Uttar Pradesh Rajarshi Tandon
Open University

BBA-121

Research Methodology

NODE-110/NODE-124/NFT-063

Uttar Pradesh Rajarshi Tandon
Open University

BBA-121

Research Methodology

BLOCK

1

FUNDAMENTALS OF RESEARCH

UNIT-1

INTRODUCTION

UNIT-2

RESEARCH PROBLEMS

UNIT-3

RESEARCH DESIGN

UNIT-4

DATA COLLECTION

परिशिष्ट–4

आन्तरिक कवर–दो का प्ररूप

Format of the II Inner Covers

लेखक Dr. Piyali Ghosh, Asst. Professor, School of Management, MNNIT, Allahabad

सम्पादक Prof. H.K. Singh, Professor, Department of Commerce, BHU Varansi.

परिमापक

सहयोगी टीम

संयोजक Dr. Gaurav Sankalp, SoMS, UPRTOU, Allahabd.

प्रूफ रीडर

©UPRTOU, Prayagraj-2020
ISBN : 978-93-83328-54-3

NODE-110/NODE-124/NFT-063

BLOCK INTRODUCTION

In **Block-1** you learnt about the fundamentals of research including the topics like identifying research problems, preparing research design, data collection etc.

Unit-1 discusses about meaning, significance of research, types of research, difference between research methods & research methodology and topics used in business management.

Unit-2 explains meaning and components of research problem, sources of identifying research problem and ethical issues relating to business research.

Unit-3 deals with concept of research design, features of research design, classification of research design and factors affecting research design.

Unit-4 deals with difference between primary data and secondary data, methods of collecting primary data, methods of collecting secondary data, questionnaire and schedule construction, basic rules for questionnaire item construction and limitations of secondary data.

UNIT-1 INTRODUCTION

Unit Framework

1.1 OBJECTIVE

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

- Meaning, Objectives and Types of Research
- Qualities of Researcher
- Significance of Research
- Empirical Research
- Research Methods vs Research Methodology
- Topics Used in Business Management

1.2 INTRODUCTION : MEANING AND DEFINITION

The word research is composed of two syllables, re and search re is a prefix meaning again, anew or over again search is a verb meaning to examine closely and carefully, to test and try, or to probe. Together they form a noun describing a careful, systematic, patient study and

investigation in some field of knowledge, undertaken to establish facts or principles.

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

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

According to **Redman and Mory** (1923), research is a *"systematized effort to gain new knowledge"*. It is an academic activity and therefore the term should be used in a technical sense.

According to **Clifford Woody** (kothari, 1988), research comprises "*defining and redefining problems, formulating hypotheses or suggested solutions; collecting, organizing and evaluating data; making deductions and reaching conclusions; and finally, carefully testing the conclusions to determine whether they fit the formulated hypotheses"*.

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

1.3 CHARACTERISTICS OF RESEARCH

It is clear that research is an unbiased, structured, systematic and sequential method of enquiry directed towards a clear or implicit objective. But to qualify as research it must contain various characteristics. Henry Johnson mentions that research has following main characteristics:

1. It is empirical, i.e., it is based on observation and reasoning and not on speculation.
2. It is theoretical, i.e., it summarises data precisely giving logical relationship between propositions which explain causal relationship.
3. It is cumulative, i.e., generalizations / theories arc corrected, rejected and newly developed theories are built upon one another.
4. It is non-ethical, i.e., scientists do not say whether particular things/events/ phenomena/ institutions /systems/structures are good or bad. They only explain them.

A good research study contains the following features:

- **Objectivity:** The main purpose of a research is to answer the research question. Research should be objective which helps in necessitating the formulation of a proper hypothesis. Lack of objectivity leads to a poor formulation of hypothesis and the entire process thereafter lacks any congruency between the research questions and the hypothesis.

- **Control:** A good research must be able to control all the variables. This requires randomization at all stages, e.g., in selecting the subjects, the sample size and the experimental treatments. This shall ensure an adequate control over the independent variables. This is the basic technique in all scientific experimentation—allowing one variable to vary while holding all other variables constant. Unless all variables except one have been controlled, one cannot be sure which variable has produced the results.

- **Rigorous**: you must be scrupulous in ensuring that the procedures followed to find answers to questions are relevant, appropriate and justified. Again, the degree of rigor varies markedly between the physical and social sciences and within the social sciences.

- **Systematic**: This implies that the procedure adopted to undertake an investigation follow a certain logical sequence. The different steps cannot be taken in a haphazard way. Some procedures must follow others.

- **Valid and verifiable:** This concept implies that whatever you conclude on the basis of your findings is correct and can be verified by you and others.

- **Empirical: T**his means that any conclusions drawn are based upon hard evidence gathered from information collected from real life experiences or observations.

- **Critical**: Critical scrutiny of the procedures used and the methods employed is crucial to a research enquiry. The process of investigation must be foolproof and free from drawbacks. The process adopted and the procedures used must be able to withstand critical scrutiny. For a process to be called research, it is imperative that it has the above characteristics.

- **Precision:** A research should never be ambiguous. One has to make it as exact as necessary. The facts and figures should be exact to the extent possible. Precision does not restrict to just the data or the facts and findings, but also extends to the measurement factor too.

1.4 SIGNIFICANCE OF RESEARCH

- **To Gather Necessary Information:** Research provides you with all necessary information in field of your work, study or operation

before you begin working on it. For example, most companies do research before beginning a project in order to get a basic idea about the things they will need to do for the project. Research also helps them get acquainted with the processes and resources involved and reception from the market. This information helps in the successful outcome of the project.

- **To Make Changes:** Sometimes, there are in-built problems in a process or a project that is hard to discover. Research helps us find the root cause and associated elements of a process. The end result of such a research invokes a demand for change and sometimes is successful in producing changes as well. For example, many U.N. researches have paved way for changes in environmental policies.

- **Improving Standard of Living:** Only through research can new inventions and discoveries come into life. It was C.V Raman's research that prompted invention of radio communication. Imagine how you would have communicated had Graham Bell not come out with the first ever practical telephone! Forget telephones, what would have happened if Martin Cooper did not present the world the concept of mobile phones! Addicted as we are to mobile phones, we need to understand that all the luxuries and the amenities that are now available to us are the result of research done by someone. And with the world facing more and crisis each day, we need researchers to find new solutions to tackle them.

- **For A Safer Life:** Research has made ground breaking discoveries and development in the field of health, nutrition, food technology and medicine. These things have improved the life expectancy and health conditions of human race in all parts of the world and helped eradicate diseases like polio, smallpox completely. Diseases that were untreatable are now history, as new and new inventions and research in the field of medicine have led to the advent of drugs that not only treat the once-incurable diseases, but also prevent them from recurring.

- **To Know the Truth:** It has been proved time and again that many of established facts and known truths are just cover ups or blatant lies or rumors. Research is needed to investigate and expose these and bring out the truth.

- **Explore Our History:** Research about our planets history and human history has enabled us to learn and understand more about our forefathers and helped us learn from their mistakes and absorb good things from their life. Research about the planet's history and existence has told us a lot about how things will shape up in years to come and how we need to respect our planet and work closely together to stop global warming and other scenarios of destruction.

- **Understanding Arts:** This helps us in understanding the work of artists in literature, paintings, sculptures and everything that can be

attributed with artistic touch. If no research is conducted into any of these, we will never be able to understand any of these as per the artist's imagination. Also, a lot of great artistic work is hidden in the shadows of history, which needs to be drawn out.

- **Motives to do Research:** Research is the result of advancing knowledge created in the past. There are people from all walks of life that contribute to gathered information. These are ordinary people and extraordinary people. They include teachers, students, scientists, professors, scholars, business owners, librarians, book keepers, writers, politicians and many more unknown out there. These are everyday citizens we interact with. They all help with the flow information that people use for self-help.

1.5 TYPES OF RESEARCH

Research can be classified from three perspectives:

1. Application of research study
2. Objectives in undertaking the research
3. Inquiry mode employed

1. APPLICATION :

From the point of view of application, there are two broad categories of research:

- Pure research and
- Applied research.
- **Pure research** involves developing and testing theories and hypotheses that are intellectually challenging to the researcher but may or may not have practical application at the present time or in the future. The knowledge produced through pure research is sought in order to add to the existing body of research methods.
- **Applied research** is done to solve specific, practical questions; for policy formulation, administration and understanding of a phenomenon. It can be exploratory, but is usually descriptive. It is almost always done on the basis of basic research. Applied research can be carried out by academic or industrial institutions. Often, an academic institution such as a university will have a specific applied research program funded by an industrial partner interested in that program.

2. OBJECTIVES :

From the viewpoint of objectives, a research can be classified as –

- **Descriptive** research attempts to describe systematically a situation, problem, phenomenon, service or programme, or provides information about , say, living condition of a community, or describes attitudes towards an issue.

- **Correlation** research attempts to discover or establish the existence of a relationship/ interdependence between two or more aspects of a situation.

- **Explanatory** research attempts to clarify why and how there is a relationship between two or more aspects of a situation or phenomenon.

- **Exploratory** research is undertaken to explore an area where little is known or to investigate the possibilities of undertaking a particular research study (feasibility study / pilot study). In practice most studies are a combination of the first three categories.

3. INQUIRY MODE :

From the process adopted to find answer to research questions the two approaches are:

Structured approach

Unstructured approach

- **Structured approach:** The structured approach to inquiry is usually classified as quantitative research. Here everything that forms the research process- objectives, design, sample, and the questions that you plan to ask of respondents- is predetermined.

  It is more appropriate to determine the extent of a problem, issue or phenomenon by quantifying the variation e.g. how many people have a particular problem? How many people hold a particular attitude?

- **Unstructured approach:** The unstructured approach to inquiry is usually classified as qualitative research. This approach allows flexibility in all aspects of the research process. It is more appropriate to explore the nature of a problem, issue or phenomenon without quantifying it.

  Main objective is to describe the variation in a phenomenon, situation or attitude. e,g., description of an observed situation, the historical enumeration of events, an account of different opinions different people have about an issue, description of working condition in a particular industry.

  Both approaches have their place in research. Both have their strengths and weaknesses. In many studies you have to

combine both qualitative and quantitative approaches. For example, suppose you have to find the types of cuisine accommodation available in a city and the extent of their popularity.

1.6 PROBLEMS AND PRECAUTIONS IN EFFECTIVE RESEARCH

Business Research in India is facing a lot of difficulties especially in case of Exploratory Research, it may be of varied reasons, and some are discussed herein under:

(a) It is important to be aware that exploratory research should never take the place of quantitative research. Doing so, 'could *Introduction to Research* lead to misinterpretations and poor judgments.

(b) The greatest concern, however, is that of rejecting a good idea during the initial stages of exploratory research, thus voiding it from being analyzed and targeted correctly.

(c) As secondary data has been collected for purposes other than those outlined in the research study, its usefulness may be restricted in a few ways.

(d) Biased research transpires when the research process IS executed improperly, resulting in incorrect findings.

(e) The Researchers are not scientifically trained. So they generally make error while selecting the right Research method.

(f) Choosing of population for survey is difficult. For example, if you want to make survey of consumer goods in the urban areas it will give a different result than if you make the survey in the rural areas.

(g) Due to collection of large data, sometimes the researcher is unable to make correct interpretation leading to false result calculation.

(h) As the decision is based on the paradigm of the researcher, so sometimes it is biased and incorrect.

(i) While making a Business Research, the secrecy of the business is generally leaked to its competitors. So, in such case the data are not provided to the researcher.

(j) Research is time taking process. It may require time which may span over many years. Also in the beginning of the Research, the researcher is not sure, how much result will he get at the end of the research.

(k) Since Research is such a long process. So, it requires lot of funds, and that fund has no return unless the research work is complete.

(l) Many Social Research face stiff objection of the society, if it is about the changing trend in the society. In such case, it becomes very difficult for the researcher to collect data and interpret it. (m) Researchers group in India generally face the problem of discipline, where due to large volume of data, a researcher may predict the data to be collected.

This leads to a bias data collection. Data should be collected initially and at the time of making the decisions or conclusion, interpretations should be made.

1.7 EMPIRICAL RESEARCH

Empirical research is a way of gaining knowledge by means of direct and indirect observation or experience. Empirical evidence (the record of one's direct observations or experiences) can be analyzed quantitatively or qualitatively.

Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions, which should be clearly defined and answerable with the evidence collected (usually called data). Research design varies by field and by the question being investigated. Many researchers combine qualitative and quantitative forms of analysis to better answer questions which cannot be studied in laboratory settings, particularly in the social sciences and in education.

In some fields, quantitative research may begin with a research question (e.g., "Does listening to vocal music during the learning of a word list have an effect on later memory for these words?") which is tested through experimentation in a lab.

Usually, a researcher has a certain theory regarding the topic under investigation. Based on this theory some statements, or hypotheses, will be proposed (e.g., "Listening to vocal music has a negative effect on learning a word list."). From these hypotheses predictions about specific events are derived (e.g., "People who study a word list while listening to vocal music will remember fewer words on a later memory test than people who study a word list in silence."). These predictions can then be tested with a suitable experiment. Depending on the outcomes of the experiment, the theory on which the hypotheses and predictions were based will be supported or not.

STEPS IN EMPIRICAL RESEARCH

The ideal research proposal should be comprehensive enough to enable the reader to know everything that could be expected to happen if the project were actually carried out including anticipated obstacles as well as anticipated benefits. In order to design a research project, you may wish to ask yourself the following series of questions:

1. **PROBLEM STATEMENT, PURPOSES, and BENEFITS.**

- What exactly do I want to find out?
- What is a researchable problem?
- What are the obstacles in terms of knowledge, data availability, time, or resources?
- Do the benefits outweigh the costs?

2. THEORY, ASSUMPTIONS, and BACKGROUND LITERATURE

- What does the relevant literature in the field indicate about this problem?
- To which theory or conceptual framework can I link it?
- What are the criticisms of this approach, or how does it constrain the research process?
- What do I know for certain about this area?
- What is the history of this problem that others need to know?

3. VARIABLES AND HYPOTHESES

- What will I take as given in the environment?
- Which are the independent and which are the dependent variables?
- Are there control variables?
- Is the hypothesis specific enough to be researchable yet still meaningful?
- How certain am I of the relationship(s) between variables?

4. OPERATIONAL DEFINITIONS AND MEASUREMENT

- What is the level of aggregation?
- What is the unit of measurement?
- How will the research variables be measured?
- What degree of error in the findings is tolerable?
- Will other people agree with my choice of measurement operations?

5. RESEARCH DESIGN AND METHODOLOGY

- What is my overall strategy for doing this research?
- Will this design permit me to answer the research question?

- What other possible causes of the relationship between the variables will be controlled for by this design?
- What are the threats to internal and external validity?

6. SAMPLING

- How will I choose my sample of persons or events?
- Am I interested in representativeness?
- If so, of whom or what, and with what degree of accuracy or level of confidence?

7. INSTRUMENTATION

- How will I get the data I need to test my hypothesis?
- What tools or devices will I use to make or record observations?
- Are valid and reliable instruments available, or must I construct my own?

8. DATA COLLECTION AND ETHICAL CONSIDERATIONS

- Are there multiple groups, time periods, instruments, or situations that will need to be coordinated as steps in the data collection process?
- Will interviewers, observers, or analysts need to be trained?
- What level of inter-rater reliability will I accept?
- Do multiple translations pose a potential problem?
- Can the data be collected and subjects' rights still preserved?

9. DATA ANALYSIS

- What combinations of analytical and statistical process will be applied to the data?
- Which will allow me to accept or reject my hypotheses?
- Do the findings show numerical differences, and are those differences important?

10. CONCLUSIONS, INTERPRETATIONS, RECOMMENDATIONS

- Was my initial hypothesis supported?
- What if my findings are negative?
- What are the implications of my findings for the theory base, for the background assumptions, or relevant literature?

- What recommendations can I make for public policies or programs in this area?
- What suggestions can I make for further research on this topic?

1.8 TOPICS USED IN BUSINESS MANAGEMENT

- ❖ Money Management: How to Save Money?
- ❖ Recession: Opportunities in Recession
- ❖ Stress Management: Stress Freedom
- ❖ The Genuine Professional
- ❖ Life Management: Success in Life
- ❖ Self-awareness
- ❖ Management in Life, Profession, Family and Society
- ❖ Recession
- ❖ Competency Matrix/Competency Mapping
- ❖ Training of Trainers
- ❖ Human Resource Management (HRM)
- ❖ Training Games
- ❖ Outsourcing
- ❖ Performance Management
- ❖ Impression Management
- ❖ Johari Window
- ❖ ABC Analysis or Pareto Analysis
- ❖ Downsizing
- ❖ Everything You Wanted to Know on Leadership and Management
- ❖ Democracy
- ❖ Efficient Work Methods
- ❖ Management Case Studies
- ❖ Basic Statistics
- ❖ Business Plan
- ❖ Brainstorming

- Public Speaking
- Body Language
- Self-Motivation
- Self-concept
- Financial Services
- Financial Ratios and Financial Ratio Analysis
- Human Software (H Software)
- Management Anecdotes
- Entrepreneurship
- Career Planning within Organizations
- Group Discussion
- Positive Strokes
- Life Positions and OKness
- Safety and Health Management
- Human Relations
- Corporate Governance
- Group Dynamics
- Organizational Culture
- Recession Management
- Email Etiquette
- Supply and Demand
- Management Universe

1.9 RESEARCH METHODS VS RESEARCH METHODOLOGY

Research Methods and Research Methodology are two terms that are often confused as one and the same. Strictly speaking they are not so and they show differences between them. One of the primary differences between them is that research methods are the methods by which you conduct research into a subject or a topic. On the other hand research methodology explains the methods by which you may proceed with your research.

Research methods involve conduct of experiments, tests, surveys and the like. On the other hand research methodology involves the

learning of the various techniques that can be used in the conduct of research and in the conduct of tests, experiments, surveys and critical studies. This is the technical difference between the two terms, namely, research methods and research methodology.

In short it can be said that research methods aim at finding solutions to research problems. On the other hand research methodology aims at the employment of the correct procedures to find out solutions. It is thus interesting to note that research methodology paves the way for research methods to be conducted properly. Research methodology is the beginning whereas research methods are the end of any scientific or non-scientific research.

Let us take for example a subject or a topic, namely, 'employment of figures of speech in English literature'. In this topic if we are to conduct research, then the research methods that are involved are study of various works of the different poets and the understanding of the employment of figures of speech in their works. On the other hand research methodology pertaining to the topic mentioned above involves the study about the tools of research, collation of various manuscripts related to the topic, techniques involved in the critical edition of these manuscripts and the like.

If the subject into which you conduct a research is a scientific subject or topic then the research methods include experiments, tests, study of various other results of different experiments performed earlier in relation to the topic or the subject and the like. On the other hand research methodology pertaining to the scientific topic involves the techniques regarding how to go about conducting the research, the tools of research, advanced techniques that can be used in the conduct of the experiments and the like. Any student or research candidate is supposed to be good at both research methods and research methodology if he or she is to succeed in his or her attempt at conducting research into a subject.

1.10 SUMMARY

Research is designed to solve a particular existing problem so there is a much larger audience eager to support research that is likely to be profitable or solve problems of immediate concern. We also must understand how research impacts our decision making. Most people make decisions without gathered information to back them up. Only few do. The problem is most people aren't patient enough to put in the effort. Research requires time, effort, and sometimes money to have the evidence you need to make a sound decision that's why many avoid it. The role of research in the important areas of management has been briefly covered. The areas include marketing, production, banking, materials, human resource development, and government.

In conclusion research is very vital to our everyday decision making. It arms you from wrong information and save time and money. It

is important to your success as you take on life's challenges and career decisions making. But be careful though, because too much research without action on what you're learning is not good either. The question is how much information is enough? How much information can you afford? Information obesity can be research problem just my advice. Research plus action will most likely guarantee a successful research.

1.11 SELF-ASSESSMENT QUESTIONS

1. Define research.

2. What are the objectives of research?

3. State the significance of research.

4. What is the importance of knowing how to do research?

5. Highlight the different research approaches.

6. Explain the different types of research.

7. What do you understand by Research Methodology? Why is it needed? Explain.

8. Discuss the characteristics of a research study.

9. Explain the following:

(a). Scientific Method

(b). Empirical Research

10. Differentiate between Research methods and Research methodology.

11. What problems do come in doing effective research and how we can overcome them?

1.12 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, "Business Research Methods", McGraw Hill Inc.
- Alan Bryman, and Emma Bell, "Business Research Methods". Oxford University Press Publication.
- Brown, F.E. "Marketing Research, a structure for decision making", Addison - Wesley Publishing Company.
- Saunders, "Research Methods for Business Students", Pearsons Education Publications.
- Naresh Malhotra, "Marketing Research: An Applied Orientation", Prentice Hall International Edition.

- T.S. Wilkinson, and P.L. Bhandarker, “Methodology and Techniques of Social Research”, Himalaya Publishing House, New Delhi.
- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- Stockton and Clark, "Introduction to Business and Economic Statistics" D.B. Taraporevala Sons and Co. Private Limited, Bombay.
- Dollard, J., “Criteria for the Life-history”, Yale University Press, New York, 1935.
- Marie Jahoda, Morton Deutsch and Staurt W. Cook, “Research Methods in Social Relations”.
- Pauline V. Young, Scientific Social Surveys and Research.
- L.V. Redman and A.V.H. Mory, “The Romance of Research”, 1923.
- The Encylopaedia of Social Sciences, Vol. IX, Macmillan, 1930.

UNIT-2 RESEARCH PROBLEMS

Unit Framework

2.1 OBJECTIVE

After studying this unit, you will be able to:

- Understand the meaning of research problem
- Discuss the different factors to consider on research problem
- Explain process of research and describe the different levels of measurement
- Recognize what makes for good measurement
- Describe the sources of identifying research problem
- Explain ethical and non-ethical issues relating to business research and

- Understand environmental conditions

2.2 INTRODUCTION : MEANING OF RESEARCH PROBLEM

A research problem in general refers to some difficulty which a researcher experiences in the context of either a theoretical or practical situation and wants a solution for the same.

