=y_1Y_1 + y_2Y_2 + yjY_j + y$ has a maximum common variance. The process of finding the weights requires factor analysis with two matrices. The resulting canonical correlation solution then gives an overall description of the presence or absence of a relationship between the two sets of variables.

III. **Multidimensional scaling:** Multidimensional scaling (MDS) allows a researcher to measure an item in more than one dimension at a time. The basic assumption is that people perceive a set of objects as being more or less similar to one another on a number of dimension (usually uncorrelated with one another) instead of only

one. There are several MDS techniques for dimensional reduction) often used for the purpose of revealing patterns of one sort or another in interdependent data structures. If data happen to be non-metric, MDS involves rank ordering each pair of objects in terms of similarity. Then the judged similarities are transformed into distances through statistical manipulations and are consequently shown in n-dimensional space in a way that the inter-point distances best preserve the original inter-point proximities. After this sort of mapping is performed, the dimension is usually interpreted and labeled by the researcher.

The significance of MDS lies in the fact that it enables the researchers to study "The perceptual structure of a set of stimuli and the cognitive processes underlying the development of this structure. MDS provides a mechanism for determining the truly salient attributes without forcing the judge to appear irrational". With MDS, one can scale objects, individuals or both with a minimum of information. The MDS analysis will reveal the most salient attributes which happen to be the primary determinants for making a specific decision.

IV. **Multivariate ANOVA:** Multivariate analysis of variance is an extension of bivariate analysis of variance in which the ratio of among groups variance to within groups variance is calculated on a set of variables instead of a single variable. This technique is considered appropriate when severe; metric dependent variables are involved in a research study along with many non-metric explanatory variables. (But if the study has only one metric dependent variable and several non-metric explanatory variables, then we use the ANOVA). In other words, multivariate analysis of variance is specially applied whenever the researcher wants to test hypothesis concerning multivariate differences in group responses to experimental manipulations. For example, the market researcher may be interested in using one test market and one control market to examine the effect of an advertising campaign on scale as well as awareness, knowledge and attitudes. In that case he/she should use the technique of multivariate analysis of variance for meeting his/her objective.

V. **Latent structure analysis:** This type of analysis share both the objectives of the factor analysis viz. to extract latent factors and express relationship of observed (manifest) variables with these factors as their indicators and to classify a population of respondents into pure types. This type of analysis is appropriate when the variables involved in a study do not possess dependency relationship and happen to be non-metric.

13.8. SUMMARY

In this unit we have discussed various aspects of cluster and multivariate analysis. So far you have learnt that:

- Cluster analysis is also known as clustering. This is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (clusters). Cluster analysis itself is not one specific algorithm, but the general task to be solved. The purpose of cluster analysis is to divide large group of objects or observations, like customers or products, into smaller groups such that the observations within each groups are similar or close (or homogeneous) and the observation in different groups are dissimilar or far away. These similar groups are called clusters. Thus, resulting cluster exhibit high internal homogeneity and high external (between clusters) heterogeneity.
- Cluster analysis is an inter dependence technique and makes no distinction between dependent (study) and independent (explanatory) variables. Note that, we do not combine the variables here as done in factor analysis. Nearest neighbor is defined as the smallest distance between two cases in the different clusters. Furthest neighbor (complete linkage) is distance between two cluster is defined as the distance between the two further points.
- UPGMA is defined as the average of the distance between all pairs of cases in which one members of the pair id form each of the cluster. Since, it uses information about all pairs of distance, and it is preferred to the single linkage and complete linkage.

- In Wards method, first we calculate the mean of all variables for each cluster. Then, for each case the squared Euclidean distance to the cluster means is calculated. These distances are summed for all of the cases. At each step, those two clusters are combined which result in the smallest increase in the overall sum of the squared within –cluster distances.
- In Centroid method, first we calculate the mean of all variables for each cluster. The distance between two clusters is given by the sum of distance between cluster means.
- The word dendrogram is made by two Greek words first is Dendron which means tree and second is gramma which means drawing. Dendrogram is used to illustrate the arrangement of the cluster produced by hierarchical clustering. It gives a visual representation of the distance at which cluster are combined. It is read from left to right.
- Multivariate techniques are largely empirical and deal with the reality. They possess the ability to analyze complex data. Accordingly, in most of the applied and behavioural researches, we generally restore to multivariate analysis techniques for realistic results.
- The basic objective underlying multivariate techniques is to present a collection of massive data in a simplified way.
- Today, there exit a great variety of multivariate techniques which can be conveniently classified into two broad categories viz. dependence method and interdependence methods.
- There are various multivariate techniques such as path analysis, canonical correlation, multivariate ANOVA, multidimensional scaling and latent structure analysis.

TERMINAL QUESTIONS

1. (a) Fill in the blank spaces with appropriate words.
Today, there exit a great variety of …….. …….. which can be conveniently classified into …….broad categories viz. ……method and ……methods.

This sort of classification depends upon the question: Are some of the involved variables dependent upon other? If the answer is Yes we have ……method but in case the answer is ……. we have ……..methods. Two more questions are relevant for understanding the nature of ……..techniques. Firstly, in case some variables are……, the question is how many variables are……..? The other question is, whether the data are metric or……..? This means whether the data are……., collected on intervals or ratio scale or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for given situation depends upon the answer to all these very questions. ……. …. ……..in his article on "The multivariate revolution in marketing research" has given the flow chart that clearly exhibits the nature of some important multivariate techniques.

2. (a) What is cluster analysis?

(b) Discuss about the cluster algorithm.

3. (a) What is agglomerative clustering?

(b) Write a short note on dendrogram?

4. (a) Discuss the multivariate analysis.

5. (a) Discuss characteristics and applications of multivariate analysis?

6. (a) The term Path analysis was first introduced by the biologist **……. ……** in the year ……. in connection with decomposing the total ……between any two variables in a casual system. The technique of ……. …….. is based on series of multiple ……analysis with the added assumption of casual relationship between ……..and dependent variables. This technique lays relatively heavier emphasis on the heuristic use of ……diagram, technically described as …….. diagram.

7. (a) Write a short note on path analysis.

(b) Write about canonical correlation

(c) Write about multidimensional scaling and multivariate ANOVA?

ANSWERS

1. (a) multivariate techniques, two, dependence, interdependence, dependence, no, interdependence, multivariate, dependent, dependent, non-metric, quantitative, Jadish N. Seth

2. (a) see section 13.2

(b) See section 13.3

3. (a) See section 13.4

(b) See section13.4 under heading dendrogram

4. (a) See section 13.5

5. (a) See the section 13.6

6. (a) Sewall Wright, 1934, correlation, path analysis, regression, independent, visual, path

7. (a) See the section 13.7 under heading path analysis

(b) See the section 13.7under heading canonical correlation

(c) See the section 13.7 under headings multidimensional scaling and multivariate ANOVA

Unit 14: Applications of remote sensing and GIS in Environmental studies: Case study of land use and land cover change; urban sprawling; mining

Unit Structure

14.0. Learning Objectives

After studying this unit, you will be able to answer the following questions:

- What is Remote sensing?
- What are advantages and disadvantages of remote sensing?
- What is GIS?
- What are advantages and disadvantages of GIS?
- Describe the applications of remote sensing.
- Describe the applications of GIS in environmental studies.
- What is urban sprawling?
- What is mining hazards?