"The term problem means a question or issue to be examined". The term problem originates from the Greek word 'Probellim' – meaning anything that thrown forwards, a question proposed for solution, a matter stated for examination.

What is formulation? Formulation means "translating and transforming the selected Research problem in to a scientifically researchable question".

A research problem or phenomenon as it might be called in many forms of qualitative research is the topic you would like to address, investigate, or study, whether descriptively or experimentally. It is the focus or reason for engaging in your research. It is typically a topic, phenomenon, or challenge that you are interested in and with which you are at least somewhat familiar.

In other words, defining a research problem is the fuel that drives the scientific process, and is the foundation of any research method and experimental design, from true experiment to case study.

2.3 POINTS TO CONSIDER ON RESEARCH PROBLEM

The following points should be kept in mind while defining a research problem:

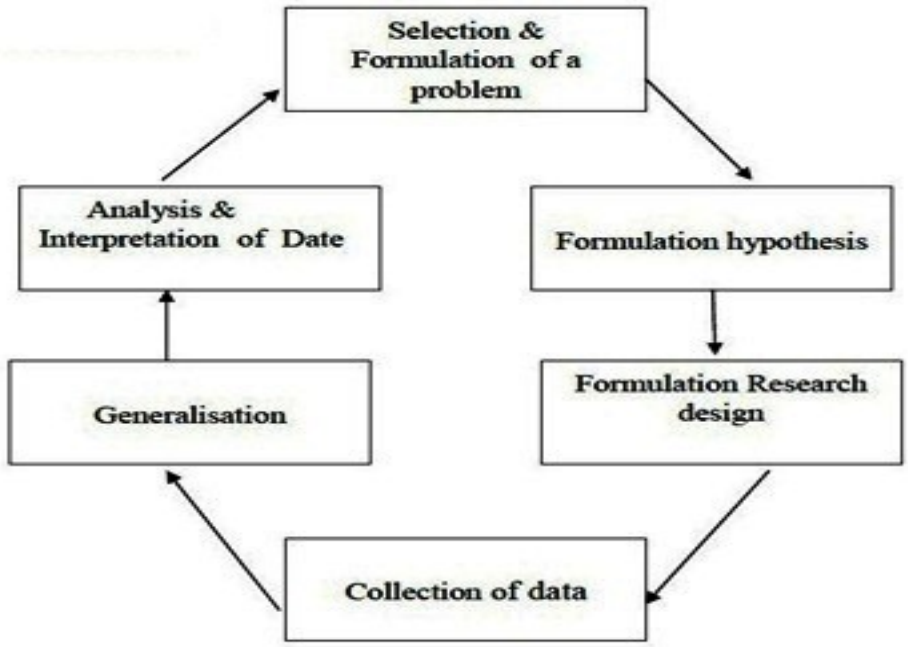
- The right question must be addressed if research is to aid decision makers. A correct answer to the wrong question leads either to poor advice or to no advice.
- Very often in research problem we have a tendency to rationalize and defend our actions once we have embarked upon a particular research plan. The best time to review and consider alternative approaches is in the planning stage. If this is done needless cost of false start and redoing work could be avoided.
- A good starting point in problem definition is to ask what the decision maker would like to know if the requested information could be obtained without error and without cost.

- Another good rule to follow is "Never settling on a particular approach" without developing and considering at least one alternative".
- The problem definition step of research is the determination and structuring of the decision maker's question. It must be the decision maker's question and not the researcher's question.
- What decision do you face? If you do not have decision to make, there is no research problem.
- What are your alternatives? If there are no alternatives to choose, again there is no research problem.
- What are your criteria for choosing the best alternative? If you do not have criteria for evaluation, again there is no research problem.
- The researcher must avoid the acceptance of the superficial and the obvious.

2.4 THE RESEARCH PROCESS

The Research Process is the Paradigm of research project. In a research project there are various scientific activities. The research process is a system of interrelated activities. Usually research begins with the selection of a problem. The various stages in the research process are explained in the above diagram. Research is a cyclical process. If the Data do not support the hypothesis, research is repeated again.

C.R. Kothari in his book, “Research Methodology: Methods & Techniques” presents a brief overview of a research process. He has given the following order concerning the Research Process.



- ❖ Formulation the Research problem
- ❖ Extensive Literature survey
- ❖ Developing the hypothesis
- ❖ Preparing the research design
- ❖ Determining sample design
- ❖ Collection of Data
- ❖ Execution of the Project
- ❖ Analysis of Data
- ❖ Hypothesis testing
- ❖ Generalisation & Interpretation
- ❖ Preparation of the report.

2.5 LEVELS OF MEASUREMENT

Measurement may be classified into four different levels, based on the characteristics of order, distance and origin.

1. **Nominal measurement:** This level of measurement consists in assigning numerals or symbols to different categories of a variable. The example of male and female applicants to an MBA program mentioned earlier is an example of nominal measurement. The numerals or symbols are just labels and have no quantitative value. The number of cases under each category is counted. Nominal measurement is therefore the simplest level of measurement. It does not have characteristics such as order, distance or arithmetic origin.

2. **Ordinal measurement:** In this level of measurement, persons or objects are assigned numerals which indicate ranks with respect to one or more properties, either in ascending or descending order.

   Example: Individuals may be ranked according to their "socio-economic class", which is measured by a combination of income, education, occupation and wealth. The individual with the highest score might be assigned rank 1, the next highest rank 2, and so on, or vice versa.

   The numbers in this level of measurement indicate only rank order and not equal distance or absolute quantities. This means that the distance between ranks 1 and 2 is not necessarily equal to the distance between ranks 2 and 3.

   Ordinal scales may be constructed using rank order, rating and paired comparisons. Variables that lend themselves to ordinal measurement include preferences, ratings of organizations and economic status. Statistical techniques that are commonly used to analyze ordinal scale data are the median and rank order correlation coefficients.

3. **Interval measurement:** This level of measurement is more powerful than the nominal and ordinal levels of measurement, since it has one additional characteristic, which is equality of distance. However, it does not have an origin or a true zero. This implies that it is not possible to multiply or divide the numbers on an interval scale.

   Example: The Centigrade or Fahrenheit temperature gauge is an example of the interval level of measurement. A temperature of 50 degrees is exactly 10 degrees hotter than 40 degrees and 10 degrees cooler than 60 degrees.

   Since interval scales are more powerful than nominal or ordinal scales, they also lend themselves to more powerful statistical techniques, such as standard deviation, product moment correlation and "t" tests and "F" tests of significance.

4. **Ratio measurement:** This is the highest level of measurement and is appropriate when measuring characteristics which have an absolute zero point. This level of measurement has all the three characteristics order, distance and origin.

   Examples: Height, weight, distance and area etc. are measured by natural numbers. Since there is a natural zero, it is possible to multiply and divide the numbers on a ratio scale. Apart from being able to use all the statistical techniques that are used with the nominal, ordinal and interval scales, techniques like the geometric mean and coefficient of variation may also be used.

The different levels of measurement and their characteristics may be summed up. In the table below:

LEVELS OF MEASUREMENT	CHARACTERISTICS
Nominal	No order, distance or origin
Ordinal	Order, but no distance or origin
Interval	Both order and distance, but no origin
Ratio	Order, distance and origin

2.6 CHARACTERISTICS OF GOOD MEASUREMENT

A good measurement tool must possess the following characteristics –

1. **Uni-dimensionality** – This means that the measurement scale should not measure more than one characteristic at a time. For example, a scale should measure only length and not both length and temperature at the same time.

2. **Linearity** – A good measurement scale should follow the straight line model.

3. **Validity** – This means that a measurement scale should measure what it is supposed to measure.

4. **Reliability** – This refers to consistency. The measurement scale should give consistent results.

5. **Accuracy and Precision** – The measurement scale should give an accurate and precise measure of what is being measured.

6. **Simplicity** – A measurement tool should not be very complicated or elaborate.

7. **Practicability** – The measurement tool should be easy to understand and administer. There should be proper guidelines regarding its purpose and construction procedure, so that the results of a test can be interpreted easily.

Of the above characteristics, validity and reliability are the most important requirements of a measurement scale and will be explained in more detail.

2.6.1 VALIDITY

Validity may be classified into different types, as described below. The degree of validity of each type is determined by applying logic, statistical procedures or both.

1. **Content validity :** This type of validity may be of two types – a) Face validity and b) Sampling validity. **Face validity** is determined through a subjective evaluation of a measuring scale. For example, a researcher may develop a scale to measure consumer attitudes towards a brand and pre-test the scale among a few experts. If the experts are satisfied with the scale, the researcher may conclude that the scale has face validity. However, the limitation of this type of validity is that it is determined by opinions, rather than through a statistical method. **Sampling validity** refers to how representative the content of the measuring instrument is. In other words, the measuring instrument's content must be representative of the content universe of the characteristic being measured.

2. **Predictive validity :** This type of validity refers to the extent to which one behavior can be predicted based on another, based on the association between the results yielded by the measuring instrument and the eventual outcome. **Example** – In the case of an admission test designed for prospective MBA students, the

predictive validity of the test would be determined by the association between the scores on the test and the grade point average secured by students during the first semester of study. A statistical measure of this association the correlation coefficient could be computed to determine the predictive validity of the admission test. Predictive validity would be strong if the coefficient is greater than .50.

One limitation of determining predictive validity using this statistical association is that the eventual outcome, in this case, the grade point average of students during the first semester, may be influenced by other "extraneous" variables or factors. Therefore, predicting behavior from one situation to another is not always accurate.

3. **Construct validity :** A construct is a conceptual equation that is developed by the researcher based on theoretical reasoning. Various kinds of relationships may be perceived by the researcher between a variable under study and other variables. These relationships must be tested in order to determine the construct validity of a measuring instrument. The instrument may be considered to have construct validity only if the expected relationships are found to be true. When determining the validity of a particular measurement instrument, all the three types of validity discussed above should be determined.

2.6.2 RELIABILITY

- This refers to the ability of a measuring scale to provide consistent and accurate results**.** To give a simple example, a weighing machine may be said to be reliable if the same reading is given every time the same object is weighed.

- There are two dimensions of reliability – stability and equivalence or non-variability. **Stability** refers to consistency of results with repeated measurements of the same object, as in the weighing machine example. **Non variability** refers to consistency at a given point of time among different investigators and samples of items.

- The problem of reliability is more likely to arise with measurements in the social sciences than with measurements in the physical sciences, due to factors such as poor memory or recall of respondents, lack of clear instructions given to respondents and irrelevant contents of the measuring instrument.

- The desired level of reliability depends on the research objectives, as well as the homogeneity of the population under study. If precise estimates are required, the higher will be the desired level of accuracy. In the case of a homogeneous population, a lower level of reliability may be sufficient, since there is not much variation in the data.

- Reliability and validity are closely interlinked. A measuring instrument that is valid is always reliable, but the reverse is not true. That is, an instrument that is reliable is not always valid. However, an instrument that is not valid may or may not be reliable and an instrument that is not reliable is never valid.

2.7 SOURCES OF IDENTIFYING RESEARCH PROBLEM

Researchers are, by nature, curious. But if you're just starting out as a researcher, trying to identify a problem that requires researching may prove more challenging and overwhelming than you thought. You need to choose a topic that is of interest to you, is relevant to its field and is need of clarification. There are some tips you can follow, though, that will help you identify a problem that is worthy of your time and energy.

Instructions :

1. Search for a problem in your everyday life. Look around you! Problems suitable for research exist everywhere. You might see them in your professional practice or personal life. Make a habit of asking you questions about what you see and hear. Why does such-and-such happen?

2. Read more about your field of studies. You definitely have topics that interest you in your chosen disciple, so look through professional journals and magazines, textbooks and dissertations to find out more about these topics. This will give you a clear idea about what is already known in your area of interest -- and what is still unknown. Reading also gives you theoretical base for your study and gives you information about a variety of research methods.

3. Take notes, or keep a research journal. Write down ideas that spark a possible research topic, such as an unexpected and contradictory finding of previous studies, suggestions that other established researches have given in the books for future research, perspectives and interesting project types that can be applied in new situations.

4. Seek professional advice. New researchers should learn from established ones. Attending professional conferences helps you make contacts with specialists and also gives you an idea of "what is hot" in the field. Approach the experts, and let them know that you are familiar with their work and you want to get some advice from them. Consult with a valued professor, as well.

5. Think about what interests you. Your topic needs to motivate you and capture the attention of others. Your research will likely take months or years of your time and effort, so it has to be something

you are passionate about, that you feel strongly needs to be shared with the public and that has possible practical applications.

2.8 ETHICAL ISSUES RELATING TO BUSINESS RESEARCH

There are a number of key phrases that describe the system of ethical protections that the contemporary social and medical research establishments have created to try to protect better the rights of their research participants. The principle of **voluntary participation** requires that people not be coerced into participating in research. This is especially relevant where researchers had previously relied on 'captive audiences' for their subjects -- prisons, universities, and places like that.

Confidentiality -- they are assured that identifying information will not be made available to anyone who is not directly involved in the study. The stricter standard is the principle of **anonymity** which essentially means that the participant will remain anonymous throughout the study even to the researchers themselves.

Clearly, the anonymity standard is a stronger guarantee of privacy, but it is sometimes difficult to accomplish, especially in situations where participants have to be measured at multiple time points (e.g., a pre-post study). Increasingly, researchers have had to deal with the ethical issue of a person's **right to service**. Good research practice often requires the use of a no-treatment control group a group of participants who do not get the treatment or program that is being studied. But when that treatment or program may have beneficial effects, persons assigned to the no-treatment control may feel their rights to equal access to services are being curtailed.

Even when clear ethical standards and principles exist, there will be times when the need to do accurate research runs up against the rights of potential participants. No set of standards can possibly anticipate every ethical circumstance.

Furthermore, there needs to be a procedure that assures that researchers will consider all relevant ethical issues in formulating research plans. To address such needs most institutions and organizations have formulated an **Institutional Review Board (IRB)**, a panel of persons who reviews grant proposals with respect to ethical implications and decides whether additional actions need to be taken to assure the safety and rights of participants. By reviewing proposals for research, IRBs also help to protect both the organization and the researcher against potential legal implications of neglecting to address important ethical issues of participants.

2.9 NON-ETHICAL ISSUES FOR RESEARCH

The related non-ethical issues for research are:

- **Fabrication :** In scientific inquiry and academic research, **fabrication** is the intentional misrepresentation of research results by making up data, such as that reported in a journal article. As with other forms of scientific misconduct, it is the intent to deceive that marks fabrication as highly unethical and different from scientists deceiving themselves. In some jurisdictions, fabrication may be illegal.

- **Falsification :** Falsification is manipulating research materials, equipment, or processes, or changing or omitting/suppressing data or results without scientific or statistical justification, such that the research is not accurately represented in the research record. This would include the "misrepresentation of uncertainty" during statistical analysis of the data.

- **Plagiarism :** Plagiarism is the "wrongful appropriation" and "purloining and publication" of another author's "language, thoughts, ideas, or expressions," and the representation of them as one's own original work. The idea remains problematic with unclear definitions and unclear rules. The modern concept of plagiarism as immoral and originality as an ideal emerged in Europe only in the 18th century, particularly with the Romantic Movement. Plagiarism is considered academic dishonesty and a breach of journalistic ethics. It is subject to sanctions like expulsion. Plagiarism is not a crime per se but in academia and industry it is a serious ethical offense, and cases of plagiarism can constitute copyright infringement.

2.10 ENVIRONMENTAL CONDITIONS

Environmental conditions fall within the category of relevant characteristics, but they comprise a special type of relevant characteristic. The characteristics of interest are the target variables. The research is undertaken in order to discover their values. Environmental conditions, however, are of concern because of their possible relationship with the characteristics of interest. What would sales be if prices were Rs. 169? Rs. 149? What would competitor do if we increased our advertising by 25%? Or decreased it by 25%? How would A's action affect our sales and profits? What would happen to the supply of oil if the depletion allowance were cut in half/ were removed completely?

The environmental conditions specified in the research problem are of two types; (1) those beyond the firm's control and (2) those within the firm's control. The firm must adjust to the first and choose wisely with respect to the second. Neither is possible without knowing how the particular variables influence the characteristics of interest. Therefore both types of variables must be introduced into the research problem.

For example, the research cannot study every price, every level and type of advertising support, or every sales training programme. Only a few

alternatives can be researched. The research problem must specify those which seem most promising. These specifications are critical; the research cannot answer unasked questions.

2.11 SUMMARY

A research problem or phenomenon as it might be called in many forms of qualitative research is the topic you would like to address, investigate, or study, whether descriptively or experimentally. It is the focus or reason for engaging in your research. It is typically a topic, phenomenon, or challenge that you are interested in and with which you are at least somewhat familiar.

The Research Process is the Paradigm of research project. In a research project there are various scientific activities. The research process is a system of interrelated activities. Usually research begins with the selection of a problem.

Measurement is an important concept in research and is a difficult task. It refers to the assignment of numerals to objects in order to measure the characteristics or properties of objects**.** Measurement may be classified into four different levels, based on three characteristics – order, distance and origin. The lowest level of measurement is nominal measurement and involves assigning numerals or labels to different categories of a variable. The next level is ordinal measurement in which objects are rank ordered with respect to a specific characteristic. The interval level of measurement has the characteristics of order, distance and equality of interval but no origin. The highest level of measurement is ratio measurement which is suitable for measuring properties which have an absolute zero point. It permits the use of advanced statistical techniques to analyze the data.

The characteristics of good measurement are uni-dimensionality, linearity, validity, reliability, accuracy, precision, simplicity and practicability. Validity refers to how effective an instrument is in measuring a property which it intends to measure. There are three types of validity – content validity, predictive validity and construct validity.

Reliability of a measuring instrument refers to its ability to provide consistent and accurate results with repeated measurements. Reliability and validity are closely associated. An instrument that is valid is also reliable, but not vice versa.

The environmental conditions specified in the research problem are of two types; (1) those beyond the firm's control and (2) those within the firm's control. The firm must adjust to the first and choose wisely with respect to the second. Neither is possible without knowing how the particular variables influence the characteristics of interest.

2.12 SELF-ASSESSMENT QUESTIONS

1. What is a research problem?

2. Elaborate the various sources of identifying research problem.

3. Discuss the ethical issued regarding business research.

4. Write a short note on Plagiarism

5. Distinguish between Fabrication and Falsification.

6. Differentiate between nominal, ordinal, interval and ratio scales, with an example of each.

7. What is meant by validity? How does it differ from reliability and what are its types?

8. What are the purposes of measurement in social science research?

2.13 TEXT AND REFERENCES

- Alan Bryman, and Emma Bell, “Business Research Methods”. Oxford University Press Publication.
- Brown, F.E. "Marketing Research, a structure for decision making", Addison - Wesley Publishing Company.
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- Stockton and Clark, "Introduction to Business and Economic Statistics" D.B. Taraporevala Sons and Co. Private Limited, Bombay.
- Pauline V. Young, Scientific Social Surveys and Research.

UNIT-3 RESEARCH DESIGN

Unit Framework

3.1 OBJECTIVE

After studying this unit you should be able to understand:

- Concepts of Research Design
- Characteristics of Research Design
- Classifications of Research Design
- Research Design in Case of Exploratory Research Studies
- Research Design in case of Descriptive and Diagnostic Research Studies

- Research Design in case of Observational Method
- Factors Affecting Research Design

3.2 INTRODUCTION : CONCEPT OF RESEARCH DESIGN

MEANING

Research is the study of materials, sources and data in order to get conclusions. Getting the research design right is the first step towards organized research, which is more likely to be good research. A research design is the program that guides the investigator in the process of collecting, analyzing and interpreting observations. It provides a systematic plan of procedure for the researcher to follow.

DEFINITION

- "It constitutes the blue print for the collection, measurement and analysis of data" - **Philips Bernard S**
- It "provides a systematic plan of procedure for the researcher to follow" -**Best John N**
- "The design research from controlling general scientific model into varied research procedure"- **P.V. Young**
- "A research design is "the programme that guides the investigator in the process of collecting, analysis and interpreting observations". – **David and Shava**
- "A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure."- **Elltiz, Jahoda and Destsch and Cook**

CONCEPT OF RESEARCH DESIGN

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

1. The nature of the study
2. The purpose of the study
3. The location where the study would be conducted
4. The nature of data required
5. From where the required data can be collected
6. What time period the study would cover
7. The type of sample design that would be used
8. The techniques of data collection that would be used
9. The methods of data analysis that would be adopted and
10. The manner in which the report would be prepared

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

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

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

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

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

3.3 FEATURES OF RESEARCH DESIGN

When a researcher has formulated a research problem, he/she has to focus on developing a good design for solving the problem. A good design is one that minimizes bias and maximizes the reliability of the data. It also yields maximum information, gives minimum experimental error, and provides different aspects of a single problem. A research design depends on the purpose and nature of the research problem. Thus, one single design cannot be used to solve all types of research problem, i.e., a particular design is suitable for a particular problem. Some of the important features are as follows:

1. It is a plan that specifies the objectives of study and the hypothesis to be tested.
2. It is an outline that specifies the sources and types of information relevant to the research question.

3. It is a blueprint specifying the methods to be adopted for gathering and analysis of data.

4. It is a scheme defining the procedure involved in a research process.

5. It is a series of guide posts to keep one going in the right direction.

6. It reduces wastage of time and cost.

7. It encourages co-ordination and effective organization.

8. It is a tentative plan which undergoes modifications, as circumstances demand, when the study progresses, new aspects, new conditions and new relationships come to light and insight into the study deepens.