14.1. Introduction

The word Environment is from the French word “Environ” which means surrounding. The Environment is anything which is surrounding us. Environment may be air, water, food, biological diversity and other non-living components. The scientists/researchers conduct researches on different aspects of environment and use various types of techniques to conduct researches or monitor the environmental components. Remote sensing and GIS are a powerful technique for data generation, survey, analysis, and management of natural resources (land, water, forest etc.). Environmentalists use remote sensing for variety of purposes such as detect environmental changes in an area; to measure changes in land use in an area.

Remote sensing can also be used for evaluation of urbanization. Urbanization is process in which rural area converted into urban area and sometimes called urban sprawling. Urban sprawling has positive as well as negative consequences on the environment. Urban sprawling can also be evaluated by remote sensing. Land use can also be analyzed by remote sensing. Land use is term which is used for purpose of land and land cover is something which cover the land. This cover may be building, agriculture crop, solid waste etc. Land uses can be categorized into many category such as agriculture, constriction of buildings etc. In this unit you will learn about the Remote sensing, GIS and their uses in environmental studies. You will also learn about the urban sprawling, land use, land cover and mining hazards.

14.2. Meaning and definitions of Remote sensing

As you know the word “remote” means far away. The instrument or sensor which can sense the obejct from the distance without physical contact is known as remote sensing. Remote sensing technique can be used in different sciences such as earth science, geology, geography, hydrobiology, oceanography, glaciology, etc. In many contries, remote sensing is also used for security purposes. Remote sensing is satellite based technology in which we take the aerial photographs of earth.

There are two types of remote sensing i.e. active remote sensing and passive remote sensing. On the basis of source of electromagnetic energy, remote sensing can be

classified as passive or active remote sensing. In passive remote sensing, source of energy is naturally available such as the Sun. Most of the remote sensing systems work in passive mode using solar energy as the source of electromagnetic radiation (EMR). In active remote sensing, energy is generated and sent from the remote sensing platform towards the targets. The energy reflected back from the targets are recorded using sensors onboard the remote sensing platform. Most of the microwave remote sensing is done through active remote sensing. As a simple analogy, passive remote sensing is similar to taking a picture with an ordinary camera whereas active remote sensing is analogous to taking a picture with camera having built-in flash

Remote sensing is a technique of science of obtaining information about an object without physical contact with that object. People apply remote sensing in their day-to-day business, through their vision, hearing and sense of smell. The data collected can be of many forms: variations in acoustic wave distributions, variations in force distributions, variations in electromagnetic energy distributions etc. These remotely collected data through various sensors may be analyzed to obtain information about the objects or features under observation. Thus, remote sensing is the process of inferring surface parameters from measurements of the electromagnetic radiation (EMR), which can either be reflected or emitted from the Earth's surface. In other words, remote sensing is detecting and measuring electromagnetic (EM) energy emanating or reflected from distant objects made of various materials, so that we can identify and categorize these objects by class or type, substance and spatial distribution.

Specific object reflects different amounts of energy in different bands of the electromagnetic spectrum. The amount of energy reflected depends on the properties of both the material and the incident energy (angle of incidence, intensity and wavelength). Detection and discrimination of objects or surface features is identified through the uniqueness of the reflected or emitted electromagnetic radiation from the object. A device to detect this reflected or emitted electro-magnetic radiation from an object is called a "sensor". A vehicle used to carry the sensor is called a "platform" (e.g., aircrafts and satellites).

Characters of Real Remote sensing: Real remote sensing systems has following characters:

1. Energy Source: Energy sources for remote sensing are typically non-uniform over several wavelengths and are also differ with time and space. Energy source has major impact on the passive remote sensing. The spectral distribution of reflected sunlight varies both temporally and spatially. Materials on earth also emit different kind of energy. A remote sensing system requires calibration for source characteristics.
2. The Atmosphere: Atmosphere is also an important character of real remote sensing system. The atmosphere modifies the spectral distribution and potency of the energy received or emitted. The effect of atmospheric interaction varies with the wavelength associated, sensor used and the sensing application. Calibration is required to eliminate or compensate these atmospheric effects.
3. The Energy/Matter Interactions: Remote sensing is based on the principle that each and every material reflects or emits energy in a unique way. However, spectral signatures may be similar for different material types. This makes differentiation difficult. Also, the knowledge of most of the energy/matter interactions for earth surface features is either at elementary level or even completely unknown.
4. Sensor: Sensors have fixed limits of spectral sensitivity i.e.; they are not sensitive to all wavelengths. Also, they have limited spatial resolution. Selection of a sensor requires a trade-off between spatial resolution and spectral sensitivity. For example, while photographic systems have very good spatial resolution and poor spectral sensitivity, non-photographic systems have poor spatial resolution.
5. The Data Handling System: Human involvement is essential for processing sensor data; even though machines are also included in data handling. This makes the idea of real time data handling almost unfeasible. The amount of data generated by the sensors far exceeds the data handling capacity.
6. The Multiple Data Users: The success of any remote sensing mission relies on the user who ultimately transforms the data into information. This is possible only if the

user understands the problem thoroughly and has a wide knowledge in the data generation. The user should know how to interpret the data generated and should know how best to use them.

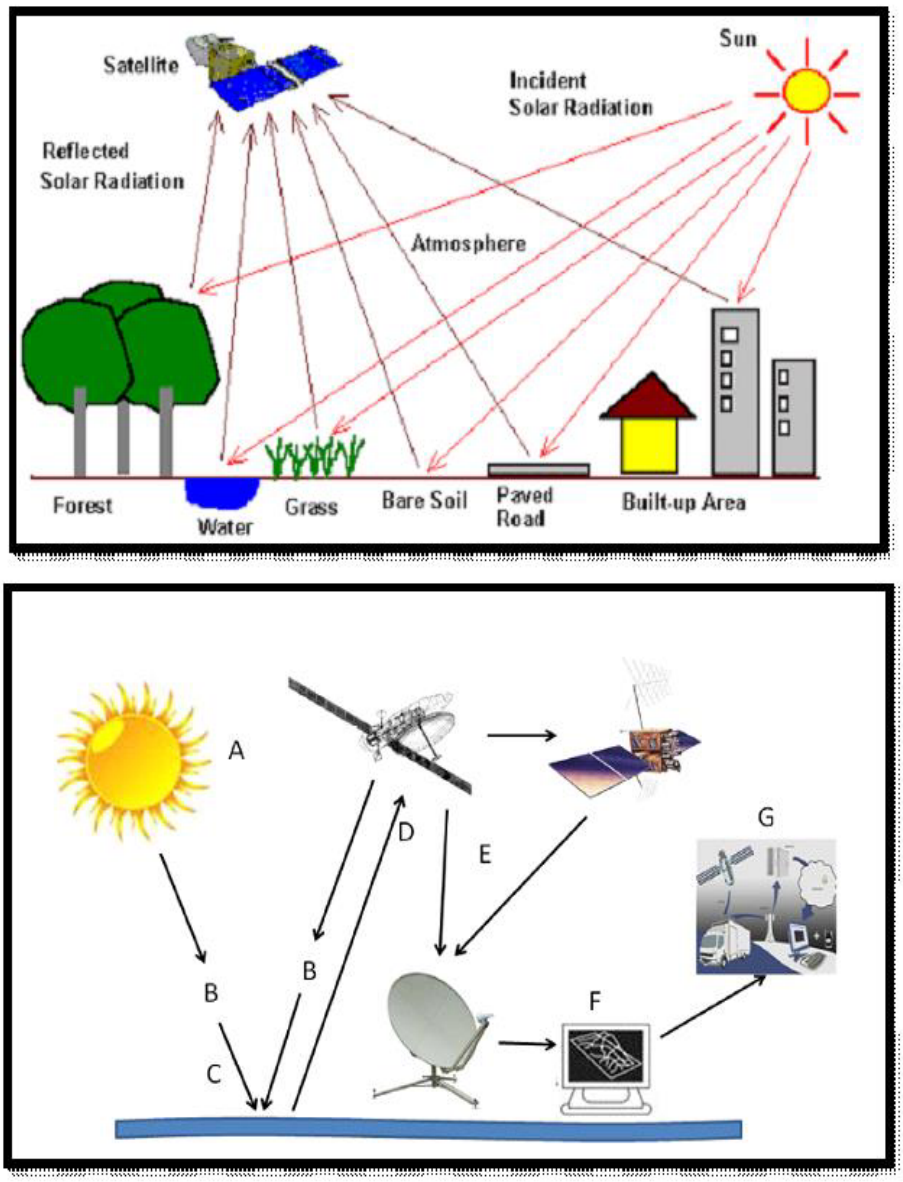


Fig- 1: Showing components of working procedure and components of Remote sensing (Source: http://maps.unomaha.edu/Peterson/gis/notes/RS2.htm)

[A - Energy source or illumination B – Radiation and atmosphere C – Interaction with the target D – Recording of the energy by the sensor E – Transmission, reception and processing F – Interpretation and analysis G – Application]

14.3. Advantages and Disadvantages of Remote Sensing

There are various advantages and disadvantages of remote sensing which are summarized in Fig-2 and also given below:

Advantages:

- It provides data of large areas
- It provides data of very remote and inaccessible regions
- Remote sensing is able to obtain imagery of any area over a continuous period of time through which the any anthropogenic or natural changes in the landscape can be analyzed.
- It is relatively inexpensive when compared to employing a team of surveyors
- It is easiest method of data collection.
- Remote sensing can produce large amount of maps for interpretation.

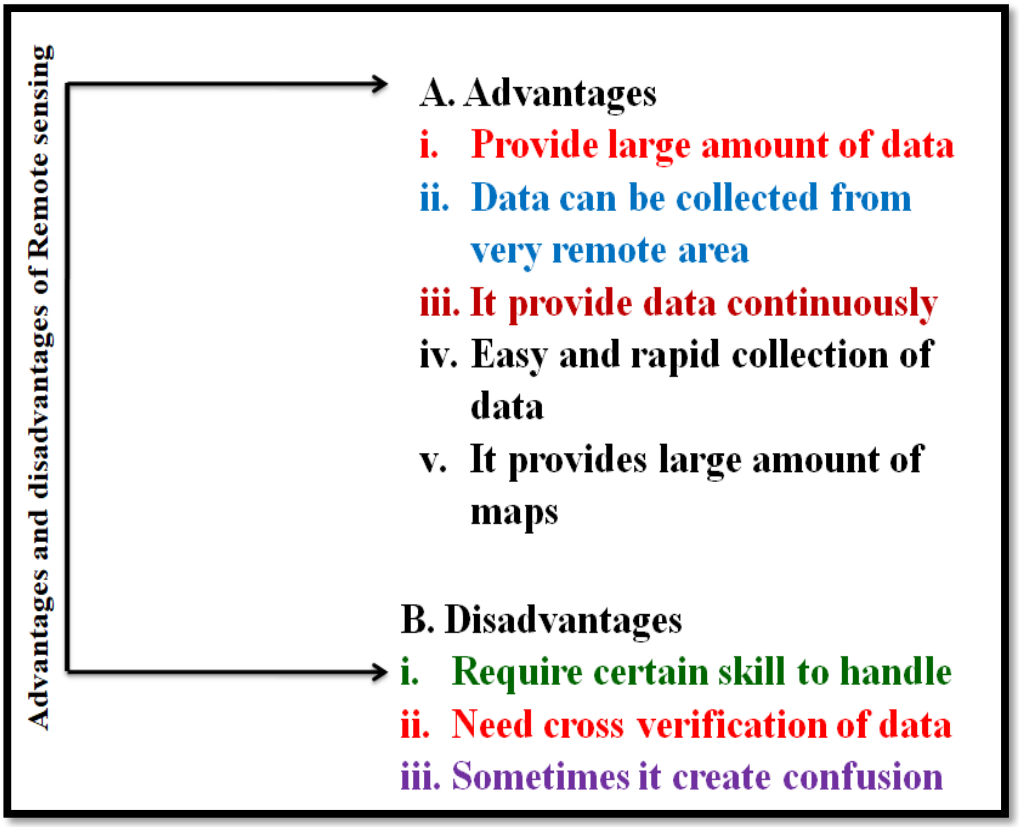


Fig-2: Showing advantages and disadvantages of remote sensing

Disadvantages:

- The interpretation of imagery requires a qualified manpower with required skill
- Remote sensing data needs cross verification with ground (field) survey data
- Data from multiple sources may create confusion
- Objects can be misclassified or confused
- Distortions may occur in an image due to the relative motion of sensor and source

14.4. Meaning and definition of GIS

GIS is a system design to capture, store, analyze and monitor the geographical data. GIS data includes imagery and base maps linked to spreadsheets and tables. GIS is a computer system build to capture geographical data. In simple words we can understand that GIS is an image that is referenced to the earth. GIS is well structured. There are two common data of GIS viz. Vector and Raster data. Vector data is classified in to three categories viz. Polygon, Line or arc and Point data. Polygon feature are most commonly distinguished using either a colour or pattern. Line data is used to represent linear characters such as rivers, streets and trains. Point data most commonly used to represent non-adjacent characters. Raster data is also known as grid data it is celled based and include satellite and aerial imagery. Raster data may be of two types viz. continuous and discrete data. Population density is an example of discrete raster data and temperature, elevation etc. are example of continuous data.