9. It has to be geared to the availability of data and the cooperation of the informants.

10. It has also to be kept within the manageable limits

If a research study is an exploratory or formulative one, i.e., it focuses on discovery of ideas and insights, the research design should be flexible enough to consider different aspects of the study. Similarly, if the study focuses on accurate description or association between variables, the design should be accurate with minimum bias and maximum reliability. However, in practice, it is difficult to categorize a particular study into a particular group. A study can be categorized only on the basis of its primary function and accordingly, its design can be developed. Moreover, the above mentioned factors must be given due weightage while working on the details of the research design.

3.4 CLASSIFICATION OF RESEARCH DESIGN

What are the different major types of research designs? We can classify designs into a simple threefold classification by asking some key questions. First, does the design use random assignment to groups? [Don't forget that random assignment is not the same thing as random selection of a sample from a population!] If random assignment is used, we call the design a **randomized experiment** or **true experiment**. If random assignment is not used, then we have to ask a second question: Does the design use **either** multiple groups or multiple waves of measurement? If the answer is yes, we would label it a **quasi-experimental design**. If no, we would call it a **non-experimental design**. This threefold classification is especially useful for describing the design with respect to internal validity. A randomized experiment generally is the strongest of the three designs when your interest is in establishing a cause-effect relationship. A non-experiment is generally the weakest in this respect. I have to hasten to add here; that I don't mean that a non-experiment is the weakest of the three designs overall, but only with respect to internal validity or causal

assessment. In fact, the simplest form of non-experiment is a one-shot survey design that consists of nothing but a single observation

There are a number of crucial research choices, various writers advance different classification schemes, some of which are:

1. Experimental, historical and inferential designs (American Marketing Association).
2. Exploratory, descriptive and causal designs (Selltiz, Jahoda, Deutsch and Cook).
3. Experimental and expose fact (Kerlinger)
4. Historical method, and case and clinical studies (Goode and Scates)
5. Sample surveys, field studies, experiments in field settings, and laboratory experiments (Festinger and Katz)
6. Exploratory, descriptive and experimental studies (Body and Westfall)
7. Exploratory, descriptive and casual (Green and Tull)
8. Experimental, „quasi-experimental designs " (Nachmias and Nachmias)
9. True experimental, quasi-experimental and non-experimental designs (Smith).
10. Experimental, pre-experimental, quasi-experimental designs and Survey Research (Kidder and Judd).

These different categorizations exist, because „research design " is a complex concept. In fact, there are different perspectives from which any given study can be viewed. They are:

1. The degree of formulation of the problem (the study may be exploratory or formalized)
2. The topical scope-breadth and depth-of the study(a case or a statistical study)
3. The research environment : field setting or laboratory (survey, laboratory experiment)
4. The time dimension (one-time or longitudinal)
5. The mode of data collection (observational or survey)

6. The manipulation of the variables under study (experimental or expose facto)

7. The nature of the relationship among variables (descriptive or causal)

3.5 EXPLORATORY RESEARCH DESIGN

3.5.1 EXPLORATORY RESEARCH

This research is conducted for a problem that has not been clearly defined. Exploratory research helps determine the best research design, data collection method and selection of subjects. It should draw definitive conclusions only with extreme caution. Given its fundamental nature, exploratory research often concludes that a perceived problem does not actually exist.

Given below are some of the uses of exploratory research:

- Formulating a problem or defining a problem more precisely.
- Identifying alternative courses of action.
- Developing hypothesis.
- Isolating key variables and relationships for further examination.
- Gaining insights for developing an approach to the problem.
- Establishing priorities for further research.

The following are the circumstances in which exploratory study would be ideally suited:

1. To gain an insight into a problem.
2. To list out all possibilities, from which one can prioritize that possibility which seems likely.
3. To develop hypotheses.
4. It can also be used to increase the analyst's familiarity with the problem, particularly when the analyst is new to the problem area. Example: A market researcher working for (new entrant) a company for the first time.
5. Exploratory studies may be used to clarify concepts and help in formulating precise problems.
6. To pre-test a draft questionnaire.

7. In general, exploratory research is appropriate to any problem about which very little is known. This research is the foundation for any future study.

Given below are some of characteristics of exploratory research:

1. Exploratory research is often the front end of total research design.
2. It is flexible, unstructured and very versatile.
3. Experimentation is not a requirement.
4. Cost incurred to conduct study is low.
5. This type of research allows very wide exploration of views.
6. Research is interactive in nature and also it is open ended.

Hypothesis Development at Exploratory Research Stage

At exploratory stage:

1. Sometimes, if the situation is being investigated for the first time, it may not be possible to develop any hypothesis at all. This is because of non-availability any previous data.
2. Sometimes, some information may be available and it may be possible to formulate a tentative hypothesis.
3. In other cases, most of the data is available and it may be possible to provide answers to the problem.

3.5.2 DESCRIPTIVE RESEARCH

Descriptive research is used to describe characteristics of a population or phenomenon being studied. It does not answer questions about how/when/why the characteristics occurred. Some of the questions that need to be answered before data collection for this descriptive study are as follows:

- **Who:** unit of analysis
- **What:** information need from the respondent
- **When:** the information should be obtained, before shopping, after shopping, etc.
- **Where:** the respondent should be contacted
- **Why:** why we are getting the info from the respondent
- **Way:** how to get the info, questionnaire, survey, etc.

Hence, research cannot describe what caused a situation. Thus, Descriptive research cannot be used to as the basis of a causal relationship, where one variable affects another. In other words, descriptive research

can be said to have a low requirement for internal validity. There are three main types of descriptive methods: observational methods, case study methods and survey methods.

Descriptive research is used:

- To describe the characteristics of relevant groups, such as consumers, salespeople, organizations, or market areas.
- To estimate the percentage of units in a specified population exhibiting a certain behavior.
- To determine the perceptions of product characteristics.
- To determine the degree to which marketing variables are associated.
- To make specific predictions.

3.5.3 OBSERVATIONAL METHOD

With the observational method (sometimes referred to as field observation) animal and human behavior is closely observed. There are two main categories of the observational method — naturalistic observation and laboratory observation.

The biggest advantage of the naturalistic method of research is that researchers view participants in their natural environments. This leads to greater ecological validity than laboratory observation, proponents say.

Ecological validity refers to the extent to which research can be used in real-life situations.

Proponents of laboratory observation often suggest that due to more control in the laboratory, the results found when using laboratory observation are more meaningful than those obtained with naturalistic observation.

Laboratory observations are usually less time-consuming and cheaper than naturalistic observations. Of course, both naturalistic and laboratory observation are important in regard to the advancement of scientific knowledge.

3.5.4 CASE STUDY METHOD

Case study research involves an in-depth study of an individual or group of individuals. Case studies often lead to testable hypotheses and allow us to study rare phenomena. Case studies should not be used to determine cause and effect, and they have limited use for making accurate predictions.

There are two serious problems with case studies — expectancy effects and atypical individuals. Expectancy effects include the experimenter's underlying biases that might affect the actions taken while conducting research. These biases can lead to misrepresenting participants' descriptions. Describing atypical individuals may lead to poor generalizations and detract from external validity.

3.5.5 SURVEY METHOD

In survey method research, participants answer questions administered through interviews or questionnaires. After participants answer the questions, researchers describe the responses given. In order for the survey to be both reliable and valid it is important that the questions are constructed properly. Questions should be written so they are clear and easy to comprehend.

Another consideration when designing questions is whether to include open-ended, closed-ended, partially open-ended, or rating-scale questions. Advantages and disadvantages can be found with each type:

Open-ended questions allow for a greater variety of responses from participants but are difficult to analyze statistically because the data must be coded or reduced in some manner. Closed-ended questions are easy to analyze statistically, but they seriously limit the responses that participants can give. Many researchers prefer to use a Likert-type scale because it's very easy to analyze statistically.

In addition to the methods listed above some individuals also include qualitative (as a distinct method) and archival methods when discussing descriptive research methods.

It is important to emphasize that descriptive research methods can only describe a set of observations or the data collected. It cannot draw conclusions from that data about which way the relationship goes — Does A cause B, or does B cause A?

Unfortunately, in many studies published today, researchers forget this fundamental limitation of their research and suggest their data can actually demonstrate or "suggest" causal relationships. Nothing could be further from the truth.

3.6 FACTORS AFFECTING RESEARCH DESIGN

- Availability of scientific information
- Availability of sufficient data
- Time availability
- Proper exposure to the data source

- Availability of the money
- Manpower availability
- Magnitude of the management problem
- Degree of Top management's support
- Ability, knowledge, skill, technical understanding and technical background of the researcher
- Controllable variables
- Un–controllable variables
- Internal variables
- External variables

3.7 RELATIONSHIPS AMONG EXPLORATORY, DESCRIPTIVE, & CAUSAL RESEARCH

The distinctions among exploratory, descriptive, and causal research as major classifications of research designs are not absolute. For example a research project may involve more than one type of research design and thus serve several purposes. Which combination of research designs should be employed depends on the nature of the problem. The following are the general guidelines for choosing research designs:

A Comparison of Basic Research Designs

	Exploratory	**Descriptive**	**Causal**
Objective	Discovery of Ideas	Describes market characteristics	Determine cause and effect
Characteristics	Flexible, versatile, Front-end Research	Prior formulation of hypothesis, planned, structured design	Manipulate independent variables, Control of other variables
Methods	Secondary data	Surveys	Experiments

Differences between Exploratory and Conclusive Research

Research Project Components	Exploratory Research	Conclusive Research
Research purpose	General: to generate insights about a situation	Specific: to verify insights and aid in selecting a course of action
Data needs	Vague	Clear
Data sources	III defined	Well defined
Data collection form	Open-ended, rough	Usually structured
Sample	Relatively small; subjectively selected to maximize generalization of insights	Relatively large; objectively selected o permit generalization of findings
Data collection	Flexible; No set procedure	Rigid: well-laid-out procedure
Data analysis	Informal; typical non quantitative	Formal; typically quantitative
Inferences/recommendations	More tentative than final	More final than tentative

3.8 SUMMARY

A research design is a logical and systematic plan prepared for directing a research study. In many research projects, the time consumed in trying to ascertain what the data mean after they have been collected is much greater than the time taken to design a research which yields data whose meaning is known as they are collected. Research design is a series

of guide posts to keep one going in the right direction. It is a tentative plan which undergoes modifications, as circumstances demand, when the study progresses, new aspects, new conditions and new relationships come to light and insight into the study deepens. Exploratory research studies are also termed as formulative research studies. The main purpose of such studies is that of formulating a problem for more precise investigation or of developing the working hypothesis from an operational point of view. Descriptive research studies are those studies which are concerned with describing the characteristics of a particular individual, or of a group, whereas diagnostic research studies determine the frequency with which something occur or its association with something else.

3.9 SELF-ASSESSMENT QUESTIONS

1. What do you mean by research design?

2. State the features of research design.

3. Discuss in detail the classification of research design.

4. Define the following:

(a) Exploratory research

(b) Descriptive research

(c) Observational Method

(d) Case Study Method

(e) Survey Method

5. What are the factors affecting research design

6. What are the characteristics of a good research design?

7. What are the different types of research designs?

8. What are the features of an exploratory research design?

9. Why is research design necessary to conduct a study?

10. What are the various types of research design? Explain with examples.

11. What is exploratory research? Give Example under what circumstances, exploratory research is ideal.

12. What are the sources available for data collection at exploratory stage?

13. Distinguish exploratory from descriptive research.

3.10 TEXT AND REFERENCES

- Alan Bryman, and Emma Bell, "Business research Methods". Oxford University Press Publication.
- Saunders, "Research Methods for Business Students", Pearsons Education Publications.
- Naresh Malhotra, "Marketing Research: An applied Orientation", Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, "Methodology and techniques of Social Research", Himalaya Publishing House, New Delhi.
- Donald S. Tull and Del S. Hawkins, "Marketing Research – Measurement and Method", Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

UNIT-4 DATA COLLECTION

Unit Framework

4.1 OBJECTIVE

After reading this unit you will be able to:

- Understand the meaning of data collection.
- Explain the importance of data
- Identify the various types of data.
- Explain from where the data is collected.
- Understand the importance of primary and secondary data
- Explain advantages and disadvantages of primary and secondary data.

4.2 INTRODUCTION : MEANING AND IMPORTANCE OF DATA

DEFINITION

"Data are facts, figures and other relevant materials past and present serving as bases for study and analysis".

MEANING OF DATA

The search for answers to research questions calls collection of Data. "Data are facts, figures and other relevant materials, past and present, serving as bases for study and analysis".

TYPES OF DATA

The Data needed for social science Research may be broadly classified into:

A. Data pertaining to human beings

B. Data relating to organizations

C. Data pertaining to territorial area.

A) Personal Data (relating to Human beings) are of two types.

(a). Demographic and socio-economic characteristics of individuals. Like name, sex, race, social class, relation, education, occupation, income etc.

(b). Behavioural Variables: Attitudes, opinion knowledge, practice, intensions etc.

B) Organisation Data: - Consist of data relating to an organizations, origin ownership, function, performance etc.

C) Territorial Data: - are related to geo-physical characteristic, population, infrastructure etc. of divisions like villages, cities, taluks, distinct, state etc.

IMPORTANCE OF DATA

The data serve as the bases or raw materials for analysis without Data no specific inferences can be drawn on our study. Inferences based on imagination or guesswork cannot provide correct answers to research questions. The relevance, adequacy and reliability of data determine the quality of the findings of a study. The reliability of data determines the quality of research.

Data form the basis for testing the hypotheses formulated in a study. Data also provide the facts and figures required for constructing measurement scales and tables, which are analysed with statistical techniques. Inferences on the results of statistical, analysis and tests of significance

provide the answers to research questions. Thus the scientific process of measurement, analysis, testing and inferences depends on the availability of relevant data and their accuracy. Hence the importance of data for any research studies.

SOURCES OF DATA

The sources of data may be classified into a) primary sources b) secondary sources. Both the sources of information have their merits and demerits. The selection of a particular source depends upon the (a) purpose and scope of enquiry, (b) availability of time, (c) availability of finance, (d) accuracy required, (e) statistical tools to be used, (f) sources of information (data), and (g) method of data collection.

4.3 PRIMARY SOURCES OF DATA

Primary sources are original sources from which the researcher directly collects data that have not been previously collected e.g., collection of data directly by the researcher on brand awareness, brand preference, brand loyalty and other aspects of consumer behaviour from a sample of consumers by interviewing them. Primary data are firsthand information collected through various methods such as observation, interviewing, mailing etc.

According to **P. V. Young**, "primary sources are those data gathered at first hand and the responsibility so of their compilation and promulgations remaining under the same authority that originally gathered them."

In the words of **Watter R. Borg**, "Primary sources are direct describing occurrences by an individual who actually observed on witness for occurrences."

4.3.1 ADVANTAGE OF PRIMARY DATA

- It is original source of data
- It is possible to capture the changes occurring in the course of time.
- It flexible to the advantage of researcher.
- Researchers know its accuracy.
- Only that data are collected which meet outs the objective of research project.
- In maximum methods of primary data collection researchers know who are the respondents so face to face communication is there.
- It is most authentic since the information is not filtered or tampered.

- Extensive research study is based of primary data

4.3.2 DISADVANTAGE OF PRIMARY DATA

- Primary data is expensive to obtain
- It is time consuming
- It requires extensive research personnel who are skilled.
- It is difficult to administer.
- Chances of biasness are at great extent.
- Biasness can also be there on the part of respondent. Wrong answer can be given by then which may affect the accuracy of data.
- It may have narrow coverage. It means researchers may collect data only within his/her reach or according to his mindset.

4.3.3 METHODS OF COLLECTING PRIMARY DATA

The following are the methods of collecting Primary Data:

1. PERSONAL INTERVIEW
2. OBSERVATION METHOD
3. DATA COLLECTED THROUGH MAIL
4. WEB METHOD
5. TELEPHONE METHOD

1. **PERSONAL INTERVIEW:** Under this method the researcher personally visits the area of enquiry, establishes personal contact with the respondent and collects necessary facts and figures.

   Example: If a researcher wants to know about the family income of persons in a particular area, he goes personally to the area and collects data on the basis of personal contacts. It will be direct personal interview.

Advantages :

The main advantages of personal interview are:

(i) Generally yields highest cooperation and lowest refusal rates: Since personal visits are made to the respondents, to the refusal rates are low as better explanation about the research work can be given to the respondent.

(ii) Allows for longer, more complex interviews: Since contact is made personally and the researcher has time to explain the

about the complexity of the research, in details, to the respondent, so he may go for complex interviews where questions have to be framed at the point of the respondent.

(iii) High response quality: The response quality of the respondent can be judged at the time of collecting the data, so the quality of data is controlled.

(iv) Multi-method data collection: Under this method the person collecting the data may change the method of data collection if he is not getting appropriate data by from his existing method.

Disadvantages

The main disadvantages of personal interview are:

(i) Most costly mode of administration: Since personal visits are made for conducting the interview and also two respondents may be at far away distance, so the cost of collecting the data increases, as it will include the transportation cost of the person collecting the data.

(ii) Longer data collection period: Since the data is collected by personal visit, so it becomes a very slow process to collect data. It" takes comparative longer period to collect the data under this method.

(iii) Interviewer concerns: Since the interview is conducted by making personal visits, so there is always a chance, that the respondent may not respond to the process, or may give his own suggestion for changes in the research and provide data according to the changes.

2. OBSERVATION METHOD: Under this method, the researcher collects information directly through systematic watching and noting the phenomena as they occur in nature with regard to cause and effect or mutual relations rather than through the reports of others. It is a process of recording relevant information without asking anyone specific questions and in some cases, even without the knowledge of the respondents.

Advantages :

(i) The respondent will provide data: In any case the data is collected from the respondent by observing the response pattern and the respondents are unable or reluctant to provide information.

(ii) Data Accuracy: The method provides deeper insights into the problem and generally the data is accurate and quicker to process. Therefore, this is useful for intensive study rather than extensive study.

Disadvantages

(i) Unable to predict the occurrence of data: In many situations, the researcher cannot predict when the events will occur. So when an event occurs there may not be a ready observer to observe the event.

(ii) No true data: As the respondent may be aware of the observer and as a result may alter their behavioural pattern.

(iii) Paradigm: Since this method solely depends on the observation power of the researcher, so due to lack of training and paradi1 the researcher may not observe the things as they occur.

(iv) Not suitable for large research: This method cannot be used extensively if the inquiry is large and spread over a wide area.

3. DATA COLLECTED THROUGH MAIL: Under this method the data a collected by sending letters to the respondent. A letter may contain questionnaire and the respondent is required to respond back.

Advantages

(i) **Generally lowest cost:** As compared to other form of data collection, this method is cheapest as the questionnaire maximum of one page. Further to cut the cost, the researcher may go through print media to get its questionnaire distributed to the respondent.

(ii) **Can be administered by smaller team of people:** This research can be administered with few staffs as only the office staffs are required and no field staffs are necessary.

(iii) **Access to otherwise difficult to locate, busy populations:** As the research is done through mail, so the researcher could get the data from the respondent otherwise it is difficult to locate amongst the busy population.

(iv) **Respondents can look up information or consult with others:** As the respondent don't have to respond to the queries of the researcher, so the respondent gets enough time to understand the information required and he may even consult other person to provide the accurate data.

Disadvantages

(i) **Most difficult to obtain cooperation:** As the respondent has the option to respond back or not, so the researcher may find it difficult to collect information as all of its questionnaire

sent may not return with the data. Also there is problem of delivery of mails to the respondent.

(ii) No researcher involved in collection of data: As all the work of collection of data are through mails, so there are no researcher involved in the process, so the respondent may sometimes find it difficult to understand the queries raised by the researcher.

(iii) Need good sample: As the mails are sent with the help of some database, so the researcher may not know, how the respondent will react? The respondent mayor may not provide appropriate data as required by the researcher.

(iv) More likely to need an incentive for respondent: In order to make respondent give response to the letter sent, there should be some incentive scheme to be attached with the mail, otherwise, the response rate will be very poor.

(v) Slower data collection period: As the time needed for delivery of mails to the respondent and back, so this method is much slower than how the data is collected through telephone or personal interview or web mail.

4. WEB METHOD: Under this method, the data is collected through internet. This method can be further divided into two groups

A. Through Polling: The researcher may put the information on a web server and the respondent may require responding to the information through online poll, or blog.

B. Through Mails: The researcher may also opt to send emails to various respondents and may give them the option to respond back.

Advantages

(i) Lower cost: As the research work is completed online and the data are collected in the database, so the researcher may not require any paper, postage, mailing, data entry costs.

(ii) Can reach international populations: As the research work is done online, so the researcher may also involve international respondent for collection of data at no extra cost or effort.

(iii) Time required for implementation reduced: As the respondent is required to send the response online, and the data is also collected, so the researcher don't have to waste time in compiling the data and interpreting the data. The researcher may directly go for interpretation.

(iv) Complex patterns can be programmed: As the research work is online, so the researcher may go for complex

research activity as all the queries of the respondent are immediately handled by the researcher.

(v) Sample size can be greater: As the research work is online, the sample size for the research may be greater as the population outside the country can be included.

Disadvantages

(i) Limitation of technology: As in India, approximately 55% of homes own a computer; 30% have home e-mail, so the choice of population is restricted for the researcher.

(ii) Representative samples difficult: Since the access to the technology is difficult for general population, so the data collection activity cannot generate random samples of the population.

(iii) Differences in capabilities of people's computers and software for accessing Web surveys: Since each person has different capabilities and knowledge about the usage and utility of the web, so a good respondent may not provide the sample for the research.

(iv) Difference in people's response: The researcher cannot ascertain, if the same person has given response to the survey. If a research activity is performed by some other person, then the quality of response will be different.

(v) Different ISPs/line speeds limits extent of graphics that can be used: If the researcher is using graphic display to explain the theme or complexities of his research to the respondent, so it is quite possible that some of the respondent may not get the graphics displayed on their computer due to ISPs/line speeds limit or restriction of usage. In' such case the data collected will not be accurate.

5. TELEPHONE METHOD: Under this method, the researcher calls the respondent and collects the data over the telephone. He may use the telephone numbers available on the Telephone directory, and select the samples from the given population.

Advantages

(i) Less expensive than personal interviews: As the research can be completed over the phone, so this is less expensive than the personal interview and that the ·data are collected quickly.

(ii) Samples from general population: As telephones are accessible to the general population, than the Web Method, so more samples can be collected from larger population.

(iii) **Shorter data collection period than personal interviews:** As limited tools can be used to explain the research objective to the respondent, so it takes much lesser time to collect data. Also the larger population can be reached in very short time.

(iv) **Researcher administration:** As the researcher can explain and listen to the queries of the respondent, so in this method there is direct control of the researcher over the subject for data collection than the mails or web mails.

(v) **Better control and supervision of Researcher:** Similarly, as the researcher is contacting the respondent from his place, so the researcher may refer any literature during the process of collecting data. This is restricted in case of personal interview.

(vi) **Better response rate than mail for list samples:** For Comparative Scaling techniques this method is easy and effective as the researcher will provide the list and respondent have to choose from the available lists.

Disadvantages:

(i) **Biased against households without telephones, unlisted numbers:** As still most homes in India, don't have telephone numbers, or their numbers are not listed in the telephone directories due to having pre-paid connections. In such case such persons don't get equal opportunities to appear for research activity.

(ii) **No response:** If the respondents don't pick up the telephone on time or if the call is missed, in such case the researcher may select any other person as sample.

(iii) **Questionnaire constraints:** Complex questionnaire prepared by the researcher cannot be used in this method, if next question of the questionnaire depends on the answer of the respondent then such types of questionnaire cannot be included in the questionnaire.

(iv) **Difficult to administer questionnaires on sensitive or complex topics:** Since the researcher is getting the names of the samples from the telephone directory, and he is not aware about the economic, social or emotional situation of the samples, so it is very difficult to administer questionnaires on sensitive or complex topics.

4.4 SECONDARY SOURCES OF DATA

These are sources containing data which have been collected and compiled for another purpose. The secondary sources consists of readily compendia and already compiled statistical statements and reports whose

data may be used by researchers for their studies e.g., census reports , annual reports and financial statements of companies, Statistical statement, Reports of Government Departments, Annual reports of currency and finance published by the Reserve Bank of India, Statistical statements relating to Cooperatives and Regional Banks, published by the NABARD, Reports of the National sample survey Organization, Reports of trade associations, publications of international organizations such as UNO, IMF, World Bank, ILO, WHO, etc., Trade and Financial journals newspapers etc.

Secondary sources consist of not only published records and reports, but also unpublished records. The latter category includes various records and registers maintained by the firms and organizations, e.g., accounting and financial records, personnel records, register of members, minutes of meetings, inventory records etc.

4.4.1 FEATURES OF SECONDARY DATA

Though secondary sources are diverse and consist of all sorts of materials, they have certain common characteristics.

First, they are readymade and readily available, and do not require the trouble of constructing tools and administering them.

Second, they consist of data which a researcher has no original control over collection and classification. Both the form and the content of secondary sources are shaped by others. Clearly, this is a feature which can limit the research value of secondary sources.

Finally, secondary sources are not limited in time and space. That is, the researcher using them need not have been present when and where they were gathered.

4.4.2 USES OF SECONDARY DATA

The second data may be used in three ways by a researcher. First, some specific information from secondary sources may be used for reference purpose. For example, the general statistical information in the number of co-operative credit societies in the country, their coverage of villages, their capital structure, volume of business etc., may be taken from published reports and quoted as background information in a study on the evaluation of performance of cooperative credit societies in a selected district/state.

Second, secondary data may be used as bench marks against which the findings of research may be tested, e.g., the findings of a local or regional survey may be compared with the national averages; the performance indicators of a particular bank may be tested against the corresponding indicators of the banking industry as a whole; and so on.

Finally, secondary data may be used as the sole source of information for a research project. Such studies as securities Market Behaviour, Financial Analysis of companies, Trade in credit allocation in commercial banks, sociological studies on crimes, historical studies, and the like, depend primarily on secondary data. Year books, statistical reports of government departments, report of public organizations of Bureau of Public Enterprises, Censes Reports etc, serve as major data sources for such research studies.

4.4.3 ADVANTAGES OF SECONDARY DATA

Secondary sources have some advantages:

1. Secondary data, if available can be secured quickly and cheaply. Once their source of documents and reports are located, collection of data is just matter of desk work. Even the tediousness of copying the data from the source can now be avoided, thanks to Xeroxing facilities.
2. Wider geographical area and longer reference period may be covered without much cost. Thus, the use of secondary data extends the researcher's space and time reach.
3. The use of secondary data broadens the data base from which scientific generalizations can be made.
4. Environmental and cultural settings are required for the study.
5. The use of secondary data enables a researcher to verify the findings bases on primary data. It readily meets the need for additional empirical support. The researcher need not wait the time when additional primary data can be collected.

4.4.4 DISADVANTAGES OF SECONDARY DATA

Although secondary data are easy to access and cost-effective, they also have significant limitations:

1. The secondary data are not up-to-date and become obsolete when they appear in print, because of time lag in producing them. For example, population census data are published two or three years later after compilation and no new figures will be available for another ten years.
2. Data may be too broad-based that is, not specific enough to adequately address the firm's research questions.
3. The units in which the data are presented may not be meaningful.
4. The source of the data may not provide sufficient supporting material to allow the researcher to judge the quality of the research.

5. The data sources may lack reliability and credibility. Some secondary data may simply be inaccurate.

6. The most important limitation is the available data may not meet our specific needs. The definitions adopted by those who collected those data may be different; units of measure may not match; and time periods may also be different.

7. The available data may not be as accurate as desired. To assess their accuracy we need to know how the data were collected.

8. Finally, information about the whereabouts of sources may not be available to all social scientists. Even if the location of the source is known, the accessibility depends primarily on proximity. For example, most of the unpublished official records and compilations are located in the capital city, and they are not within the easy reach of researchers based in far off places.

4.4.5 METHODS OF COLLECTING SECONDARY DATA

The researcher may get Secondary data from two sources (a) Published; (b) Unpublished. While we are going to discuss about published data, the unpublished data may be in records of Government or private organizations, research organizations, research scholars etc. However, these data being unpublished is not freely available and the details about the sources remain with few persons.

Most of the researcher use published data as they are available from the following sources:

1. **Newspaper and Magazines:** Statistical data on a number of current socioeconomic subjects can be obtained from data collected and published by some reputed newspapers, magazines, periodicals, etc.

2. **Published Articles of an individual:** Many times some research studies are carried on at individual level and are published in magazines in the form of articles or in the form of books. Though these studies are based on research works of limited area, however they may provide useful information for further research.

3. **Government Publication and Gazetteer:** Various ministries and departments of Central Government and State Governments publish data regularly on a number of subjects. These data are considered reasonably reliable for research work.

4. **Reports from Commissions and Committees:** The government constitutes committees and commissions for the study and enquiry of various problems from time to time, which submit their reports

after collection and analysis of required data and information. The information contained in such reports is very useful and reliable.

5. **Publications from various Organizations:** Many important universities and semi-government research organizations also publish their research studies and these publications are important sources for analytical statistical information.

6. **Private Data Publication:** Recent trends have emerged where the private organizations are collecting and compiling various data for further usage as research data.

4.5 DIFFERENCE BETWEEN PRIMARY DATA AND SECONDARY DATA

(a) Primary data are collected by the researcher himself, although the secondary data has been collected previously by other researcher.

(b) Primary data are collected and used first time. However, on the secondary data, some decisions has been made previously, such decisions mayor may not be useful for the researcher now.

(c) Since primary data is collected by the researcher himself, so it relates directly to the research objective *or* may be more close to research objective. How many parts of secondary data need not to be related to the research objective?

(d) Since primary data is collected by the researcher, so it is more time taking activity for the researcher than to get the secondary data.

The students should note here that the primary data and secondary data differ in the degree of impact only and that same set of data may be secondary in the hands of one and primary in the hands of another. As Secrist has said "the distinction between primary am secondary data is largely one of degree. Data which are secondary in the hands of one party may be primary in the hands of another" Example, if population statistics are primary which it is collected under population census but it becomes secondary when it is used for some researcher.

4.6 QUESTIONNAIRE AND SCHEDULE CONSTRUCTION

QUESTIONNAIRE

A **questionnaire** is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. Although they are often designed for statistical analysis of the responses, this is not always the case. The questionnaire was invented by Sir Francis Galton. Questionnaires have advantages over some other types of surveys in that they are cheap, do not require as much effort from

the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data.

However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Thus, for some demographic groups conducting a survey by questionnaire may not be practical. As a type of survey, questionnaires also have many of the same problems relating to question construction and wording that exist in other types of opinion polls.

Question Sequence: In general, questions should flow logically from one to the next. To achieve the best response rates, questions should flow from the least sensitive to the most sensitive, from the factual and behavioural to the attitudinal, and from the more general to the more specific. There typically is a flow that should be followed when constructing a questionnaire in regards to the order that the questions are asked. The order is as follows:

1. Screens
2. Warm-ups
3. Transitions
4. Skips
5. Difficult
6. Classification

Screens are used as a screening method to find out early whether or not someone should complete the questionnaire. **Warm-ups** are simple to answer, help capture interest in the survey, and may not even pertain to research objectives. **Transition** questions are used to make different areas flow well together. **Skips** include questions similar to "If yes, then answer question 3. If no, then continue to question 5." **Difficult** questions are towards the end because the respondent is in "response mode." Also, when completing an online questionnaire, the progress bars let the respondent know that they are almost done so they are more willing to answers more difficult questions. **Classification**, or demographic, question should be at the end because typically they can feel like personal questions which will make respondents uncomfortable and not willing to finish survey.

SCHEDULE

This method of data collection is very much like the collection of data through questionnaire with little difference which likes in the fact that schedules are being filled in by enumerators who are specially appointed for this purpose. These enumerators along with schedules go to respondents, put to them the questions from the performer in the order questions are listed and record the replay in the space meant for the same Performa. This method requires the selection and training of enumerators

to fill up the schedules and they should be carefully selected. Enumerators should be intelligent and must be able to find out the truth. The enumerators should be honest sincere and hard working. This method is very useful because it yield good results. Population censes all over the world is conducted through this method.

DIFFERENCES BETWEEN SCHEDULE AND QUESTIONNAIRE

1. The Questionnaire is generally sent through mail to informants. The schedule is generally filled by the research worker.
2. To collect data through questionnaire is relatively cheap. To collect data through schedule is relatively more expensive.
3. Non- response is high in case of questionnaire whereas in schedule response is very high.
4. In Questionnaire there is no personal conducts. But in a schedule there is a face-to face contact.
5. The questionnaire method is used only when respondents are literate.
6. Along with schedules observation methods can be also used.

4.7 BASIC RULES FOR QUESTIONNAIRE ITEM CONSTRUCTION

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

Concerns with questionnaires

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

unlike interviews, the people conducting the research may never know if the respondent understood the question that was being asked. Also, because the questions are so specific to what the researchers are asking, the information gained can be minimal.

Often, questionnaires such as the Myers-Briggs Type Indicator, give too few options to answer; respondents can answer either option but must choose only one response. Questionnaires also produce very low return rates, whether they are mail or online questionnaires. The other problem associated with return rates is that often the people that do return the questionnaire are those that have a really positive or a really negative viewpoint and want their opinion heard. The people that are most likely unbiased either way typically don't respond because it is not worth their time.

4.8 SUMMARY

Data are facts and other relevant materials, past and present, serving as bases for study and analyses. The data needed for a social science research may be broadly classified into (a) Data pertaining to human beings, (b) Data relating to organization and (c) Data pertaining to territorial areas. Personal data or data related to human beings consists of: Demographic and socio-economic characteristics of individuals: Age, sex, race, social class, religion, marital status, education, occupation income, family size, location of the household life style etc.

Data may broadly be divided into two categories, namely **primary data** and **secondary data.** The primary data are those which are collected for the first time by the organisation which is using them. The secondary data, on the other hand, are those which, have already been collected by some other agency but also can be used by the organisation under consideration. Primary data maybe collected by **observation, oral investigation, and questionnaire method** or by **telephone interviews.** Questionnaires may be used for data **collection by interviewers.** They may also be mailed to **prospective respondents.** The drafting of a good questionnaire requires utmost skill. The process of interviewing also requires a great deal of tact, patience and competence to establish rapport with the respondent. Secondary data are available in various published and unpublished documents. The suitability, reliability, adequacy and accuracy of the secondary data should, however, be ensured before they are used for research problems.

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

institutions in the field of data collection, most of the apprehensions in the mind of the researcher have been removed.

4.9 SELF-ASSESSMENT QUESTIONS

1. What are the types of data?

2. What are the primary sources of data?

3. Discuss the methods of collecting primary data.

4. How is personal interview done?

5. How is data collected through mail for doing research?

6. Write a short note on the following:

- **a.** Telephone Method
- **b.** Questionnaire
- **c.** Schedule

7. How is questionnaire and schedule constructed?

8. What are the basic rules for questionnaire item construction?

9. What are the sources of secondary data?

10. How is secondary data useful to researcher?

11. What are the advantages of secondary data?

12. Describe the disadvantages of secondary data.

13. Discuss the methods of collecting secondary data.

14. Point out the difference between primary data and secondary data.

4.10 TEXT AND REFERENCES

- Hague Paul, “Market Research- a guide to planning, methodology & evaluation”; Kogan page, London.
- Khan, J.A.; “Research Methodology”; APH Publishing Corporation, New Delhi.
- Kothari, C.R.; “Research Methodology-Methods & Techniques”; New Age International (P) Limited.
- Kumar Rajendra ; “Research Methodology”; APH Publishing corporation, New Delhi.
- Kumar Ranjit; “Research methodology”; Pearson Education.
- Mehta, J.D. and Gupta, Umesh ; “Research Methods in Management”; Ramesh Book Depot, Jaipur-New Delhi.

- Mertens, Donna, M. ; “Research Methods in Education and Psychology”; Sage Publications, New Delhi.
- Murthy C.; “Research Methodology’; Vrinda Publications (P) Ltd. Delhi.
- Gopal, M.H. 1964. An Introduction to Research Procedure in Social Sciences, Asia Publishing House: Bombay.
- Sadhu, A.N. and A. Singh. 1980. Research Methodology in Social Sciences, Sterling Publishers Private Limited: New Delhi.
- Wilkinson, T.S. and P.L. Bliandarkar. 1979. Methodology and Techniques of Social Research, Himalaya Publishing House: Bombay.

Uttar Pradesh Rajarshi Tandon
Open University

BBA-121

Research Methodology

BLOCK

2

SAMPLING AND SCALING

UNIT-5

SAMPLING

UNIT-6

SCALING

UNIT-7

GRAPHS AND DIAGRAMS

UNIT-8

ADVANCED TECHNIQUES

परिशिष्ट–4

आन्तरिक कवर–दो का प्ररूप

Format of the II Inner Covers

लेखक Dr. Piyali Ghosh, Asst. Professor, School of Management, MNNIT, Allahabad

सम्पादक Prof. H.K. Singh, Professor, Department of Commerce, BHU Varansi.

परिमापक

सहयोगी टीम

संयोजक Dr. Gaurav Sankalp, SoMS, UPRTOU, Allahabd.

प्रूफ रीडर

©UPRTOU, Prayagraj-2020

ISBN : 978-93-83328-54-3

NODE-110/NODE-124/NFT-063

UNIT-5 SAMPLING

Unit Framework

5.1 OBJECTIVE

After reading this unit you will be able to:

- The concept of sampling.
- The merits and demerits of sampling.
- Various probability sampling methods along with their respective merits and demerits.
- Various non-probability sampling methods along with their respective merits and demerits.

5.2 INTRODUCTION : MEANING OF SAMPLING

Data collection stage of any research requires considerable time, effort, and money. If primary data are collected using census method, time and cost increases considerably. Sampling techniques help us in this situation. A true representative sample not only gives accurate results but also saves on time, effort, and money. This chapter is devoted to sampling methods and techniques.

The terminology "sampling" indicates the selection of a part of a group or an aggregate with a view to obtaining information about the whole. This aggregate or the totality of all members is known as Population although they need not be human beings. The selected part, which is used to ascertain the characteristics of the population, is called Sample. While choosing a sample, the population is assumed to be composed of individual units or members, some of which are included in the sample. The total number of members of the population is called Population Size and the number included in the sample is called Sample Size.

Researchers usually cannot make direct observations of every individual in the population they are studying. Instead, they collect data from a subset of individuals – a *sample* – and use those observations to make inferences about the entire population.

Ideally, the sample corresponds to the larger population on the characteristic(s) of interest. In that case, the researcher's conclusions from the sample are probably applicable to the entire population.

This type of correspondence between the sample and the larger population is most important when a researcher wants to know what proportion of the population has a certain characteristic –like a particular opinion or a demographic feature. Public opinion polls that try to describe the percentage of the population that plans to vote for a particular candidate, for example, require a sample that is highly representative of the population.

5.3 NEED OF SAMPLING

To draw conclusions about populations from samples, we must use inferential statistics which enables us to determine a population`s

characteristics by directly observing only a portion (or sample) of the population. We obtain a sample rather than a complete enumeration (a census) of the population for many reasons. Obviously, it is cheaper to observe a part rather than the whole, but we should prepare ourselves to cope with the dangers of using samples. In this tutorial, we will investigate various kinds of sampling procedures. Some are better than others but all may yield samples that are inaccurate and unreliable. We will learn how to minimize these dangers, but some potential error is the price we must pay for the convenience and savings the samples provide.

ESSENTIALS OF SAMPLING :

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

1. **It must be representative:** The sample selected should possess the similar characteristics of the original universe from which it has been drawn.
2. **Homogeneity:** Selected samples from the universe should have similar nature and should not have any difference when compared with the universe.
3. **Adequate samples:** In order to have a more reliable and representative result, a good number of items are to be included in the sample.
4. **Optimization:** All efforts should be made to get maximum results both in terms of cost as well as efficiency. If the size of the sample is larger, there is better efficiency and at the same time the cost is more. A proper size of sample is maintained in order to have optimized results in terms of cost and efficiency.

5.4 ADVANTAGES OF SAMPLING

The sampling only chooses a part of the units from the population for the same study. The sampling has a number of advantages as compared to complete enumeration due to a variety of reasons. Sampling has the following advantages:

1. **Cost effective:** This method is cheaper than the Census Research because only a fraction of the population is studied in this method.
2. **Time saving:** There is saving in time not only in conducting the sampling enquiry but also in the decision making process
3. **Testing of Accuracy:** Testing of accuracy of samples drawn can be made by comparing two or more samples.
4. **Detailed Research is Possible:** Since the data collected under this method is limited but homogeneous, so more time could be spend on decision making.

5. **Reliability:** If samples are taken in proper size and on proper grounds the results of sampling will be almost the same which might have been obtained by Census method.

6. **Exclusive methods in many circumstances**: Where the population is infinite, then the sampling method is the only method of effective research. Also, if the population is perishable or testing units are destructive, then we have to complete our research only through sampling. Example: Estimation of expiry dates of medicines.

7. **Administrative convenience:** The organization and administration of sample survey are easy for the reasons which have been discussed earlier.

8. **More scientific:** Since the methods used to collect data are based on scientific theory and results obtained can be tested, sampling is a more scientific method of collecting data.

5.5 LIMITATIONS OF SAMPLING

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

1. **Biased Conclusion**: If the sample has not been properly taken then the data collected and the decision on such data will lead to wrong conclusion. Samples are like medicines. They can be harmful when they are taken carelessly or without knowledge off their effects.

2. **Experienced Researcher is required**: An efficient sampling requires the services of qualified, skilled and experienced personnel. In the absence of these the results of their search will be biased.

3. **Not suited for Heterogeneous Population**: If the populations are mixed or varied, then this method is not suited for research.

4. **Small Population**: Sampling method is not possible when population size is too small.

5. **Illusory conclusion:** If a sample enquiry is not carefully planned and executed, the conclusions may be inaccurate and misleading.

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

7. **Lack of Experts:** As there are lack of experts to plan and conduct a sample survey, its execution and analysis, and its results would be unsatisfactory and not trustworthy.

8. **Conditions of Complete Coverage:** If the information is required for each and every item of the universe, then a complete enumeration survey is better.

5.6 PROBABILITY AND NON-PROBABILITY SAMPLING

A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.) In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, and Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and

2. Involves random selection at some point.

Non-probability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'under covered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, non-probability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing

limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Non-probability sampling methods include accidental sampling, quota sampling and purposive sampling. In addition, non-response effects may turn *any* probability design into a non-probability design if the characteristics of non-response are not well understood, since non response effectively modifies each element's probability of being sampled.

5.7 SAMPLING TECHNIQUES

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

5.7.1 PROBABILITY OR RANDOM SAMPLING

Probability sampling is based on the theory of probability. It is also known as random sampling. It provides a known nonzero chance of selection for each population element. It is used when generalization is the objective of study, and a greater degree of accuracy of estimation of population parameters is required. The cost and time required is high hence the benefit derived from it should justify the costs.

5.7.2 SIMPLE RANDOM SAMPLING

This sampling technique gives each element an equal and independent chance of being selected. An equal chance means equal probability of selection. An independent chance means that the draw of one element will not affect the chances of other elements being selected.

The procedure of drawing a simple random sample consists of enumeration of all elements in the population.

1. Preparation of a List of all elements, giving them numbers in serial order 1, 2, B, and so on, and
2. Drawing sample numbers by using (a) lottery method, (b) a table of random numbers or (c) a computer.

Suitability: This type of sampling is suited for a small homogeneous population.

5.7.3 STRATIFIED SAMPLING

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub population, out of which individual elements can be randomly selected. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group),

stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

Advantages : Stratified random sampling enhances the representativeness to each sample, gives higher statistical efficiency, easy to carry out, and gives a self-weighing sample.

Disadvantages : A prior knowledge of the composition of the population and the distribution of the population, it is very expensive in time and money and identification of the strata may lead to classification of errors.