14.5. Advantages and disadvantages of GIS

There are various advantages and disadvantages of GIS which are summarized in Fig-3 and also discussed below:

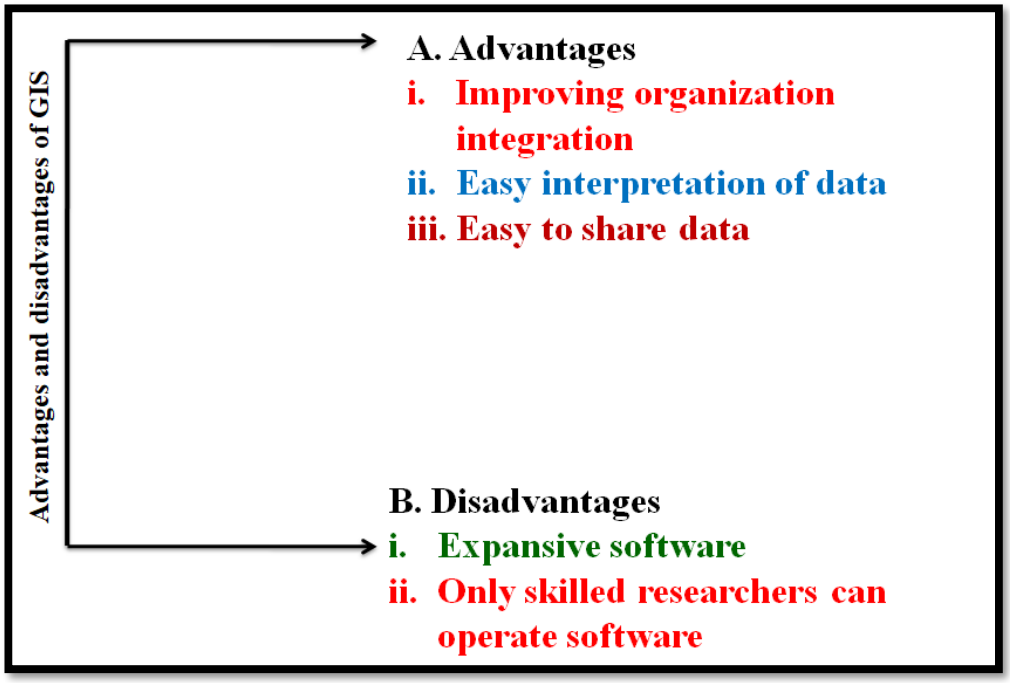


Fig-3: Showing advantages and disadvantages of GIS

Advantages:

- GIS has the ability of improving the organization integration. GIS would then integrate software, hardware and also data in order to gather, monitor and display all forms of facts being geographically referenced.
- GIS allow questioning, understanding and interpreting the data into numbers of ways which will observe correlation pattern in to form of globes, maps and reports.
- GIS is to provide help in answering queries as well as solve problems through looking at the data in way which is easily and quickly shared.

Disadvantages:

- This technology is costly.
- It requires large amount of data.
- Sometimes it may create confusion.

14.6. Application of Remote sensing in environmental studies

There are various applications of remote sensing in environmental studies. Some of the important applications are summarized in Fig-4 and also described below:

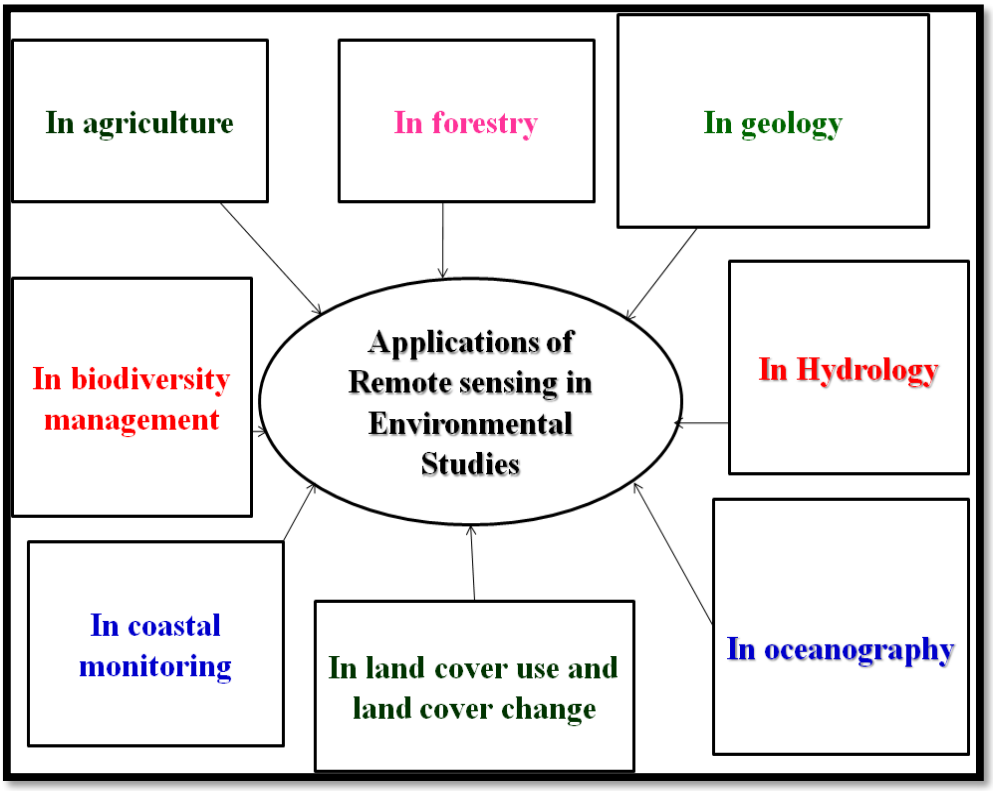


Fig-4: Showing applications of remote sensing in environmental studies.

In Agriculture: Remote sensing images are used as mapping tools to categorized crops, monitor health of plants, analyze the viability of plants, and observe farming practices. Remote sensing is important tool in agricultural applications include crop condition evaluation, crop yield assessment, mapping of soil.

In Forestry: It is well known that forests are a valuable resource. Forestry applications of remote sensing include the reconnaissance mapping, commercial forestry, harvest information, updating inventory information for timber supply, broad forest type, vegetation density, and biomass measurements. Remote sensing is used for monitoring the quantity, health, and diversity of the earth's forests.