5.7.4 SYSTEMATIC SAMPLING

Systematic sampling relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every k^{th} element from then onwards. In this case, k = (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the k^{th} element in the list. A simple example would be to select every 10^{th} name from the telephone directory (an 'every 10^{th}' sample, also referred to as 'sampling with a skip of 10').

Suitability: Systematic selection can be applied to various populations such as students in a class, houses in a street, telephone directory etc.

Advantages: The advantages are it is simpler than random sampling, easy to use, easy to instruct, requires less time, it's cheaper, easier to check, sample is spread evenly over the population, and it is statistically more efficient.

Disadvantages: The disadvantages are it ignores all elements between two k^{th} elements selected, each element does not have equal chance of being selected, and this method sometimes gives a biased sample.

5.7.5 CLUSTER SAMPLING

Sometimes it is more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.) For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

Suitability: The application of cluster sampling is extensive in farm management surveys, socio-economic surveys, rural credit surveys, demographic studies, ecological studies, public opinion polls, and large scale surveys of political and social behaviour, attitude surveys and so on.

Advantages: The advantages of this method is it is easier and more convenient, cost of this is much less, promotes the convenience of field work as it could be done in compact places, it does not require more time, units of study can be readily substituted for other units and it is more flexible.

Disadvantages: The cluster sizes may vary and this variation could increase the bias of the resulting sample. The sampling error in this method of sampling is greater and the adjacent units of study tend to have more similar characteristics than do units distantly apart.

5.7.6 AREA SAMPLING

This is an important form of cluster sampling. In larger field surveys cluster consisting of specific geographical areas like districts, talluks, villages or blocks in a city are randomly drawn. As the geographical areas are selected as sampling units in such cases, their sampling is called area sampling. It is not a separate method of sampling, but forms part of cluster sampling.

5.7.7 PROBABILITY-PROPORTIONAL-TO-SIZE SAMPLING

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1.

In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490students respectively (total 1500 students), and we want to use student population as the basis fora PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school(1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of

500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

Advantages: The advantages are clusters of various sizes get proportionate representation, PPS leads to greater precision than would a simple random sample of clusters and a constant sampling fraction at the second stage, equal-sized samples from each selected primary cluster are convenient for field work.

Disadvantages: PPS cannot be used if the sizes of the primary sampling clusters are not known.

5.7.8 DOUBLE SAMPLING AND MULTIPHASE SAMPLING

Double sampling refers to the subsection of the final sample form a preselected larger sample that provided information for improving the final selection. When the procedure is extended to more than two phases of selection, it is then, called multi-phase sampling. This is also known as sequential sampling, as sub-sampling is done from a main sample in phases. Double sampling or multiphase sampling is a compromise solution for a dilemma posed by undesirable extremes. “The statistics based on the sample of ‘n’ can be improved by using ancillary information from a wide base: but this is too costly to obtain from the entire population of N elements. Instead, information is obtained from a larger preliminary sample n_L which includes the final sample n.

5.7.9 NON-PROBABILITY OR NON RANDOM SAMPLING

Non-probability sampling or non-random sampling is not based on the theory of probability. This sampling does not provide a chance of selection to each population element.

Advantages: The only merits of this type of sampling are simplicity, convenience and low cost.

Disadvantages: The demerits are it does not ensure a selection chance to each population unit. The selection probability sample may not be a representative one. The selection probability is unknown. It suffers from sampling bias which will distort results.

5.7.10 QUOTA SAMPLING

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just a sin stratified sampling. Then judgement is used to select the subjects or units from each segment based on a specified

proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for several years.

Suitability: It is used in studies like marketing surveys, opinion polls, and readership surveys which do not aim at precision, but to get quickly some crude results.

Advantage: It is less costly, takes less time, non-need for a list of population, and field work can easily be organized.

Disadvantage: It is impossible to estimate sampling error, strict control if field work is difficult, and subject to a higher degree of classification.

5.7.11 CONVENIENCE OR ACCIDENTAL SAMPLING

Accidental sampling (sometimes known as **grab**, **convenience** or **opportunity sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey were to be conducted at different times of day and several times per week.

Suitability: Though this type of sampling has no status, it may be used for simple purposes such as testing ideas or gaining ideas or rough impression about a subject of interest.

Advantage: It is the cheapest and simplest, it does not require a list of population and it does not require any statistical expertise.

Disadvantage: The disadvantage is that it is highly biased because of researcher's subjectivity, it is the least reliable sampling method and the findings cannot be generalized.

5.7.12 PURPOSIVE (OR JUDGMENT) SAMPLING

This method means deliberate selection of sample units that conform to some pre-determined criteria. This is also known as judgment sampling. This involves selection of cases which we judge as the most appropriate ones for the given study. It is based on the judgment of the researcher or some expert. It does not aim at securing a cross section of a population. The chance that a particular case be selected for the sample depends on the subjective judgment of the researcher.

Suitability: This is used when what is important is the typicality and specific relevance of the sampling units to the study and not their overall representativeness to the population.

Advantage: It is less costly and more convenient and guarantees inclusion of relevant elements in the sample.

Disadvantage: It is less efficient for generalizing, does not ensure the representativeness, requires more prior extensive.

5.7.13 SNOW-BALL SAMPLING

This is the colourful name for a technique of Building up a list or a sample of a special population by using an initial set of its members as informants. This sampling technique may also be used in socio-metric studies.

Suitability: It is very useful in studying social groups, informal groups in a formal organization, and diffusion of information among professional of various kinds.

Advantage: It is useful for smaller populations for which no frames are readily available.

Disadvantage: The disadvantage is that it does not allow the use of probability statistical methods. It is difficult to apply when the population is large. It does not ensure the inclusion of all the elements in the list.

5.8 SUMMARY

A statistical sample ideally purports to be a miniature model or replica of the collectivity or the population. Sampling helps in time and cost saving. If the population to be studied is quite large, sampling is warranted. However, the size is a relative matter. The decision regarding census or sampling depends upon the budget of the study. Sampling is opted when the amount of money budgeted is smaller than the anticipated cost of census survey.

Sampling techniques help us in this situation. Sampling method is the only method that can be used in certain cases. There are some cases in which the census method is inapplicable and the only practicable means is

provided by the sample method. Despite various advantages of sampling, it is not completely free from limitations. A sample survey must be carefully planned and executed otherwise the results obtained may be inaccurate and misleading. Even if a complete count care is taken still serious errors may arise in sampling, if the sampling procedure is not perfect. Sampling generally requires the services of experts, even only for consultation purposes. In the absence of qualified and experienced persons, the information obtained from sample surveys cannot be relied upon. Shortage of experts in the sampling field is a serious hurdle in the way of reliable statistics. Sampling techniques may be classified into two broad categories namely probability and non-probability sampling. Non-probability sampling methods are those which do not provide every item in the universe with a known chance of being included in the sample. These include judgment, quota, cluster and convenience sampling techniques.

5.9 SELF-ASSESSMENT QUESTIONS

1. What do you mean by Sampling? State the purpose of sampling.
2. Define Probability and non-probability sampling
3. Elaborate and discuss the sampling techniques.
4. What are the advantages and disadvantages of stratified sampling?
5. Define:
    a. Cluster sampling
    b. Quota sampling
    c. Accidental sampling
6. Discuss the advantages and limitations of sampling techniques.
7. “Sampling is necessary under certain conditions”. Explain this with suitable examples.
8. Critically examine the various probability sampling methods.
9. What are the methods of sampling? How do you select a sample to study consumer behaviour?
10. Distinguish between random sampling, purposive sampling and stratified sampling. How is a random sample obtained?

5.10 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, “Business Research Methods”, Tata McGraw Hill Inc.
- Alan Bryman, and Emma Bell, “Business research Methods”. Oxford University Press Publication.

- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- S.P. Gupta, “*Statistical methods*” Sultan Chand & Sons publication, New Delhi
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Naresh Malhotra, “Marketing Research: An applied Orientation”, Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, “Methodology and techniques of Social Research”, Himalaya Publishing House, New Delhi.

UNIT-6 SCALING

Unit Framework

6.1 OBJECTIVE

After studying this unit, you should be able to understand:

- The concept of scales of measurement.
- Various types of scales of measurement.

- Common type of measurement scales that are used to measure attitudes.
- Application of scales of measurement.
- Measurement scales and statistical tests suitable to them.
- Some important measurement errors.

6.2 INTRODUCTION : MEANING OF SCALE

Scale analysis is a set of methods to analyze survey data, in which responses to questions are combined to measure a latent variable. These items can be dichotomous (e.g. yes/no, agree/disagree, correct/incorrect) or polytomous (e.g. disagree strongly /disagree /neutral /agree/agree strongly). Any measurement for such data is required to be reliable, valid, and homogeneous with comparable results over different studies It is nothing but a standardized process of assigning numbers or symbols to certain characteristics of the objects of interest, according to some predefined rules. Measurement actually is a pre-requisite to any mathematical or statistical analysis of data. Measurement scales are used to categorize and/or quantify variables and help in collection and analysis phase of any research. This chapter describes the four scales of measurement that are commonly used in statistical analysis: nominal, ordinal, interval, and ratio scales.

6.3 MEASUREMENT OF SCALE

Measurement scales are used to categorize and/or quantify variables. This lesson describes the four scales of measurement that are commonly used in statistical analysis: nominal, ordinal, interval, and ratio scales.

Properties of Measurement Scales

Each scale of measurement satisfies one or more of the following properties of measurement.

- **Identity:** Each value on the measurement scale has a unique meaning.
- **Magnitude**: Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- **Equal intervals**: Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
- **A minimum value of zero**: The scale has a true zero point, below which no values exist.

6.4 NOMINAL SCALE OF MEASUREMENT

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude.

Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

6.5 ORDINAL SCALE OF MEASUREMENT

The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale.

An example of an ordinal scale in action would be the results of a horse race, reported as "win", "place", and "show". We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a mile.

6.6 INTERVAL SCALE OF MEASUREMENT

The interval scale of measurement has the properties of identity, magnitude, and equal intervals. A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.

With an interval scale, you know not only whether different values are bigger or smaller, you also know *how much* bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday; you also know that it was 10 degrees hotter.

6.7 RATIO SCALE OF MEASUREMENT

The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and a minimum value of zero. The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and the scale has a minimum value of zero.

Weight scales have a minimum value of zero because objects at rest can be weightless, but they cannot have negative weight.

6.8 SCALING TECHNIQUES

6.8.1 LIKERT SCALE

A **Likert scale** is a psychometric scale commonly involved in research that employs questionnaires. It is the most widely used approach to scaling responses in survey research, such that the term is often used interchangeably with *rating scale*, or more accurately the **Likert-type scale**, even though the two are not synonymous. The scale is named after its inventor, psychologist Rensis Likert. Likert distinguished between a scale proper, which emerges from collective responses to a set of items (usually eight or more), and the format in which responses are scored along a range. Technically speaking, a Likert scale refers only to the former. The difference between these two concepts has to do with the distinction Likert made between the underlying phenomenon being investigated and the means of capturing variation that point to the underlying phenomenon. When responding to a Likert questionnaire item, respondents specify their level of agreement or disagreement on a symmetric agree disagree scale for a series of statements. Thus, the range captures the intensity of their feelings for a given item. A scale can be created as the simple sum questionnaire responses over the full range of the scale. In so doing, Likert scaling assumes that distances on each item are equal. Importantly, "*All items are assumed to be replications of each other or in other words items are considered to be parallel instruments*"

6.8.2 SAMPLE QUESTION PRESENTED USING A FIVE-POINT LIKERT ITEM

An important distinction must be made between a *Likert scale* and a *Likert item*. The Likert scale is the sum of responses on several Likert items. Because Likert items are often accompanied by a visual analog scale (e.g., a horizontal line, on which a subject indicates his or her response by circling or checking tick-marks), the items are sometimes called scales themselves. This is the source of much confusion; it is better, therefore, to reserve the term *Likert scale* to apply to the summed scale, and *Likert item* to refer to an individual item.

A Likert item is simply a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured. It is considered symmetric or "balanced" because there are equal amounts of positive and negative positions. Often five ordered response levels are used, although many psychometricians advocate using seven or nine levels; a recent empirical study found that a 5- or 7- point scale may produce slightly higher mean scores relative to the highest possible

attainable score, compared to those produced from a 10-point scale, and this difference was statistically significant. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis.

The format of a typical five-level Likert item, for example, could be:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Likert scaling is a bipolar scaling method, measuring either positive or negative response to a statement. Sometimes an even-point scale is used, where the middle option of "Neither agree nor disagree" is not available. This is sometimes called a "forced choice" method, since the neutral option is removed. The neutral option can be seen as an easy option to take when a respondent is unsure, and so whether it is a true neutral option is questionable. A 1987 study found negligible differences between the use of "undecided" and "neutral" as the middle option in a 5-point Likert scale.

Likert scales may be subject to distortion from several causes. Respondents may avoid using extreme response categories (*central tendency bias*); agree with statements as presented (*acquiescence bias*); or try to portray themselves or their organization in a more favorable light (*social desirability bias*). Designing a scale with balanced keying (an equal number of positive and negative statements) can obviate the problem of acquiescence bias, since acquiescence on positively keyed items will balance acquiescence on negatively keyed items, but central tendency and social desirability are somewhat more problematic.

6.8.3 MULTIDIMENSIONAL SCALING

Multidimensional scaling (**MDS**) is a means of visualizing the level of similarly of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. An MDS algorithm aims to place each object in *N*-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the *N* dimensions. Unlike principal component analysis wherein most of the variance in the data is captured in the first axis with each subsequent axis containing progressively less information, axes in MDS are arbitrary and distance units along each axis do not reflect equal quantitative distances at other sections of the same axis. The number of dimensions of an MDS plot *N*

can exceed 2 and are specified a priori. Choosing *N*=2 optimizes the object locations for a two dimensional scatterplot.

TYPES OF MULTIDIMENSIONAL SCALING

1. **Classical multidimensional scaling:** Also known as Principal Coordinates Analysis, Torgerson Scaling or Torgerson–Gower scaling. Takes an input matrix giving dissimilarities between pairs of items and outputs a coordinate matrix whose configuration minimizes a loss function called *strain*.

2. **Metric multidimensional scaling:** A superset of classical MDS that generalizes the optimization procedure to a variety of loss functions and input matrices of known distances with weights and so on. A useful loss function in this context is called *stress*, which is often minimized using a procedure called stress majorization.

3. **Non-metric multidimensional scaling:** In contrast to metric MDS, non-metric MDS finds both a non-parametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items, and the location of each item in the low-dimensional space. The relationship is typically found using isotonic regression. Louis Guttman's smallest space analysis (SSA) is an example of a non-metric MDS procedure.

4. **Generalized multidimensional scaling:** An extension of metric multidimensional scaling, in which the target space is an arbitrary smooth non-Euclidean space. In cases where the dissimilarities are distances on a surface and the target space is another surface, GMDS allows finding the minimum-distortion embedding of one surface into another.

6.8.4 SELECTION OF APPROPRIATE ATTITUDE MEASUREMENT SCALE

Now that we have looked at a number of attitude measurement scales, a natural question arises: “Which is most appropriate?” As in the selection of a basic research design, there is no single best answer for all research projects. The answer to this question is relative, and the choice of scale will depend on the nature of the attitudinal object to be measured, the manager’s problem definition, and the backward and forward linkages to choices already made (for example, telephone survey versus mail survey). However, several questions will help focus the choice of a measurement scale:

- Is a ranking, sorting, rating, or choice technique best?
- Should a monadic or a comparative scale be used?
- What type of category labels, if any, will be used for the rating scale?

- How many scale categories or response positions are needed to accurately measure an attitude?
- Should a balanced or unbalanced rating scale be chosen?
- Should a scale that forces a choice among predetermined options be used?
- Should a single measure or an index measure be used?

6.8.5 RANKING, SORTING, RATING, OR CHOICE TECHNIQUE

The decision whether to use ranking, sorting, rating, or a choice technique is determined largely by the problem definition and especially by the type of statistical analysis desired. For example, ranking provides only ordinal data, limiting the statistical techniques that may be used.

Monadic or Comparative Scale

If the scale to be used is not a ratio scale, the researcher must decide whether to include a standard of comparison in the verbal portion of the scale. Consider the following rating scale:

Now that you've had your automobile for about one year, please tell us how satisfied you are with its engine power and pickup.

Completely Dissatisfied	Dissatisfied	Somewhat	Satisfied	Completely Satisfied
[ ]	[ ]	[ ]	[ ]	[ ]

This is a monadic rating scale, because it asks about a single concept (the brand of automobile the individual actually purchased) in isolation. The respondent is not given a specific frame of reference. A comparative rating scale asks a respondent to rate a concept, such as a specific amount of responsibility or authority, in comparison with a benchmark—perhaps another similar concept—explicitly used as a frame of reference. In many cases, the comparative rating scale presents an ideal situation as a reference point for comparison with the actual situation. For example:

Please indicate how the amount of authority in your present position compares with the amount of authority that would be ideal for this position:

Too Much [ ] About Right [ ] Too Little [ ] What Type of Category Labels, If Any?

We have discussed verbal labels, numerical labels, and unlisted choices. Many rating scales have verbal labels for response categories because researchers believe they help respondents better understand the response positions. The maturity and educational levels of the respondents will influence this decision. The semantic differential, with unlabeled response categories between two bipolar adjectives, and the numerical scale, with numbers to indicate scale positions, often are selected because the researcher wishes to assume interval-scale data.

How Many Scale Categories or Response Positions?

Should a category scale have four, five, or seven response positions or categories? Or should the researcher use a graphic scale with an infinite number of positions? The original developmental research on the semantic differential indicated that five to eight points is optimal. However, the researcher must determine the number of meaningful positions that is best for the specific project. This issue of identifying how many meaningful distinctions respondents can practically make is basically a matter of sensitivity, but at the operational rather than the conceptual level.

6.8.6 BALANCED OR UNBALANCED RATING SCALE

The fixed-alternative format may be balanced or unbalanced. For example, the following question, which asks about parent-child decisions relating to television program watching, is a **balanced rating scale**:

Who decides which television programs your children watch?

- Child decides all of the time. [ ]
- Child decides most of the time. [ ]
- Child and parent decide together. [ ]
- Parent decides most of the time. [ ]
- Parent decides all of the time. [ ]

This scale is balanced because a neutral point, or point of indifference, is at the center of the scale.

Unbalanced rating scales may be used when responses are expected to be distributed at one end of the scale. Unbalanced scales, such as the following one, may eliminate this type of “end piling”:

Completely Dissatisfied	Dissatisfied	Somewhat	Satisfied	Completely Satisfied
[ ]	[ ]	[ ]	[ ]	[ ]

Notice that there are three "satisfied" responses and only two "dissatisfied" responses above. The choice of a balanced or unbalanced scale generally depends on the nature of the concept or the researcher's knowledge about attitudes toward the stimulus to be measured.

6.8.7 USE A SCALE THAT FORCES A CHOICE AMONG PREDETERMINED OPTIONS

In many situations, a respondent has not formed an attitude toward the concept being studied and simply cannot provide an answer. If a forced-choice rating scale compels the respondent to answer, the response is merely a function of the question. If answers are not forced, the midpoint of the scale may be used by the respondent to indicate unawareness as well as indifference. If many respondents in the sample are expected to be unaware of the attitudinal object under investigation, this problem may be eliminated by using a non-forced-choice scale that provides a "no opinion" category, as in the following example:

- How does the Bank of Commerce compare with the First National Bank? [ ]
- Bank of Commerce is better than First National Bank [ ]
- Bank of Commerce is about the same as First National Bank [ ]
- Bank of Commerce is worse than First National Bank [ ]
- Can't say [ ]

Asking this type of question allows the investigator to separate respondents who cannot make an honest comparison from respondents who have had experience with both banks. The argument for forced choice is that people really do have attitudes, even if they are unfamiliar with the banks, and should be required to answer the question. Still, the use of forced-choice questions is associated with higher incidences of "no answer." Internet surveys make forced-choice questions easy to implement because the delivery can be set up so that a respondent cannot go to the next question until the previous question is answered. Realize, however, if a respondent truly has no opinion, and the no opinion option is not included, he or she may simply quit responding to the questionnaire.

6.8.8 SINGLE MEASURE OR AN INDEX MEASURE

Whether to use a single measure or an index measure depends on the complexity of the issue to be investigated, the number of dimensions the issue contains, and whether individual attributes of the stimulus are part of a holistic attitude or are seen as separate items. Very simple concepts that do not vary from context to context can be measured by single items. However, most psychological concepts are more complex and require multiple-item measurement. Additionally, multiple-item measures are easier to test for construct validity. The researcher's conceptual definition will be helpful in making this choice.

The researcher has many scaling options. Generally, the choice is influenced by plans for the later stages of the research project. Again, problem definition becomes a determining factor influencing the research design.

6.9 SUMMARY

Scaling is nothing but a standardized process of assigning numbers or symbols to certain characteristics of the objects of interest, according to some pre-defined rules. Scales are used to categorize and/or quantify variables and help in collection and analysis phase of any research. They are used for measuring qualitative responses of respondents such as those related to their feelings, perception, likes, dislikes, interests and preferences. It involves creating a continuum upon which measured objects are located.

They are of four types: nominal, ordinal, interval, and ratio. The rules for assigning numbers constitute the essential criteria for defining each scale. As we move from nominal to ratio scales, we must meet increasingly restrictive rules. A brief discussion of multidimensional scaling followed. Finally, the issues of the selection of an appropriate attitude measurement scale and the limitations of these research tools were discussed. As the rules become more restrictive, the kinds of arithmetic operations for which the numbers can be used are increased.