In Geology: As you know geology entails the study of landforms, structures, and the subsurface to understand physical processes that create and modify the earth's crust. Geological applications of remote sensing include the bedrock mapping, lithological mapping, structural mapping, sand and gravel exploration/ exploitation, mineral exploration, hydrocarbon exploration, environmental geology, geo-botany, baseline infrastructure, sedimentation monitoring, event/monitoring, geo-hazard mapping, and planetary mapping.

Hydrology: Hydrology is the study of water and water bodies. Remote sensing play very important role in analysis of wetlands, soil moisture estimation, snow pack monitoring, measuring snow thickness, determining the snow-water equivalent, ice monitoring, flood monitoring, glacier dynamics monitoring (surges, ablation), river/delta change detection, drainage basin mapping, watershed modeling, irrigation canal leakage detection, and irrigation scheduling.

Land Cover and Land Use: Remote sensing used in natural resource management, wildlife habitat protection, baseline mapping for GIS input, urban expansion, to analyze seismic activities, damage description, legal boundaries of land, target detection, and identification of landing strips, roads, clearings, bridges, and land/water interface. Remote sensing can recognize the changes occurred in the land cover.

Mapping: As you know mapping constitutes an integral component of land resource management, with mapped information the common product of the analysis of remotely sensed data. Mapping applications of remote sensing include planimetry, which is land surveying techniques escorted by the use of a remote sensing can be used to meet high accuracy requirements, but limitations include cost effectiveness and difficulties in attempting to map large or remote areas. Remote sensing provides a means of identifying planimetric data in an efficient manner, so imagery is available in varying scales to meet the requirements of many different users.

In Oceans and Coastal Monitoring: The oceans provide valuable food-biophysical resources, serve as transportation routes, are crucially important in weather system formation and CO_2 storage, and are an important link in the earth's hydrological balance. Coastlines are environmentally sensitive interfaces between the ocean and land, and they

respond to changes brought about by economic development and changing land-use patterns. Often coastlines are also biologically diverse inter-tidal zones and can be highly urbanized. Ocean applications of remote sensing include the following:

14.7. Application of GIS in environmental studies

There are various applications of GIS in environmental studies. Some of the important applications are summarized in Fig-5 and also described below.

Environmental Impact Assessment (EIA): As you know EIA is an important policy initiative to protect natural resources and environmental components. Many anthropogenic activities create negative or harmful environmental effects which include the construction and operation of highways, railway tracks, pipelines, airports, radioactive waste disposal etc. Environmental impact assessment is generally required to contain specific information on the magnitude and characteristics of environmental impact. GIS plays important role in conducting the EIA.

Disaster Management: There are various types of natural disasters such as earthquake, flood, volcanoes, landslides etc. Disasters are not only harmful to human being but destroy the whole ecosystem. Today well-developed GIS systems are used to detect the natural disasters. It has become an integrated and well-developed tool in disaster management. GIS can also be important to mitigate the impacts of natural disasters. GIS can assist with risk management and analysis by exhibiting which areas are likely to be prone to disasters. When such disasters are recognized, preventive measures can be developed.

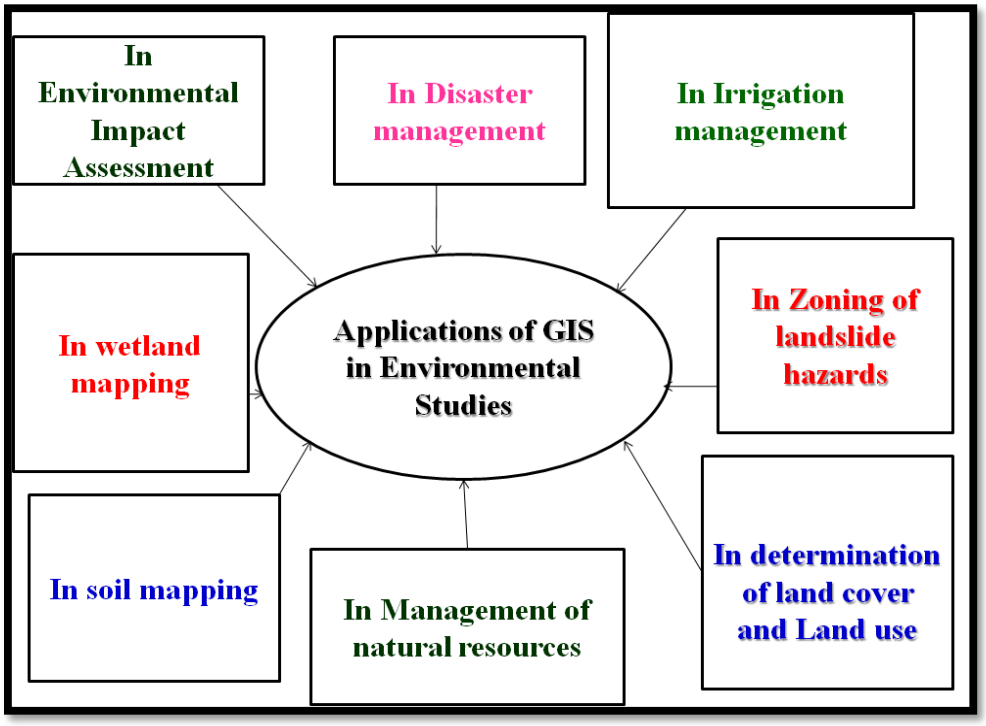


Fig-5: Showing different applications of GIS in environmental studies

Zoning of Landslides hazard: As you know that landslide hazard zonation is the process of ranking different parts of an area according to the degrees of actual or potential hazard from landslides. GIS used in the zonation of landslides hazards. The evaluation of landslide hazard is a difficult task. It has become possible to efficiently collect, manipulate and integrate a variety of spatial data such as geological, structural, surface cover and slope characteristics of an area, which can be used for hazard zonation.

Determination of land cover and land use: The actual meaning of land cover is the feature that covers the land. The land covers with the rocks, forests, grasses etc. On the other hand, land use means the land utilized for particular use. The role of GIS technology in land use and land cover applications is that we can determine land use/land cover changes in the different areas. GIS can detect and estimate the changes in the land use/ land cover pattern with time. It enables to find out sudden changes in land use and land cover either by natural forces or by manmade activities.

Estimation of flood damage: GIS helps to document the need for federal disaster relief funds, when appropriate and can be utilized by insurance agencies to assist in

assessing monetary value of property loss. A local government need to map flooding risk areas for evaluate the flood potential level in the surrounding area. The damage can be well estimated and can be shown using digital maps.