6.10 SELF-ASSESSMENT QUESTIONS

1. What do you mean by scale?
2. Explain the following:
    a) Nominal Scale of Measurement
    b) Ordinal Scale of Measurement
    c) Interval Scale of Measurement
    d) Ratio Scale of Measurement

3. Explain scaling techniques.
4. How is appropriate attitude measurement scale selected?
5. How is scaling ranked and rated?
6. What is single measure or an index measure?
7. Discuss various types of scales of measurement highlighting their respective usage.
8. Differentiate between ordinal and interval scales citing suitable examples.
9. Write short notes on the following:
    a) Likert scale
    b) Multidimensional scale
    c) Balanced or Unbalanced Rating Scale
10. Discuss various scales that may be used to measure attitudes.

6.11 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, "Business Research Methods", McGraw Hill Inc.
- Alan Bryman, and Emma Bell, "Business research Methods". Oxford University Press Publication.
- Saunders, "Research Methods for Business Students", Pearsons Education Publications.
- Naresh Malhotra, "Marketing Research: An applied Orientation", Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, "Methodology and techniques of Social Research", Himalaya Publishing House, New Delhi.
- Donald S. Tull and Del S. Hawkins, "Marketing Research – Measurement and Method", Prentice Hall of India Pvt. Ltd., New Delhi.
- Churchill, Gilbert A., 1983, Marketing Research: Methodological Foundations, The Dryden Press, New York.
- Nunally, Jum C., 1978, Psychometric Theory, Tata McGraw-Hill, New Delhi. Feber, Robert 1974, Handbook of Marketing Research, McGraw-Hill, New York.
- Meister, David, 1985, Behavioural Analysis and Measurement Methods, John Wiley, New York.

- Rodger, Lesile W., 1984, Statistics for Marketing, McGraw-Hill (UK), London.
- Lundstrom, William J.;et. al. November 1976. "The development of a scale to measure consumer discontent", Journal of Marketing Research, Vol. 13, pp. 373-381.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- S.P. Gupta, “Statistical methods” Sultan Chand & Sons publication, New Delhi.

UNIT-7 GRAPHS AND DIAGRAMS

Unit Framework

7.1 OBJECTIVE

After studying this unit you should be able to understand:

- Checking for process analysis

- Editing
- Coding
- Classification
- Tabulation
- Construction of Frequency Table
- Components of a table
- Graphs and diagrams
- Types of graphs and general rules
- Time Series Graphs
- Constructing a Time Series Graph

7.2 INTRODUCTION : DATA PROCESSING

This data preparation for research analysis is teamed as processing of data. Further selections of tools for analysis would to a large extent depend on the results of this data processing. Data processing is an intermediary stage of work between data collections and data interpretation. The data gathered in the form of questionnaires/interview schedules/field notes/data sheets is mostly in the form of a large volume of research variables. The research variables recognized is the result of the preliminary research plan, which also sets out the data processing methods beforehand. Processing of data requires advanced planning and this planning may cover such aspects as identification of variables, hypothetical relationship among the variables and the tentative research hypothesis.

The various steps in processing of data may be stated as:

7.2.1 Editing

7.2.2 Coding

7.2.3 Classification

7.2.4 Tabulation

7.2.1 EDITING

Editing of data is a process of examining the collected raw data (specially in surveys) to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and/or schedules. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly

entered, as completed as possible and have been well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and central editing. Field editing consists in the review of the reporting forms by the investigator for completing (translating or rewriting) what the latter has written in abbreviated and/or in illegible form at the time of recording the respondents' responses. This type of editing is necessary in view of the fact that individual writing styles often can be difficult for others to decipher. This sort of editing should be done as soon as possible after the 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 or 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 large inquiry. Editor(s) may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule.

A t times, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final results, especially in the context of mail surveys.

Editors must keep in view several points while performing their work: They should be familiar with instructions given to the interviewers and coders as well as with the editing instructions supplied to them for the purpose. While crossing out an original entry for one reason or another, they should just draw a single line on it so that the same may remain legible. They must make entries (if any) on the form in some distinctive colour and that too in a standardised form. They should initial all answers which they change or supply. Editor's initials and the date of editing should be placed on each completed form or schedule.

7.2.2 CODING

Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to be

observed is that of uni-dimensionality by which is meant that every class is defined in terms of only one concept.

Coding is necessary 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 usually be taken at the designing stage of the questionnaire. This makes it possible to pre-code the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch from the original questionnaires. But in case of hand coding some standard method may be used. One such standard method is to code in the margin with a coloured pencil. The other method can be to transcribe the data from the questionnaire to a coding sheet. Whatever method is adopted, one should see that coding errors are altogether eliminated or reduced to the minimum level.

7.2.3 CLASSIFICATION

Most research studies result in a large volume of raw data which must be reduced into homogeneous 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 characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes. Classification can be one of the following two types, depending upon the nature of the phenomenon involved:

Classification according to attributes: As stated above, data are classified on the basis of common characteristics which can either be descriptive (such as literacy, sex, honesty, etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively; only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are known as statistics of attributes and their classification is said to be classification according to attributes.

Such classification can be simple classification or manifold classification. In simple classification we consider only one attribute and divide the universe into two classes— one class consisting of items possessing the given attribute and the other class consisting of items which do not possess the given attribute. But in manifold classification we consider two or more attributes simultaneously, and divide that data into a number of classes (total number of classes of final order is given by 2n, where n = number of attributes considered). Whenever data are classified according to attributes, the researcher must see that the attributes are defined in such a manner that there is least possibility of any doubt/ambiguity concerning the said attributes.

Classification according to 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 category. Such data are known as statistics of variables and are classified on the basis of class intervals. For instance, persons whose incomes, say, are within Rs 201 to Rs 400 can form one group; those whose incomes are within Rs 401 to Rs 600 can form another group and so on. In this way the entire data may be divided into a number of groups or classes or what are usually called, 'class-intervals.' Each group of class interval, thus, has an upper limit as well as a lower limit which are known as class limits.

The difference between the two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution.

Classification according to class intervals usually involves the following three main problems:

1. How many classes should be there?
2. What should be their magnitudes?
3. There can be no specific answer with regard to the number of classes.

The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class-intervals should be of equal magnitudes, but in some cases unequal magnitudes may result in better classification.

Hence researcher's objective judgement plays an important part in this connection. Multiples of 2, 5 and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by **H.A. Sturges**, determining the size of class interval:

$i = R/(1 + 3.3 \log N)$ where

i = size of class interval;

R = Range (i.e., difference between the values of the largest item and smallest item among the given items);

N = Number of items to be grouped.

It should also be kept in mind that in case one or two or very few items have very high or very low values, one may use what are known as open-

ended intervals in the overall frequency distribution. Such intervals may be expressed like under Rs 500 or Rs 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researcher must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are to be classified.

How to choose class limits?

While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by 2) of a class-interval and the actual average of items of that class interval should remain as close to each other as possible. Consistent with this, the class limits should be located at multiples of 2, 5, 10, 20, 100 and such other figures. Class limits may generally be stated in any of the following forms:

Exclusive type class intervals: They are usually stated as follows:

10–20

20–30

30–40

40–50

The above intervals should be read as under:

10 and under 20

20 and under 30

30 and under 40

40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to the upper limit of a class are grouped in the next higher class. For example, an item whose value is exactly 30 would be put in 30–40 class intervals and not in 20–30 class intervals. In simple words, we can say that under exclusive type class intervals, the upper limit of a class interval is excluded and items with values less than the upper limit (but not less than the lower limit) are put in the given class interval.

Inclusive type class intervals: They are usually stated as follows:

11–20

21–30

31–40

41–50

In inclusive type class intervals the upper limit of a class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11–20 class intervals.

The stated upper limit of the class interval 11–20 is 20 but the real limit is 20.99999 and as such 11–20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be a discrete one (i.e., can be measured and stated only in integers), then we should adopt inclusive type classification. But when the phenomenon happens to be a continuous one capable of being measured in fractions as well, we can use exclusive type class intervals.

How to determine the frequency of each class?

This can be done either by tally sheets or by mechanical aids. Under the technique of tally sheet, the class-groups are written on a sheet of paper (commonly known as the tally sheet) and for each item a stroke (usually a small vertical line) is marked against the class group in which it falls. The general practice is that after every four small vertical lines in a class group, the fifth line for the item falling in the same group is indicated as horizontal line through the said four lines and the resulting flower (IIII) represents five items. All this facilitates the counting of items in each one of the class groups. Alternatively, class frequencies can be determined, especially in case of large inquires and surveys, by mechanical aids i.e., with the help of machines viz., sorting machines that are available for the purpose. Some machines are hand operated, whereas other work with electricity. There are machines which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

7.2.4 TABULATION

When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarizing raw data and displaying the same in compact form (i.e., in the form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows. Tabulation is essential 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 mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulating machines or computers. In

relatively large inquiries, we may use mechanical or computer tabulation if other factors are favourable and necessary facilities are available. Hand tabulation is usually preferred in case of small inquiries where the number of questionnaires is small and they are of relatively short length. Hand tabulation may be done using the direct tally, the list and tally or the card sort and count methods. When there are simple codes, it is feasible to tally directly from the questionnaire. Under this method, the codes are written on a sheet of paper, called tally sheet, and for each response a stroke is marked against the code in which it falls. Usually after every four strokes against a particular code, the fifth response is indicated by drawing a diagonal or horizontal line through the strokes. These groups of five are easy to count and the data are sorted against each code conveniently. In the listing method, the code responses may be transcribed onto a large work-sheet, allowing a line for each questionnaire. This way a large number of questionnaires can be listed on one work sheet. Tallies are then made for each question. The card sorting method is the most flexible hand tabulation. In this method the data are recorded on special cards of convenient size and shape with a series of holes. Each hole stands for a code and when cards are stacked, a needle passes through particular hole representing a particular code. These cards are then separated and counted. In this way frequencies of various codes can be found out by the repetition of this technique. We can as well use the mechanical devices or the computer facility for tabulation purpose in case we want quick results, our budget permits their use and we have a large volume of straight forward tabulation involving a number of cross-breaks.

7.3 RULES FOR CONSTRUCTING DIAGRAMS

1. The diagrams should be simple.

2. Each diagram must be given a clear, concise and suitable title without damaging clarity.

3. A proper proportion between height and width must be maintained in order to avoid an unpleasant look.

4. Select a proper scale; it should be in even numbers or in multiples of five or ten. e.g. 25,50, 75 or 10, 20, 30, 40, ... etc. But there are no fixed rules.

5. In order to clear certain points, always put footnotes.

6. An index, explaining different lines, shades and colors should be given.

7. Diagrams should be absolutely neat and clean.

7.4 ADVANTAGES AND DISADVANTAGES OF DIAGRAMS

ADVANTAGES

1. Quick way for the audience to visualize what you are saying -- numbers, trends, up or down
2. Forceful -- emphasizes main point
3. Convincing -- proves a point, see and hear
4. Compact way to convey information
5. More interesting than just talk or print (Remember to use as many of the five senses as possible)

DISADVANTAGES

1. Time consuming to make -- decisions must be made in advance for layout, color, materials, etc.
2. Technical in nature -- audience knowledge to interpret, or understand
3. Costly -- depending on the medium used (poster board, transfer letters, etc.)

7.5 PICTOGRAM

A **pictogram**, also called a **pictogramme** or **pictograph**, is an ideogram that conveys its meaning through its pictorial resemblance to a physical object. Pictographs are often used in writing and graphic systems in which the characters are to a considerable extent pictorial in appearance.

7.6 HISTOGRAM

A **histogram** is a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson. A histogram is a representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval.

The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size. The rectangles of a

histogram are drawn so that they touch each other to indicate that the original variable is continuous.

Histograms are used to plot the density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the lengths of the intervals on the x-axis are all 1, then a histogram is identical to a relative frequency plot. An alternative to the histogram is kernel density estimation, which uses a kernel to smooth samples. This will construct a smooth probability density function, which will in general more accurately reflect the underlying variable.

7.7 GRAPHS VS DIAGRAMS

Sometimes, the information purported to be understood is too long and complex. To make it interesting and understandable in an exciting manner, different visual representations are used.

Graphs and diagrams are two of the common means to visually represent information that is either repetitive in nature or too complex. There are similarities in these techniques that confuse many to treat them as similar.

7.7.1 DIAGRAMS

We are too well aware of the use of diagrams to explain information and facts that are presented in the form of text. If you need to explain the parts of a machine or the principle of its working, it becomes difficult to make one understand the concept through text only. This is where diagrams in the form of sketches come into play. Similarly, diagrams are made heavy use of in biology where students have to learn about different body parts and their functions. Visual representation of concepts through diagrams has better chances of retention in the memory of students than presenting them in the form of text.

Diagrams are resorted to right from the time a kid enters a school as even alphabets are presented to him in a more interesting and attractive manner with the help of diagrams.

7.7.2 GRAPHS

Whenever there are two variables in a set of information, it is better to present the information using graphs as it makes it easier to understand the data. For example, if one is trying to show how the prices of commodities have increased with respect to time, a simple line graph would be a more effective and interesting way rather than putting all this information in the form of text which is hard to remember whereas even a layman can see how prices have gone up or down in relation to time.

Graphs make use of graph paper which has precise squares and presents the information in an accurate manner and the reader can see the effect of one variable on another in a very simple manner.

7.7.3 DIFFERENCE BETWEEN GRAPHS AND DIAGRAMS

1. All graphs are a diagram but not all diagrams are graph. This means that diagram is only a subset of graph.
2. Graph is a representation of information using lines on two or three axes such as x, y, and z, whereas diagram is a simple pictorial representation of what a thing looks like or how it works.
3. Graphs are representations to a scale whereas diagrams need not be to a scale
4. Diagrams are more attractive to look at which is why they are used in publicity whereas graphs are for the use of statisticians and researchers.
5. Values of mean and median can be calculated through graphs which is not possible with diagrams
6. Graphs are drawn on graph paper whereas diagrams do not need a graph paper
7. For frequency distribution, only graphs are used and it cannot be represented through diagrams

7.8 TYPES OF GRAPHS

One goal of statistics is to present data in a meaningful way. It's one thing to see a list of data on a page; it's another to understand the trends and details of the data. Many times data sets involve millions (if not billions) of data values. This is far too many to print out in a journal article or sidebar of a magazine story. One effective tool in the statistician's toolbox is to depict data by the use of a graph.

They say a picture is worth a thousand words. The same thing could be said about a graph. Good graphs convey information quickly and easily to the user. Graphs highlight salient features of the data. They can show relationships that are not obvious from studying a list of numbers. Graphs can also provide a convenient way to compare different sets of data.

7.8.1 LIST OF COMMON GRAPHS IN STATISTICS

Different situations call for different types of graphs, and it helps to have a good knowledge of what graphs are available. Many times the type of data determines what graph is appropriate to use. Qualitative data, quantitative data and paired data each use different types of graphs.

Seven of the most common graphs in statistics are listed below:

1. **Pareto Diagram or Bar Graph** - A bar graph contains a bar for each category of a set of qualitative data. The bars are arranged in order of frequency, so that more important categories are emphasized.

2. **Pie Chart or Circle Graph** - A pie chart displays qualitative data in the form of a pie. Each slice of pie represents a different category.

3. **Histogram -** A histogram in another kind of graph that uses bars in its display. This type of graph is used with quantitative data. Ranges of values, called classes, are listed at the bottom, and the classes with greater frequencies have taller bars.

4. **Stem and Left Plot** - A stem and left plot breaks each value of a quantitative data set into two pieces, a stem, typically for the highest place value, and a leaf for the other place values. It provides a way to list all data values in a compact form.

5. **Dot Plot -** A dot plot is a hybrid between a histogram and a stem and leaf plot. Each quantitative data value becomes a dot or point that is placed above the appropriate class values.

6. **Scatter Plots** - A scatter plot displays data that is paired by using a horizontal axis (the *x* axis), and a vertical axis (the *y* axis). The statistical tools of correlation and regression are then used to show trends on the scatterplot.

7. **Time-Series Graphs** - A time-series graph displays data at different points in time, so it is another kind of graph to be used for certain kinds of paired data. The horizontal axis shows the time and the vertical axis is for the data values. These kinds of graphs can be used to show trends as time progresses.

7.9 TIME SERIES GRAPHS

Suppose that we want to study the climate of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. But all of these methods ignore a portion of the data that we have collected.

The feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don‘t have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that

recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

7.9.1 CONSTRUCTING A TIME SERIES GRAPH

To construct a time series graph, we must look at both pieces of our paired data set. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values variable that we are measuring. By doing this each point on the graph corresponds to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

7.9.2 USES OF A TIME SERIES GRAPH

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot. These trends are important as they can be used to project into the future.

In addition to trends, the weather, business models and even insect populations exhibit cyclical patterns. The variable being studied does not exhibit a continual increase or decrease, but instead goes up and down depending upon the time of year. This cycle of increase and decrease may go on indefinitely. These cyclical patterns are also easy to see with a time series graph.

An Example of a Time Series Graph

We use the data set in the table below to construct a time series graph. The data is from the U.S. Census Bureau and reports the U.S. resident population from 1900 to 2000. The horizontal axis measures time in years and the vertical axis represents the number of people in the U.S. The graph shows us a steady increase in population that is roughly a straight line. Then the slope of the line becomes steeper during the Baby Boom.

7.10 SUMMARY

In this unit, we learnt about processing of data. Irrespective of the methods of data collection, the information is called ‘raw data’. The processing of data includes all operations undertaken on a collected data set, which start s with editing and ends with presentation. The editing is basically clearing your data, for consistency, accuracy, homogeneity and completeness. The coding of data which involves developing a code book, pre testing it and verifying the coded data, is next step. The classification categorizes the information and arranges it in summarized form and makes it easily understandable. The tabulation is the easiest method of presenting data in the form of rows and columns. Coding process assigns numerals or other symbols to the several responses of the data set. It is therefore a pre-

requisite to prepare a coding scheme for the data set. It has a title describing the type of data it contains, headings of columns and rows, properly arranged information in the columns and rows. Tabulation is easily understandable, yet graphic presentations are used to make data presentation more attractive and understandable.

7.11 SELF-ASSESSMENT QUESTIONS

1. What are the various steps in processing of data?
2. How is data editing is done at the time of recording of data?
3. What are types of coding?
4. What is data classification?
5. What are the components of a table?
6. Explain the following:
    a. Editing
    b. Coding
    c. Classification
    d. Tabulation
7. List the rules for constructing diagrams
8. Give the significance and limitations of diagrams.
9. What are the advantages and disadvantages of diagrams?
10. Define:
    a. Pictogram
    b. Histogram
11. What are the differences between graphs and diagrams?
12. Explain the types of graphs.
13. How is a time series graph constructed?
14. What are the principles of table construction?
15. What are the fundamentals of frequency distribution?
16. What are the types and general rules for graphical representation of data?

7.12 TEXT AND REFERENCES

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, “Business Research Methods”, McGraw Hill Inc.

- Alan Bryman, and Emma Bell, “Business research Methods”. Oxford University Press Publication.
- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- S.P. Gupta, “Statistical methods” Sultan Chand & Sons publication, New Delhi.
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Naresh Malhotra, “Marketing Research: An applied Orientation”, Prentice Hall International Edition.
- T.S. Wilkinson, and P.L. Bhandarker, “Methodology and techniques of Social Research”, Himalaya Publishing House, New Delhi.

UNIT-8 ADVANCED TECHNIQUES

Unit Framework

8.1 OBJECTIVE

After studying this unit, you will be able to:

- Understand the meaning of clustering
- Methods of Clustering
- Partitioning Methods
- Hierarchical Agglomerative methods
- The Single Link Method (SLINK)
- The Complete Link Method (CLINK)
- The Group Average Method
- Meta-analysis
- Conjoint analysis

8.2 CLUSTERING

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into *partitioning* methods, in which the classes are mutually exclusive, and the less common *clumping* method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into *agglomerative* or *divisive* methods.

In *agglomerative* methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common *divisive* methods begin with all objects

in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

8.3 METHODS OF CLUSTERING

Some of the important data clustering methods are described below:

8.3.1 PARTITIONING METHODS

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.

2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.

3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order *O(N.logN)* for order *O(logN)* clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

8.3.2 HIERARCHICAL AGGLOMERATIVE METHODS

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster

2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.

3. If more than one cluster remains, return to step 2 Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below:

- In the second matrix approach , an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to O(n3) if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.

- The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an O(N3) time requirement.

8.3.3 THE SINGLE LINK METHOD (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name *single link* thus refers to the joining of pairs of clusters by the single shortest link between them.

8.3.4 THE COMPLETE LINK METHOD (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

8.3.5 THE GROUP AVERAGE METHOD

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster.

8.3.6 TEXT BASED DOCUMENTS

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

8.4 STEPS IN CLUSTER ANALYSIS

The objective of cluster analysis is to group observations into clusters such that each cluster is as homogenous as possible with respect to the clustering variables. The various steps in cluster analysis are:

1. Select a measure of similarity.
2. Decision is to be made on the type of clustering technique to be used
3. Type of clustering method for the selected technique is selected
4. Decision regarding the number of clusters
5. Cluster solution is interpreted.

8.5 META-ANALYSIS

In statistics, a meta-analysis refers to methods focused on contrasting and combining results from different studies, in the hope of identifying patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. In its simplest form, this is normally by identification of a common measure of effect size, of which a weighted average might be the output of a meta-analysis. The weighting might be related to sample sizes within the individual studies. More generally there are other differences between the studies that need to be allowed for, but the general aim of a meta-analysis is to more powerfully estimate the true effect size as opposed to a less precise effect size derived in a single study under a given single set of assumptions and conditions.

Meta-analyses are often, but not always, important components of a systematic review procedure. For instance, a meta-analysis may be conducted on several clinical trials of a medical treatment, in an effort to obtain a better understanding of how well the treatment works. Here it is convenient to follow the terminology used by the Cochrane Collaboration, and use "meta-analysis" to refer to statistical methods of combining evidence, leaving other aspects of 'research synthesis' or 'evidence

synthesis', such as combining information from qualitative studies, for the more general context of systematic reviews.

Meta-analysis forms part of a framework called estimation statistics which relies on effect sizes, confidence intervals and precision planning to guide data analysis, and is an alternative to null hypothesis significance testing.