Management of Natural Resources: By the help of GIS technology the agricultural, water and forest resources can be well maintained and manage. Foresters can easily monitor forest condition. Agricultural land includes managing crop yield, monitoring crop rotation, and more. GIS can be used for monitoring and manage the agricultural resources. Water is one of the most essential constituents of the environment. GIS is used to analyze geographic distribution of water resources. GIS is also used to monitor and manage the forestation.

Soil Mapping: Soil mapping provides resource information about an area. It helps in understanding soil suitability for various land use activities. It is essential for preventing environmental deterioration associated with misuse of land. GIS helps to identify soil types in an area and to delineate soil boundaries. It is used for the identification and classification of soil. Soil map is widely used by the farmers in developed countries to retain soil nutrients and earn maximum yield.

Wetland Mapping: Wetlands contribute to a healthy environment and retain water during dry periods, thus keeping the water table high and relatively stable. During the flooding they act to reduce flood levels and to trap suspended solids and attached nutrients. GIS provide options for wetland mapping and design projects for wetland conservation quickly with the help of GIS. Integration with Remote Sensing data helps to complete wetland mapping on various scale. We can create a wetland digital data bank with spices information using GIS.

Irrigation management: Water availability for irrigation purposes for any area is vital for crop production in that region. It needs to be properly and efficiently managed for the proper utilization of water. GIS can be used to develop irrigation map.

Identification of Volcanic Hazard: Volcanic hazard to human life and environment include hot avalanches, hot particles gas clouds, lava flows and flooding. Potential volcanic hazard zone can be recognized by the characteristic historical records of volcanic activities, it can incorporate with GIS. Thus, an impact assessment study on volcanic

hazards deals with economic loss and loss of lives and property in densely populated areas.

14.8. Case study of Land use and Land cover change

Land cover refers to the physical characteristics of Earth's surface, captured in the distribution of vegetation, water, soil and other physical features. Land use refers to the way in which land has been used by humans and their habitats**.** Vegetation and structure that cover land area is known as land cover. Human activities change land cover, especially in urban area. These changes have environment and economic effect. The land cover may be classified into six categories Rangeland, forest land, cropland, parks and preserves, wetland mountain desert, and urban land and these land areas using by human being for grazing livestock, harvesting wood, for growing plants, recreation, preservation of native animal and plants and residence etc. respectively (Table 1).

Table-1: The land Cover type and human use of land

Sl. No.	Land cover type	Human use of land
1.	Rangeland	Grazing livestock
2.	Forest land	Harvesting wood, wild life, fishes and other resources
3.	Cropland	Growing plants for food and filter
4.	Parks and preserves	Recreation, preservation of native land and animal communication and ecosystem
5.	Wet lands mountain desert & others	Preservation of native animal and plant communication and ecosystem
6.	Urban land	Residence, other building and road

There are various reasons of land cover changes. These reasons may include industrialization, agriculture, urbanization, increase population etc. Land use is generally inferred based on the cover, yet both the terms land use and land cover being closely related and are used interchangeable. The surface temperatures are increasing globally as a consequence of anthropogenic climate change. However, it is known that observed changes are a result of both climate forcing and numerous other feedbacks including

LULC. The LULC could change as a response to climate and also act as a feedback. In addition to these natural forcing and feedback cycles, there are also additional aspects that are linked to anthropogenic activities. This results in further modification to the LULC and meteorological responses thereupon. These LULC changes and their effects are mostly discernible over regions having higher population density, industrialization, urbanization, deforestation, agricultural diversification etc. Thus, the most visible effect of anthropogenic activities regionally and locally are changes in the LULC which modifies the surface energy balance which in turn affects the surface temperature altering the region's micro-climate. The changes in LULC also modulate the incidence of heat/cold waves, clouds and rainfall patterns. In addition, LULC change have also been linked to atmospheric aerosol emissions which can modify the surface temperature through both direct and indirect effects, thereby modulating rainfall which can also result in droughts or floods through changes to extreme events under certain favorable circumstances.

Over the Indian region, there are only a few scientific investigations that have attempted to discern LULC induced temperature changes, but they are either limited to the major metropolitan cities or have only focused on aspects related to urbanization. For example, the surface temperature over western India is found to be warming by ~0.13 °C/decade due to the combined effect of greenhouse gases and LULC change of which ~50% was attributed to LULC change. Also, in 2001 an area covering 26.4% of New Delhi had a diurnal temperature range (DTR) below 11 °C whereas in 2011 65.3% of New Delhi had a DTR below 11 °C which was attributed to the increase in built up area by 53%. Furthermore, the LULC has also been linked to Indian monsoon rainfall changes. Studies linking LULC to surface temperature changes are limited over Eastern India though this region is among the most rapidly changing landscape over the entire Indian region. The region is also rich in mineral deposits. Exploitation for mineral wealth has accelerated LULC change in the past few decades. In addition, Odisha being one of the most natural disaster-prone regions of India, a very few studies have investigated the relationship between LULC change and surface temperature, heat waves, extreme rainfall etc.

14.9. Urban sprawling

Urban sprawling is unrestricted growth in urban area. This growth may be for construction of houses, roads etc. Urban sprawling is part of urbanization. As you know that urban area is that area where all the commodities are easily available and easily accessible. Population growth in urban area may lead into urban sprawling. Urban sprawling has various consequences on environment. There are various causes of urban sprawling which are given here:

1. Advance infrastructure: There is increased spending on certain kinds of infrastructure such as roads, electrical supplies etc. Some urban area doesn't have these infrastructures and continuously working for the development.
2. Rise in standard of living: As you know that standards of living also high in urban area and it continuously increasing day by day. High living standards also lead into urban sprawling.
3. High population growth: Rise in population growth is a factor which is responsible for urban sprawling. High population needs more and more area for living. As number of cities grows in urban area the local community also spread further and further from the Centre of cities. This condition also leads into encroachment.

There are various consequences of urban sprawling which are summarized in Fig-6 and also discussed below:

1. Increase in Public Expenditure: Public expenditure also increases due to urban sprawling. Urban sprawling can increase the expanses of people. It is because urban sprawling happened by the money of tax payers.
2. Consumption of high energy: As you know demand of energy increasing day by day. We use energy in every sector. The consumption of energy is higher in the urban area as compared to rural area. Urban sprawling increases demand of energy. In urban zone we need energy for operating many types of equipment such as washing machine, transportation, refrigerators, air conditioning etc. Urban sprawling changes the life style of people. To maintain the urban life style, people use more and more energy.

3. Generation of solid waste: Solid waste is any garbage or waste which is not in use in present time. Urban sprawling leads into generating high amount of solid waste. As you know solid waste can cause, air, water and soil pollution.

4. Environmental Pollution: Urban sprawling causes environmental pollution. As you know environmental pollution is any undesirable changes in physical chemical and biological properties of environment. Environmental pollution may be air, water, soil, noise and thermal pollution. Urban sprawling changes the physical, chemical and biological properties of air, water and soil and consequently leads into environmental pollution. Urban sprawling reduces the number of trees which cause air pollution.

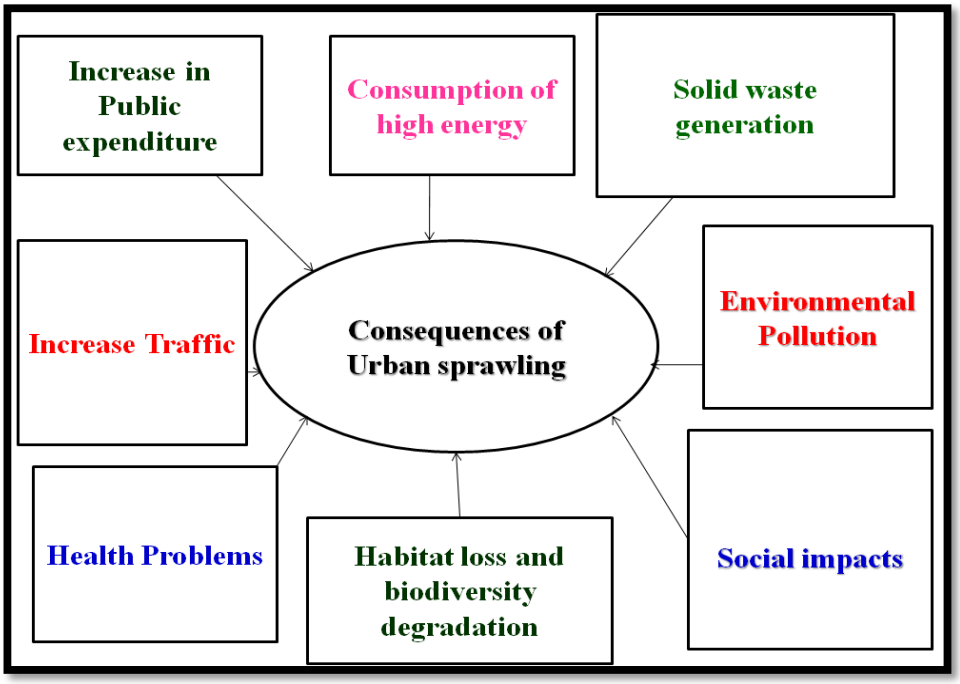


Fig- 6 Showing impacts of urban sprawling

1. Habitat loss and biodiversity degradation: Urban sprawling may lead into habitat fragmentation which is one of the major threats to biodiversity. As you know many species live on the trees and plants. Urban sprawling certainly leads into deforestation which consequently declines the population of many species. Researchers also found that population of house sparrow (*Passer domesticus*) have been declined due to urban sprawling.

2. Health Problems: Health problem also increases due to urban sprawling. Many diseases such as cardiovascular diseases, neurological diseases, psychological disorders, nausea, skin problem, respiratory diseases etc. are also increases due to urban sprawling.
3. Increase Traffic: Urban sprawling also may lead into traffic in the cities. Due to urban sprawling people buy more and more vehicle to fulfill the requirements. Thus, traffic load also increases.
4. Social impacts: When people come in the contact of cities, they do not have neighbors therefore, social life of the people changed due to urban sprawl. It is also noticed that, traditional values are also declining due to urban sprawling.

14.10. Mining hazards

Mining hazard is any of the danger to the people due to exposure of coal and other minerals. As you know mining is extraction of minerals from the earth. These minerals may include metals, coal, oil, shale, rock salt, gravel, clay, chalk etc. The extraction of minerals is possible through mining. Mining is vitally important because these mined minerals can use for various purposes. Mining provides various minerals to society but unfortunately mining has some negative impacts on environment. There are various impacts of mining on environment. Some of the important impacts of mining on environment are summarized in Fig-7 and also discussed below:

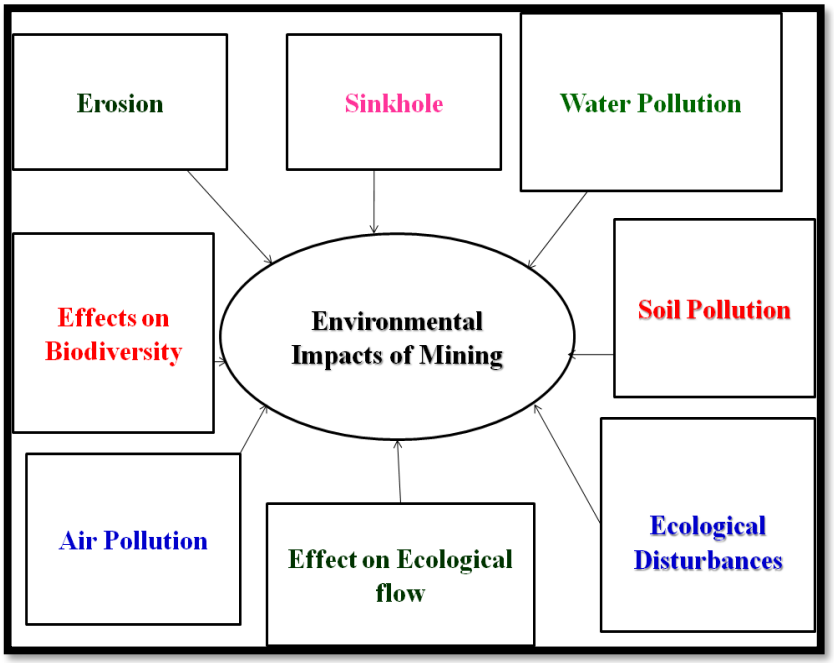


Fig-7: Showing negative impacts of mining on Environment

1. Erosion: Mining may cause soil erosion. Hillsides, mine dumps, can be exposed due to mining. It is also noticed that huge number of debris spilled out from the river which cause the soil erosion specially in the riparian or peripheral zone of the river.
2. Sinkhole: Sinkholes are also one of the major impacts of mining. These sink holes formed during mining due to extraction of minerals, weak overburden or may be due to geological discontinuities.
3. Water Pollution: Mining can pollute the water resource significantly. There are various harmful chemicals, which are also deposited in the benthic zone of river. When mining takes place in river, these chemicals spilled out in water and cause water pollution. These harmful chemicals may include mercury, arsenic, cadmium etc.
4. Effect on Biodiversity: As you know that rivers are home to many species. Various species of fishes, amphibians, reptiles, birds and mammals are live near or within aquatic bodies. Mining certainly degrades the quality of water, habitat loss which

consequently decline the population of plants and animals in water bodies. Therefore, mining is certainly responsible for biodiversity degradation.