8.5.1 ADVANTAGES OF META-ANALYSIS

Conceptually, a meta-analysis uses a statistical approach to combine the results from multiple studies. Its advantages can therefore be interpreted as follows:

- Results can be generalized to a larger population,
- The precision and accuracy of estimates can be improved as more data is used. This, in turn, may increase the statistical power to detect an effect.
- Inconsistency of results across studies can be quantified and analyzed. For instance, does inconsistency arise from sampling error, or are study results (partially) influenced by between-study heterogeneity.
- Hypothesis testing can be applied on summary estimates,
- Moderators can be included to explain variation between studies,
- The presence of publication bias can be investigated.

8.5.2 STEPS IN A META-ANALYSIS

In general, two types of evidence can be distinguished when performing a meta-analysis: Individual Participant Data (IPD) and Aggregate Data (AD). Whereas IPD represents raw data as collected by the study centers, AD is more commonly available (e.g. from the literature) and typically represents summary estimates such as odds ratios or relative risks. This distinction has raised the needs for different meta-analytic methods when evidence synthesis is desired, and has led to the development of one-stage and two-stage methods. In one-stage methods the IPD from all studies are modeled simultaneously whilst accounting for the clustering of participants within studies. Conversely, two-stage methods synthesize the AD from each study and hereto consider study weights. By reducing IPD to AD, two-stage methods can also be applied when IPD is available; this makes them an appealing choice when performing a meta-analysis. Although it is conventionally believed that one-stage and two-stage methods yield similar results, recent studies have shown that they may occasionally lead to different conclusions.

8.5.3 ASSUMPTIONS

8.5.3.1 FIXED EFFECTS

The fixed effect model provides a weighted average of a series of study estimates. The inverse of the estimates' variance is commonly used as study weight, such that larger studies tend to contribute more than smaller studies to the weighted average. Consequently, when studies within a meta-analysis are dominated by a very large study, the findings from smaller studies are practically ignored. Most importantly, the fixed effects model assumes that all included studies investigate the same population, use the same variable and outcome definitions, etc. This assumption is typically unrealistic as research is often prone to several sources of heterogeneity; e.g. treatment effects may differ according to locale, dosage levels, and study conditions.

8.5.3.2 RANDOM EFFECTS

A common model used to synthesize heterogeneous research is the random effects model of meta-analysis. This is simply the weighted average of the effect sizes of a group of studies. The weight that is applied in this process of weighted averaging with a random effects meta-analysis is achieved in two steps:

1. **Step 1:** inverse variance weighting
2. **Step 2:** Un-weighting of this inverse variance weighting by applying a random effects variance component (REVC) that is simply derived from the extent of variability of the effect sizes of the underlying studies.

This means that the greater this variability in effect sizes (otherwise known as heterogeneity), the greater the un-weighting and this can reach a point when the random effects meta-analysis result becomes simply the un-weighted average effect size across the studies. At the other extreme, when all effect sizes are similar (or variability does not exceed sampling error), no REVC is applied and the random effects meta-analysis defaults to simply a fixed effect meta-analysis (only inverse variance weighting).

The extent of this reversal is solely dependent on two factors:

1. Heterogeneity of precision
2. Heterogeneity of effect size

The most widely used method to estimate and account for heterogeneity is the Der Simonian- Laird (DL) approach. More recently the iterative and computationally intensive restricted maximum likelihood (REML) approach emerged and is catching up. However, a comparison between these two (and more) models demonstrated that there is little to gain and DL is quite adequate in most scenarios.

8.5.3.3 QUALITY EFFECTS

Some researchers introduce a new approach to adjustment for inter-study variability by incorporating a relevant component (quality) that differs between studies in addition to the weight based on the intra-study differences that is used in any fixed effects meta-analysis model.

The strength of the quality effects meta-analysis is that it allows available methodological evidence to be used over subjective random probability, and thereby helps to close the damaging gap which has opened up between methodology and statistics in clinical research. To do this a correction for the quality adjusted weight of the *i*th study called tau*i* is introduced. This is a composite based on the quality of other studies except the study under consideration and is utilized to re-distribute quality adjusted weights based on the quality adjusted weights of other studies. In other words, if study *i* is of good quality and other studies are of poor quality, a proportion of their quality adjusted weights is mathematically redistributed to study *i* giving it more weight towards the overall effect size. As studies increase in quality, re-distribution becomes progressively less and ceases when all studies are of perfect quality. This model thus replaces the untenable interpretations that abound in the literature and software is available to explore this method further.

8.6 CONJOINT ANALYSIS

Conjoint analysis, also called multi-attribute compositional models or stated preference analysis is a statistical technique that originated in mathematical psychology. Today it is used in many of the social sciences and applied sciences including marketing, product management, and operations research. It is not to be confused with the theory of conjoint measurement.

8.6.1 METHODOLOGY

Conjoint analysis requires research participants to make a series of trade-offs. Analysis of these trade-offs will reveal the relative importance of component attributes. To improve the predictive ability of this analysis, research participants should be grouped into similar segments based on objectives, values and/or other factors.

The exercise can be administered to survey respondents in a number of different ways. Traditionally it is administered as a ranking exercise and sometimes as a rating exercise (where the respondent awards each trade-off scenario a score indicating appeal). In more recent years it has become common practice to present the trade-offs as a choice exercise (where the respondent simply chooses the most preferred alternative from a selection of competing alternatives - particularly common when simulating consumer choices) or as a constant sum allocation exercise (particularly common in pharmaceutical market research, where

physicians indicate likely shares of prescribing, and each alternative in the trade-off is the description a real or hypothetical therapy).

Analysis is traditionally carried out with some form of multiple regressions, but more recently the use of hierarchical Bayesian analysis has become widespread, enabling fairly robust statistical models of individual respondent decision behaviour to be developed. When there are many attributes, experiments with Conjoint Analysis include problems of information overload that affect the validity of such experiments. The impact of these problems can be avoided or reduced by using Hierarchical Information Integration.

8.6.2 EXAMPLE

A real estate developer is interested in building a high rise apartment complex near an urban Ivy League university. To ensure the success of the project, a market research firm is hired to conduct focus groups with current students. Students are segmented by academic year (freshman, upper classmen, graduate studies) and amount of financial aid received. Study participants are given a series of index cards. Each card has 6 attributes to describe the potential building project (proximity to campus, cost, telecommunication packages, laundry options, floor plans, and security features offered). The estimated cost to construct the building described on each card is equivalent.

Participants are asked to order the cards from least to most appealing. This forced ranking exercise will indirectly reveal the participants' priorities and preferences. Multi-variate regression analysis may be used to determine the strength of preferences across target market segments.

8.7 SUMMARY

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. In the partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster.

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity. In the group average method relies on the average value of the pair wise

within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods.

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. In statistics, a meta-analysis refers to methods focused on contrasting and combining results from different studies, in the hope of identifying patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. A common model used to synthesize heterogeneous research is the random effects model of meta-analysis.

The strength of the quality effects meta-analysis is that it allows available methodological evidence to be used over subjective random probability, and thereby helps to close the damaging gap which has opened up between methodology and statistics in clinical research. In Conjoint analysis, also called multi-attribute compositional models or stated preference analysis is a statistical technique that originated in mathematical psychology.

8.8 SELF-ASSESSMENT QUESTIONS

1. What is clustering? Explain its methods.
2. What are Text Based Documents?
3. Describe the Steps in Cluster analysis.
4. Explain the concept of Meta-analysis.
5. What are the advantages of meta-analysis?
6. Elaborate the steps in a meta-analysis.
7. What is conjoint analysis? Give its methodology.
8. Cite an authentic example of conjoint analysis.

8.9 TEXT AND REFERENCES

- Donald S. Tull and Del S. Hawkins, “Marketing Research – Measurement and Method”, Prentice Hall of India Pvt. Ltd., New Delhi.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- Alan Bryman, and Emma Bell, “Business Research Methods”. Oxford University Press Publication.

- Brown, F.E. "Marketing Research, a structure for decision making", Addison - Wesley Publishing Company.
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Stockton and Clark, "Introduction to Business and Economic Statistics" D.B. Taraporevala Sons and Co. Private Limited, Bombay.
- Pauline V. Young, Scientific Social Surveys and Research.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.

Uttar Pradesh Rajarshi Tandon
Open University

BBA-121

Research Methodology

BLOCK

3

CENTRAL TENDENCY, PROBABILITY AND STATISTICAL TOOLS

UNIT-9

CENTRAL TENDENCY MEASURES

UNIT-10

DISPERSION

UNIT-11

CORRELATION AND REGRESSION

UNIT-12

PROBABILITY THEORY

परिशिष्ट–4

आन्तरिक कवर–दो का प्ररूप

Format of the II Inner Covers

लेखक Dr. Piyali Ghosh, Asst. Professor, School of Management, MNNIT, Allahabad

सम्पादक Prof. H.K. Singh, Professor, Department of Commerce, BHU Varansi.

परिमापक

सहयोगी टीम

संयोजक Dr. Gaurav Sankalp, SoMS, UPRTOU, Allahabd.

प्रूफ रीडर

©UPRTOU, Prayagraj-2020

ISBN : 978-93-83328-54-3

NODE-110/NODE-124/NFT-063

BLOCK INTRODUCTION

Unit-09 : Central Tendency Measures.

Introduction, Characteristics of Good Average, Arithmetic mean, Weighted Arithmetic Mean, Median, Mode, Geometric Mean, Harmonic Mean.

Unit-10 : Dispersion.

Introduction, Range, Quartile Deviation, Mean Deviation, Standard Deviation, Coefficient of Variation, Lorenze Curve.

Unit-11 : Correlation and Regression, Significant.

Types and Methods of Determining Correlation, Concept and Types of Regression, Regression line, Distinction between Correlation and Regression.

Unit-12 : Probability Theories.

Addition and Multiplication Theorem, Premulation and Combination.

This Block (3) comprises Four Units. The First Unit of this block is related with Measures of Central Tendency, while Second Unit deals with Dispersion. The Third Unit of this block is related with Correlation and Regression. The Last Unit of this block is concerned with Probability Theories.

UNIT-09 CENTRAL TENDENCY MEASURES

Objectives

After going through this unit you should be able to know about the–

1. Various Measures of Central Tendency
2. Geometric Mean
3. Harmonic Mean

Structure

9.1. INTRODUCTION

The utility of various statistical derivatives like ratio, percentages and rates was discussed by us earlier. There it was pointed out that these measures help in reducing the size of the data and enable us to make comparative studies of related variables. But these derivatives are not enough for the proper condensation of figures and, sometimes, there are many fallacies in their use. Condensation of data is necessary in statistical analysis because a large number of big figures are not only confusing to mind but also difficult to analyse. In order to reduce the complexity of data and to make them comparable, it is essential that the various

phenomena which are being compared are reduced to one figure each. If, for example, a comparison is made between the marks obtained by a group of 200 students belonging to a university and marks obtained by another group of 200 students belonging to another university, it would be impossible to arrive at any conclusion, if the two series relating to these marks are directly compared. On the other hand, if each of these series is represented by one figure, comparison would become extremely easy. It is obvious that a figure which is used to represent a whole series should neither have the lowest value in the series nor the highest value but a value somewhere between these two limits, possibly in the centre, where most of the items of the series cluster. Such figures are called Measures of Central Tendency or Averages. The average represents a whole series and as such, its value always lies between the minimum and maximum values and, generally, it is located in the centre or middle of the distribution.

9.2 DEFINITION

The word average or the term measures of central tendency have been defined by various authors in their own way. Some of the definitions are given below :—

Simpson and Kafka observe that "A measure of central tendency is a typical value around which other figures congregate."

Lawrence J. Kaplan has defined these terms in the following words :

"One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, central tendency or central location. The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp in simple manner quickly. The single value is the point of location around which the individual items cluster."

Va-Lun Chou states that "an average is a typical value in the sense that it is, sometimes, employed to represent all the individual values in a series or of a variable."

It is, thus, clear from the above definitions that an average is a single value which represents a whole series and is supposed to contain its major characteristics.

9.3 OBJECTIVES OF AVERAGING

1. **To get a single value which is representative of the characteristics of the entire mass of data :** Averages give a bird's-eye view of the huge mass of statistical data which ordinarily are not easily intelligible.

They are devices to aid the human mind in grasping the true significance of large aggregates of facts and measurements. They set aside the unnecessary details of the data and put forward a concise picture of the complex phenomena under investigation because the human mind is not capable of grasping all the details of large numbers and their interrelationship. For example, it is not possible to keep in mind, the details of heights. weights. incomes and expenditures of even 200 students, what to talk of big figures. This difficulty of keeping all the details in mind necessitates the use of averages. An average is a single number representing the whole data and is useful in grasping the central theme of the data.

Why is an average a representative? The reason why an average is a valid representative of a series lies in the fact that ordinarily most of the items of a series cluster in the middle. On the extreme ends, the number of items is very little. In a population of 10,000 adults, there would hardly be any person who is 60 cms. high or whose height is above 240 cms. There will be a small range within which these values would vary, say, 150 cms. to 200 cms. Even within this range, a large number of persons would have a height between, say, 160 cms. and 180 cms. In other class intervals of height, the number of persons would be comparatively small. Under such circumstances if we conclude that the height of this particular group of persons would be represented by, say, 170 cms., we can reasonably be sure that this figure would, for all practical purposes, give us a satisfactory conclusion. This average would satisfactorily represent the whole group of figures from which it has been calculated.

2. **To facilitate comparison :** Since measures of central tendency or averages reduce the mass of statistical data to a single figure, they are very helpful for purposes of making comparative studies, for example, the average marks obtained by two sections of a class would give a reasonably clear picture about the level of their performance, which would not be possible if we had two full series of marks of individual students of the two sections.

However, when such a comparison is made, we have to be careful in drawing inferences, as the marks of students in one section may vary within a small range and in the other section some students may have got very high marks and others very few marks. The comparison of averages in such a case may give misleading conclusions.

9.4. CHARACTERISTICS OF A GOOD AVERAGE

1. **It should be rigidly defined:** If an average is left to the estimation of an observer and if it is not a definite and fixed value, it cannot

be representative of a series. The bias of the investigator in such cases would considerably affect the value of the average. If the average is rigidly defined, this instability in its value would be no more, and it would always be a definite figure.

2. **It should be based on all the observations of the series:** If some of the items of the series are not taken into account in its calculation, the average cannot be said to be a representative one.

3. **It should be capable of further algebraic treatment:** If an average dose not possess this quality, its use is bound to be very limited. It will not be possible to calculate, say, the combined average of two or more series from their individual averages; further, it will not be possible to study the average relationship of various parts of a variable if it is expressed as the sum of two or more variables etc.

4. **If should be easy to calculate and simple to follow:** If the calculation of the average involves tedious mathematical processes, it will not be readily understood by a person of ordinary intelligence and its use will be confined only to a limited number of persons and, hence, can never be a popular measure. As such, one of the qualities of a good average is that it should not be too abstract or mathematical and should be easy to calculate.

5. **It should not be affected by fluctuations of sampling:** If two independent sample studies are made in any particular field, the averages thus obtained, should not materially differ from each other. No doubt, when two separate enquiries are made, there is bound to be a difference in the average values calculated but, in some cases, this difference would be great while in others comparatively less. These averages in which this difference. which is technically called "fluctuation of sampling", is less, are considered better than those in which its difference is more.

9.5 TYPES OF AVERAGES

Measures of central tendency or averages are usually of the following types :

1. **Mathematical Averages:**

   (a) Arithmetic Average or Mean

   (b) Geometric Mean

   (c) Harmonic Mean

2. **Averages of Position (Positional averages) :**

   (a) Median

(b) Mode

Of the above mentioned five important averages, Arithmetic Average, Median and Mode are the most popular ones. Geometric mean and Harmonic mean come next. We shall study them in this very order.

9.6 ARITHMETIC AVERAGE

Arithmetic Average or Mean of a series is the figure obtained by dividing the sum of the values of the various items by their number. If the heights of a group of eleven persons are 164, 169, 163, 160, 165, 168, 162, 167, 170, 166, 161 centimeters, then to find the arithmetic average of the heights of these persons we shall add these figures and divide the total so obtained, by the number of items which is 11. The total of the items in this case is 1815[1] cms. and if it IS divided by 11, we get the figure of 165 cms. This is the mean or arithmetic average of the series.

Calculation of the arithmetic average in a series of individual observations

A.M. = $\frac{\text{Sum of the values}}{\text{Number of the values}} \Rightarrow$ (AM) Number of values = Sum of the values

Suppose the values of a variable are respectively X_1, X_2, X_3, ... X_n, and their arithmetic average is represented by $\overline{X}$, then

$$\overline{X} = \frac{1}{N}\left(X_1 + X_2 + X_3 \ldots X_n\right)$$

or $$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_1 \quad \text{or} \quad \overline{X} = \frac{\sum_{i=1}^{N} X_1}{N}$$

where, $\overline{X}$ = Arithmetic average; X_1's = values of the variable; $\square$ = Summation or total; N = Number of items.

The following example would illustrate this formula.

Example 1. Calculate the simple arithmetic average of the following items :

Size of items		
20	50	72
28	53	74
34	54	75
39	59	78
42	64	79

Solution. Direct Method :

Computation of arithmetic average Size of items

X : 20, 28, 34, 39, 42, 50, 53, 54, 59, 64, 72, 74, 75, 78, 79

$\sum X = 20+28+34+39+42+50+53+54+59+64+72+74+75+78+79 = 821$

Arithmetic mean = $\frac{\sum X}{N} = \frac{821}{15} = 54.73$

Calculation of arithmetic average in a discrete series

Direct Method: In a discrete series, the values of the variable are multiplied by their respective frequencies and the products so obtained are totalled. This total is divided by the number of items which, in a discrete series, is equal to the total of the frequencies. The resulting quotient is a simple arithmetic average of the series.

Algebraically,

If, f_1, f_2, f_3, etc., stand respectively for the frequencies of the values X_1, X_2, X_3, etc.

$$\overline{X} = \frac{1}{N}\left(X_1 f_1 + X_2 f_2 + X_3 f_3 + ... + X_n f_n\right)$$

or $\overline{X} = \frac{\sum fX}{N}$ or $\frac{\sum fX}{\sum f}$

Short-cut method I : A short-cut method can be used in the discrete series also. In this method, the deviations of the items from an assumed mean are first found out and they are multiplied by their respective frequencies. The total of these products is divided by the total frequencies and added to the assumed mean. The resulting figure is the actual arithmetic average.

Algebraically : $$\overline{X} = A + \frac{\sum fdx}{N}$$

where, $\sum fdx$ = the total of the products of the deviations from the assumed average and the respective frequencies of the items.

Step deviation method

In step deviation method, we define $d'x = \frac{dx}{C}$, where C is some common factor in dx values and then apply the formula :

$$\overline{X} = A + \frac{\sum f\left(\frac{dx}{C}\right)}{N} \times C.$$

This is called **short-cut method II.**

Example 2. The following table gives the marks obtained by a set of students in a certain examination. Calculate the average mark per student.

Marks	**Number of students**	**Marks**	**Number of students**
10 – 20	1	60 – 70	12
20 – 30	2	70 – 80	16
30 – 40	3	80 – 90	10
40 – 50	5	90 – 100	4
50 – 60	7		

Solution. Short-cut Method

Computation of average marks per student

Marks (X)	Mid values (m.v.)	No. of students (f)	Deviation from assumed mean (55)	Step deviations (10) (dx)	Total deviation (fdx)
10 – 20	15	1	– 40	– 4	– 4
20 – 30	25	2	– 30	– 3	– 6
30 – 40	35	3	– 20	– 2	– 6
40 – 50	45	5	– 10	– 1	– 5
50 – 60	55	7	– 0	0	0
60 – 70	65	12	+ 10	+ 1	+ 12
70 – 80	75	16	+ 20	+ 2	+ 32
80 – 90	85	10	+ 30	+ 3	+ 30
90 – 100	95	4	+ 40	+ 4	+ 16
		N = 60			$\sum$fdx = + 69

Arithmetic average or $\overline{X} = A + \left(\frac{\sum fdx}{N} \times i \right) = 55 + \left(\frac{69}{60} \times 10 \right)$ = 66.5 marks.

Merits of arithmetic average

The arithmetic average is the most popularly used measure of central tendency. There are many reasons for its popularity. In the beginning of this chapter, we had laid down certain characteristics which an ideal average should possess. We shall now see how far the arithmetic average fulfils these conditions:

1. The first condition that an average should be rigidly defined is fulfilled by the arithmetic average. It is rigidly defined and a biased investigator shall get the same arithmetic average from the series as an unbiased one. Its value is always definite.
2. The second characteristic that an average should be based on all the observations of a series is also found in this average. Arithmetic average cannot be calculated even if a single item of a series is left out.

3. Arithmetic average is also capable of further algebraic treatment. While discussing the algebraic properties of the arithmetic average, we have already seen in detail, how various mathematical processes can be applied to it for purposes of further analysis and interpretation of data. It is on account of this characteristic of the arithmetic average that:

   (a) It is possible to find the aggregate of items of a series if only its arithmetic average and the number of items is known.

   (b) It is possible to find the arithmetic average if only the aggregate of items and their number is known.

4. The fourth characteristic laid down for an ideal average that it should be easy to calculate and simple to follow, is also found in arithmetic average. The calculation of the arithmetic average is simple and it is very easily understandable. It does not require the arraying of data which is necessary in case of some other averages. In fact, this average is so well known that to a common mean average means an arithmetic average,

   Thus, the arithmetic average

   (a) is simple to calculate,

   (b) does not need arraying of data,

   (c) is easy to understand.

5. The last characteristic of an ideal average that it should be least affected by fluctuations of sampling is also present in arithmetic average to a certain extent. If the number of items in a series is large, the arithmetic average provides a good basis of comparison, as in such cases, the abnormalities in one direction are set off against the abnormalities in the other direction.