5. Soil Pollution: Soil pollution is another impact of mining. Sometimes the dredged material deposited near the aquatic body where it causes soil pollution. Mining also reduces the quality of soil. As you know mining is all about the extraction of essential minerals which reduces the quality of soil. After the mining soil becomes less productive.
6. Air Pollution: Mining operations also may lead into air pollution. Many air pollutants can mix with air during mining and create atmospheric pollution. Mining increases the concentration of SPM and RSPM in the air.
7. Ecological Disturbances: Mining also may lead into ecological disturbances. Mining may change the food chain in aquatic ecosystem. As earlier mentioned, that mining reduces the biodiversity of aquatic ecosystem and we know that every species has its own role in ecosystem. Species of ecosystem depend on other species form their food, shelter etc. Therefore, mining certainly responsible for ecological disturbance.

14.11. SUMMARY

- In this unit we have discussed various aspects of remote sensing, GIS, Land cover, Land use and mining hazards. So far you have learnt that:
- The word “remote” means far away. The instrument or sensor which can sense the obejct from the distance and without physical contact is known as remote sensing. Remote sensing have been popular in last few years. This technique can be used in different sciences such as earth science, geology, geography, hydrobiology, oceanography, glaciology, etc. In many contries, remote sensing is also used for security purposes.
- There are various advantages and disadvantages of remote sensing which are: it provides data of large area, it provides data of very remote and inaccessible regions, it is able to obtain imagery of any area over a continuous period of time through which the any anthropogenic or natural changes in the landscape can be

analyzed, it is relatively inexpensive when compared to employing a team of surveyors

- Disadvantages of Remote sensing are as it requires a certain skill, it requires cross verification, data sources may create confusion.
- GIS is a system design to capture, store, analyze and monitor the geographical data. In this technique we collect the data from the earth. GIS data includes imagery and base maps linked to spreadsheets and tables.
- GIS has the ability of improving the organization integration. GIS would then integrate software, hardware and also data in order to gather, monitor and display all forms of facts being geographically referenced.
- GIS allow questioning, understanding and interpreting the data into numbers of ways which will observe correlation pattern in to form of globes, maps and reports.
- GIS is to provide help in answering queries as well as solve problems through looking at the data in way which is easily and quickly shared.
- There are various applications of remote sensing in environmental studies. In Agriculture, In Forestry, In Geology, Hydrology, Land Cover and Land Use, Mapping, In Oceans and Coastal monitoring remote sensing used.
- There are various applications of GIS in environmental studies. Some of the important applications are Environmental Impact Assessment (EIA, Disaster Management, Zoning of Landslides hazard, Determination of land cover and land use, Estimation of flood damage, Management of Natural Resources, Soil Mapping, Wetland Mapping, Identification of Volcanic Hazard.
- Land cover refers to the physical characteristics of Earth's surface, captured in the distribution of vegetation, water, soil and other physical features. Land use refers to the way in which land has been used by humans and their habitats. Vegetation and structure that cover land area is known as land cover.
- Urban sprawling is unrestricted growth in urban area. This growth may be for construction of houses, roads etc. Urban sprawling is part of urbanization.

- There are various causes of urban sprawling such as advance infrastructure, rise in standard of living, high population growth.
- There are various consequences of urban sprawling which are increase in Public Expenditure, Consumption of high energy, Generation of solid waste, Environmental Pollution, Habitat loss and biodiversity degradation, Health Problems, Increase Traffic etc.
- Mining hazard is any of the danger to the people due to exposure of coal and other minerals. As you know mining is extraction of minerals from the earth. These minerals may include metals, coal, oil, shale, rock salt, gravel, clay, chalk etc. The extraction of minerals is certainly required because we cannot grow these minerals through agricultural processes.
- Mining is vitally important because these mined minerals can use for various purposes. Mining provides various minerals to society but unfortunately mining has some negative impacts on environment. There are various impacts of mining on environment.
- Erosion, Sinkhole, Water Pollution, Effect on Biodiversity, Soil Pollution, Air Pollution, Ecological disturbances etc. are important impacts of mining.

TERMINAL QUESTIONS

1. (a) Fill in the blank spaces with appropriate words.

Remote……. is a technique of science of obtaining ……. about an object without……..contact with that object. Men apply remote sensing in their day-to-day business, through vision, hearing and sense of smell. The data collected can be of many forms: variations in acoustic wave distributions, variations in force distributions, variations in electromagnetic energy distributions etc. These remotely collected data through various …….. may be analyzed to obtain information about the …….or features under…….. Thus, remote sensing is the process of inferring surface parameters from measurements of the …… …….. (EMR) from the Earth's surface. This EMR can either be reflected or emitted from the Earth's surface. In other words, remote sensing is detecting and measuring electromagnetic (EM) energy emanating or reflected from distant objects

made of various materials, so that we can identify and categorize these objects by class or type, substance and spatial……

2. (a) Define remote sensing.

(b) Explain GIS.

3. (a) Describe the applications of Remote sensing in environmental studies.

(b)Describe the applications of GIS in environmental studies.

4. (a) What is urban sprawling?

5. (a)What is mining? Explain impacts of mining.

6. (a) Fill the blank spaces with appropriate words.

……is a system design to capture, store, ……. .and monitor the ……..data. In this technique we collect the data from the…….. GIS data includes imagery and ……..linked to spreadsheets and tables. GIS is a computer ……..build to capture geographical data. In simple words we can understand that GIS is an image that is referenced to the…….. GIS is well structured. The use of GIS is now widespread in ……science. There are two common data of GIS viz. …….. and ……data. Vector data is classified in to three categories viz……., Line or arc and Point data. ……. feature are most commonly distinguished using either a colour or pattern. Line data is used to represent linear characters such as rivers, streets and trains. Point data most commonly used to represent non-adjacent characters. Raster data is also known as …… …….it is celled based and include satellite and aerial imagery. Raster data may be of two types viz. continuous and ……data. Population density is an example of …….. raster data and temperature, elevation etc. are example of continuous data.

(b) Remote sensing uses which of the following waves in the procedure? (Electric field/electromagnetic waves/Sonar waves/Gamma rays)

(c) In …….. remote sensing, source of energy is that naturally available such as the Sun (Passive/Active)

(d) GIS is stand for (Geological informative system/Geographic information system/Global information system/Geometric information system)

(e) What are advantages and disadvantages of Remote sensing?

7. (a) What are advantages and disadvantages of GIS?

ANSWERS

1. (a) sensing, information, physical, sensors, objects, observation, electromagnetic radiation, distribution

2. (a) see the section 14.2

(b) See the section 14.4

3. (a) See the section 14.6

(b) See the section14.7

4. (a) See the section 14.9

5. (a) See the section 14.10

6. (a) GIS, analyze, geographical, earth, base maps, system, earth, environmental, Vector, Raster, Polygon, Polygon, grid data, discrete, discrete

(b) Electromagnetic waves

(c) Passive

(d) All of the above

(e) See the section 14.3

7. (a) See the section 14.5