9.7 WEIGHTED ARITHMETIC AVERAGE

Need for weighting an average : In the calculation of simple average, each item of the series is considered equally important but there may be cases where all items may not have equal importance, and some of them may be comparatively more important than the others. The fundamental purpose of finding out an average is that it shall "fairly" represent, so far as a single figure can, the central tendency of the many varying figures from which it has been calculated. This being so, it is necessary that if some items of a series are more important than others, this fact should not be overlooked altogether in the calculation of an average. If we have to find out the average income of the employees of a certain mill and if we simply add the figures of the income of the manager, an accountant, a clerk, a labourer and a watchman and divide the total by five, the average so obtained cannot be a fair representative of the income of these people. The reason is that in a mill, there may be one manager, two accountants, six clerks, one thousand labourers and one dozen watchmen, and if it is so,

the relative importance of the figures of their income is not the same. Similarly, if we are finding out the change in the cost of living of a certain group of people and if we merely find the simple arithmetic average of the prices of the commodities consumed by them, the average would be unrepresentative. All the items of consumption are not equally important. The price of salt may increase by 500 per cent but this will not affect the cost of living to the extent to which it would be affected, if the price of wheat goes up only by 50%. In such cases, if an average has to maintain its representative character, it should take into account the relative importance of the different items from which it is being calculated. The simple average gives equal importance to all the items of a series.

Direct Method : In calculating the weighted arithmetic average, each value of the variable is multiplied by its weights and the products so obtained are aggregated. This total is divided by the total of weights and the resulting figure is the weighted arithmetic average.

Symbolically,

$$\overline{X}_w = \frac{X_1w_1 + X_2w_2 + X_3w_3 + ... + X_nw_n}{w_1 + w_2 + w_3 + ... + w_n}$$

where $\overline{X}_w$ stands for the weighted arithmetic average, X_1, X_2, etc., for the values of the variable and w_1, w_2 etc., for their respective weights :

The formula can be written in short as :

$$\overline{X}_w = \frac{\sum Xw}{\sum w}$$

where, $\sum Xw$ stands for the sum of the weights, and respective weights, and $\sum w$

Indirect method : Weighted arithmetic mean can be calculated by an indirect method also where we assume an average and take the deviation from the assumed mean. These deviations are multiplied by respective weights of items. The sum of these products is then divided by the total of weights and added to the assumed average. This figure would be the value of the weighted arithmetic average.

Symbolically,

$$\overline{X}_w = A_w + \frac{\sum (dx)w}{\sum w}$$

where, $\overline{X}_w$ = weighted arithmetic mean; A_w = assumed mean weighted; dx = deviations of items from assumed mean; w = Weights of various items.

Example 2. Calculate simple and weighted arithmetic averages from the following data and comment

Designation	Monthly salary (in Rs.)	Strength of the cadre
Class I Officers	1,500	10
Class II Officers	800	20
Subordinate Staff	500	70
Clerical Staff	250	100
Lower Staff	100	150

Solution :

Computation of Simple and Weighted A.M.

Designation	Monthly salary in Rs. (X)	Strength of the cadre (w)	(Xw)
Class I Officers	1,500	10	15,000
Class II Officers	800	20	16,000
Subordinate Staff	500	70	35,000
Clerical Staff	250	100	25,000
Lower staff	100	150	15,000
N = 5	**$\sum$X = 3,150**	**$\sum$(w) = 350**	**$\sum$(Xw) = 1,06,000**

Simple arithmetic average = $\frac{\sum X}{N} = \frac{3150}{5}$ = Rs. 630

Weighted arithmetic average = $\frac{\sum(Xw)}{\sum w} = \frac{1,06,000}{350}$ = Rs. 302.857

9.8 MEDIAN

Introduction and definition

Median is the value of the middle item of a series arranged in ascending or descending order of magnitude. Thus, if there are 9 items in a series arranged in ascending or descending order of magnitude, median will be the value of the 5^{th} item. This item would divide the series in two equal parts—one part containing values less than the median value and the other part containing values more than the median value. If, however, there are even number of items in a series, then there is no central item dividing the series in two equal parts. For example if there are 10 items in a series, the median value would be between the values of 5th and 6th items. It would, thus, be the arithmetic average of the values of 5th and 6th items or it would be equal to the value of the 5th item plus the value of the 6th item divided by two.

According to **A.L. Bowley,** "If the numbers of the group are ranked in order according to the measurement under consideration, then the measurement of the number most nearly one half is the median."

The above definitions of the median do not hold good in situations where a median value is surrounded by neighbouring values which are equal in magnitude to it. For example, in a series of values such as 12, 13, 14, 15, 16, 17 and 18, there is no value which is so located that three values are smaller than it and three are greater than it. However, value 15 is designated as median. Keeping in view such situations, **Croxton** and **Cowdon** have given a revised definition of median as, "The median is that value which divides a series so that one half or more of the items are equal to or less than it and one half or more of the items are equal to or greater than it."

Calculation of Median

The calculation of median involves two basic steps, viz. (i) the location of the middle item and (ii) finding out its value.

The middle item in series of individual observations and also in a discrete series is $\left(\frac{n+1}{2}\right)^{th}$ item, where n is the total number of observations. In case of a continuous series $\left(\frac{n}{2}\right)^{th}$ item is the middle item of the series.

Once the middle item is located, its values has to be found out. In a series of individual observations, if the total number of items is an odd figure, the value of the middle item is the median value. If the number of items is even, the median value is the average of the two items in the centre of the distribution. The examples given below would clarify these points.

Computation of Median in a series of individual observations

Example 3. Find out the median of the following items :

5, 7, 9, 12, 10, 8, 7, 15, 21

Solution. These items would first be arranged in ascending order of magnitude. The series then would he as follows:

Calculation of Median

Serial Number	Size of items
1	5
2	7
3	7
4	8
5	9
6	10
7	12
8	15
9	21

If M represents the median and N the number of items.

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Size of } \left(\frac{9+1}{2}\right)^{th} \text{ or } 5^{th} \text{ item} = 9.$$

In the above example, the number of items was odd and there was no difficulty in locating the middle items and its value. If the number of items is even, the middle item and its value would be calculated as illustrated in the following example.

Example 4. Find out the value of median from the following data :

Daily wages (R)	10	5	7	11	8
Number of Workers	15	20	15	18	12

Solution

Calculation of median

Wages in ascending order (Rs.)	Number of persons (f)	Cumulative Frequency (c.f.)
5	20	20
7	15	35
8	12	47
10	15	62
11	18	80

Median is the value of $\left(\frac{N+1}{2}\right)^{th}$ or $\left(\frac{80+1}{2}\right)^{th}$ 40.5^{th} items. All items from 35 onwards upto 47 have a value of 8. Thus, the median value would be Rs. 8.

Continuous series descending order : In such series, there is a slight change in the formula for calculating median. Since the series are cumulated in descending order, the cumulative frequency of the class preceding the median class is found out by subtracting the cumulative frequency of the median class from the total of the cumulative frequency. In other words :

$$c = (N - c.f.)$$

where c = cumulative frequency less of the class preceding the median class

N = total cumulative frequency

cf = cumulative frequency of the median class.

The following example would illustrate the point :

Example 5. Calculate median from the following data :

Age	Number of persons	Age	Number of persons
55—60	7	35—40	30
50—55	13	30—35	33
45—50	15	25—30	28
40—45	20	20—25	14
			Total 160

Solution

Computation of Median

Age	Number of persons	Cumulative frequency
55—60	7	7
50—55	13	20
45—50	15	35
40—45	20	55
35—40	30	85
30—35	33	118
25—30	28	146
20—25	14	160

In the above example, median is the value of $\left(\frac{N}{2}\right)^{th}$ or $\left(\frac{160}{2}\right)^{th}$ or 80^{th} item which lies in 35—40 class interval.

The frequency of the class preceding the median class or the value of c = Total frequency minus the cumulative frequency of the median class or

$$160—85 = 75$$

Median

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - c) = 35 + \frac{40 - 35}{30}(80 - 75) = 35 + \left(\frac{5}{30} \times 5\right) = 35.83$$

Example 6. Compute median from the following data :

Mid-values:	115	125	135	145	155	165	175	185	195
Frequency:	6	25	48	72	116	60	38	22	3

Solution. Here, we are given the mid-values of the class-intervals of a continuous frequency distribution. The difference between two mid-values is 10, hence, 10/2 = 5 is reduced from each mid-value to find the lower limit and the same is added to find the upper limit of a class. The classes are, thus, 110–120, 120–130 ..... and so on upto 190–200.

Computation of Median

Class-intervals	**Frequency**	**Cumulative frequency**
110 – 120	6	6
120 – 130	25	31
130 – 140	48	79
140 – 150	72	151
150 – 160	116	267
160 – 170	60	327
170 – 180	38	365
180 – 190	22	387
190 – 200	3	390
Total	**390**	

The middle item is $\frac{390}{2}$ or 195 which lies in the 150—160 group.

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

$$= 150 + \frac{160 - 150}{116}(195 - 151) = 150 + \frac{10}{116}(44)$$

$$= 153.8$$

9.9. MODE

Mode is the most common item of a series. Generally, it is the value which occurs largest number of times in a series. In the words of **Croxton** and **Cowden,** "The mode of a distribution' is the value at the point around which the items tend to be most heavily concentrated."

According to **A.M. Tuttle,** 'Mode is the value which has the greatest frequency density in its immediate neighbourhood.'

The above two definitions indicate that mode is a value around which there is the greatest concentration of values. It may not necessarily be the value which occurs the largest number of times in a series, as in some cases, the point of maximum concentration may be around some other value. In some cases, there may be more than one point of concentration of values and the series may be bi-modal or multi-modal. We shall discuss these cases later.

The word Mode is derived from the French word (*la mode*) which means fashion or the most popular phenomenon. Mode, thus, is the most popular item of a series around which there is the highest frequency density. When we speak of the 'average student', 'average collar size', 'average size of a shoe', we are referring to mode. When we say that, on an average, a student spends Rs. 300 per month, we imply that a very large number of students spend around Rs. 300 per month. It is the value of mode. It is the most typical or fashionable value of the series.

Example 7. Find the mode of the following data relating to the weight of 10 students :

Sl. No.	Weight in pounds	Sl. No.	Weight in pounds
1	20	6	130
2	130	7	132
3	135	8	132
4	130	9	135
5	140	10	141

Solution :

Weight in pounds	No. of Students
120	1
130	3
132	2
135	2
140	1
141	1
	10

Since item 130 occurs the largest number of times, it is the modal value.

If there are more than one point of concentration, mode cannot be found and the series is called bi-modal.

Grouping method : In discrete and continuous series, if the items concentrate at more than one value, attempts are made to find out the point of maximum concentration with the help of grouping method. In this method, values are first arranged in ascending order and the frequencies against each value are written down. These frequencies are then added in two's and the totals are written in lines between the values added.

Frequencies can be added in two's in two ways :

(i) By adding frequencies of items number 1 and 2; 3 and 4; 5 and 6 and so on.

(ii) By adding frequencies of items number 2 and 3; 4 and 5; 6 and 7 and so on. After this, the frequencies are added in three's. This can be done in three ways:

(a) By adding frequencies of items number 1, 2 and 3, 4, 5 and 6, 7, 8 and 9 and so on.

(b) By adding frequencies of items number 2, 3 and 4, 5, 6 and 7, 8, 9, and 10 and so on.

(c) By adding the frequencies of items number 3, 4 and 5, 6, 7 and 8, 9, 10 and 11 and so on.

If necessary, frequencies can be added in four's and five's also. After this, the size of items containing the maximum frequencies are noted down and the item which has the maximum frequency the largest number of times is called the mode. If grouping has been done in case of continuous series we shall be in a position to determine the modal class by this process.

Example 8. Find the mode of the following series :

Size	Frequency	Size	Frequency
5	48	13	52
6	52	14	41
7	56	15	57
8	60	16	63
9	63	17	52
10	57	18	48
11	55	19	40
12	50	—	—

Solution

Location of mode by grouping

Size of Item (x)	Frequency (f)					
	(1)	(2)	(3)	(4)	(5)	(6)
5	48	100				
6	52		108	156		
7	56	116				
8	60		123		168	179
9	63	120		180		
10	57		112		175	
11	55	105				162
12	50		102	157		
13	52	93			143	
14	41		98			
15	57	120		161		150
16	63		115		172	
17	52	100				
18	48		88	140		163
19	40					

The frequencies in column (1) are first added in two's in columns (2) and (3). Then they are added in three's in columns (4), (5) and (6). The maximum frequency in each column is indicated by thick letters. It will be observed that mode changes with the change in grouping. Thus, according to column (I), mode should be 9 or 16. To find out the point of maximum concentration, the data can be arranged in the shape of table as follows:

Analysis Table

Columns	Size of item containing maximum frequency						
(1)			9				16
(2)			9	10		15	16
(3)		8	9				
(4)		8	9	10			
(5)			9	10	11		
(6)	7	8	9				
No. of times a size occurs	1	3	9	3	1	1	2

Since the size 9 occurs the largest number of times, it is the modal size or mode is 9.

Example 9. Find the mode from the following data :

Values	Frequency	Values	Frequency
Below 50	97	Below 30	60
Below 45	95	Below 25	30
Below 40	90	Below 20	12
Below 35	80	Below 15	4

Solution. The cumulative series would first be converted into a simple continuous series as follows :

Values	Frequency	Values	Frequency
45 – 50	2	25—30	30 f_1
40 – 45	5	20—25	18 f_0
35 – 40	10	15—20	8
30 – 35	20 f_2	10—15	4

This series does not need grouping as modal class is very prominent. The maximum frequency 30 is against the class-interval 25 – 30 which is the modal class. Grouping would also give the same result. Hence,

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(l_2 - l_1)$$

$$= 25 + \frac{30 - 18}{60 - 18 - 20}(30 - 25)$$

$$= 25 + \left(\frac{12}{22} \times 5\right) = 27.72$$

Example 10. Modal marks for a group of 94 students are 54. Ten students got marks between 0—20, thirty students between 40—60 and fourteen students between 80—100. Find out the number of students getting marks between 20—40 and 60—80 if the maximum mark of the test were 100.

Solution.

Marks	No. of Students
0—20	10
20—40	x
40—60	30
60—80	y
80—100	14
	94

Mode = $l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(l_2 - l_1)$, Mode is given as 54

so $$54 = 40 + \frac{30 - x}{60 - x - y}(60 - 40)$$

or $$14 = \frac{30 - x}{60 - x - y} \times 20 = \frac{600 - 20x}{60 - x - y}$$

or $840 - 14x - 14y = 600 - 20x$ or $6x - 14y = -240$

The total number of students is 94. Therefore, the missing values (x + y) would be (94 – 10 – 30 – 14) or 40.

So, we have two equations :

$6x - 14y = -240$ or $x + y = 40$

If they are solved as simultaneous equation, we get :

$6x - 14y = -240$ ... (i)

$6x + 6y = 240$ ... (ii)

Subtracting equation (ii) from (i) we get :

$-20y = -480$ or $y = 24$

Since x + y = 40, therefore x = 40 – 24 or 16.

The missing values are, thus, 16 and 24.

9.10 GEOMETRIC MEAN

Geometric mean is defined as the n^{th} root of the product of *n* items of a series. Thus, if the geometric mean of 3, 6 and 8 is to be calculated it would be equal to the cube root of the product of these figures. Similarly, the geometric mean of 8, 9, 12 and 16 would be the 4^{th} root of the product of these four figures.

Symbolically, GM = $\sqrt[n]{X_1.X_2.X_3....X_n}$

where GM stands for the geometric mean, N for the number of items and X for the values of the variable.

The calculation of the geometric mean by this process is possible only if the number of items is very few. If the number of items is large and their

size is big, this method is more or less out of question. In such cases, calculations have to be done with the help of logarithm. In terms of logs.

$$GM = \left[X_1 . X_2 . X_3 .... X_n\right]^{\frac{1}{n}}$$

$$\Rightarrow \qquad \log GM = \frac{\log X_1 + \log X_2 + \log X_3 .... \log X_n}{N}$$

$$GM = \text{Anti-log}\left[\frac{\log X_1 + \log X_2 + \log X_3 .... \log X_n}{N}\right]$$

or $$GM = \text{Anti-log}\left[\frac{\sum \log X}{N}\right]$$

Thus, geometric mean is the anti-log of the arithmetic average of the logs of the values of a variable. It should be noted that the value of the geometric mean is always less than the value of the arithmetic average unless all the items have equal value in which case the geometric mean and arithmetic average have identical values.

The following examples would illustrate the calculation of geometric mean.

Calculation of geometric mean in a series of individual observations.

Example 11. Calculate the simple geometric mean from the following items:

133, 141, 125, 173, 182

Solution.

Calculation of the geometric mean

Size of item	**Logarithms**
133	2.1239
141	2.1492
125	2.0969
173	2.2380
183	2.2601
N = 5	**$\sum$ log s = 10.8681**

According to the formula, viz.,

Geometric Mean = $\sqrt[n]{X_1.X_2.X_3....X_n}$

$$GM = \text{Anti-log}\left[\frac{\log X_1 + \log X_2 + \log X_3....\log X_n}{N}\right]$$

$$GM = \text{Anti-log}\left(\frac{10.8681}{5}\right) = \text{Anti-log } 2,1736$$

GM = 149 (to the nearest whole number)

Thus, the geometric mean is 149.

Geometric Mean can also be calculated by assuming the logarithm of a number and taking deviations of' the logs of actual items from the assumed log. In such a case,

$$GM = \text{Anti-log}\left[\text{assumed log} + \frac{\sum \text{Deviations}}{N}\right]$$

The earlier example has been solved in this manner below:

Alternate Method

Size of item (X)	**logs X**	**Deviation from assumed log mean (2.000) dx**
133	2.1239	.1239
141	2.1492	.1492
125	2.0969	.0969
173	2.2380	.2380
182	2.2601	.2601
N = 5		**$\sum$ log dx = .8681**

$$\text{Geometric Mean} = \text{Anti-log}\left[\text{assumed log} + \frac{\sum \text{Deviations}}{N}\right]$$

$$= \text{Anti-log}\left[2 + \frac{08681}{5}\right] = \text{Anti-log } 2.1736.$$

= 149 (to the nearest whole number)

Thus, the geometric mean is 149.

Calculation of GM in discrete series

In discrete series, the geometric mean of

$$GM = \text{Anti-log } \frac{\sum f \log X}{N}$$

where, f is the frequency, X the value of the item and N the total number of items.

The steps of calculation are :

(i) Find the logarithms of variable X.

(ii) Multiply these logs with the respective frequencies, and total all such values, it would be $\sum f \log X$.

(iii) Divide $\sum f \log X$ by total frequency or N.

(iv) Find out the anti-log of this value $\frac{\sum f \log X}{N}$. It will be the value of the geometric mean.

Example 12. Find the geometric mean of the following distribution :

Values	Frequency
352	48
220	10
230	8
160	12
190	15

Solution.

Values (X)	Frequency (f)	log X	flog X
352	48	2.5465	122.2320
220	10	2.3424	23.4240
230	8	2.3617	18.8936
160	12	2.2041	26.4492
190	15	2.2788	34.1820
Total	N = 93		$\sum$ f log X = 225.1808

Geometric Mean or GM = Anti-log $\left[\frac{\sum f \log X}{N}\right]$

$$= \text{Anti-log } \frac{(225.1808)}{93} = \text{ Anti-log } 2.4134 = 263.8.$$

9.11 HARMONIC MEAN

Harmonic Mean of a series is the reciprocal of the arithmetic average of the reciprocal of the values of its various items.

Symbolically, the Harmonic mean or HM of a series :

$$\text{HM} = \text{Reciprocal of } \frac{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}{N}$$

$$= \frac{1}{\frac{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}{N}} = \frac{1}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}$$

where X stands for value of the variable.

Thus, the Harmonic Mean of 2, 4 and 8 would be reciprocal of $\frac{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}}{3}$

or reciprocal of $\frac{\frac{7}{8}}{3}$ or reciprocal of $\frac{7}{24}$ or $\frac{24}{7}$ = 3.43.

If the number of items in a series is large it would be a tedious job to find the reciprocal of each item and then to total them and then to divide the total by the number of items and then to find out the reciprocal of the value. The formula can be simplified in the following manner :

$$\text{HM} = \text{Reciprocal of } \frac{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}}{N}$$

or $$\frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}} \quad \text{or} \quad \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

Thus, H.M. is equal to the number of items or N divided by the sum of the reciprocals of different items. This formula is used in the series of individual observation.

In discrete series: We first find out the reciprocal of each value and multiply it by the concerned frequency. Then total the products and divide the total by the total frequency and then find out the reciprocal of the value, which is the harmonic mean of the series. Thus, in a discrete series:

$$\text{H.M.} = \text{Reciprocal of } \frac{f_1\left(\frac{1}{X_1}\right)+f_2\left(\frac{1}{X_2}\right)+\ldots.+f_n\left(\frac{1}{X_n}\right)}{\Sigma f}$$

$$= \frac{\Sigma f}{f_1\left(\frac{1}{X_1}\right)+f_2\left(\frac{1}{X_2}\right)+\ldots.+f_n\left(\frac{1}{X_n}\right)}$$

This formula can again be simplified as :

$$\text{H.M.} = \frac{\Sigma f}{\Sigma\left(f \times \frac{1}{X}\right)} = \frac{N}{\Sigma\left(f \times \frac{1}{X}\right)}$$

In continuous series, the value of the variable, is the mid-point of the class-interval or m.v. = X and the formula for the calculation of H.M. in such a series would be :

$$\text{H.M.} = \frac{\sum f}{\sum\left(f \times \frac{1}{M}\right)} = \frac{N}{\sum\left(f \times \frac{1}{M}\right)}$$

We shall now use these formula to calculate the H.M. in the foregoing examples.

Calculation of Harmonic in a series of individual observations.

Example 13. Calculate the Harmonic mean of the following values:

15, 250, 15.7, 157, 1.57, 105.7, 10.5, 1.06, 25.7 and 0.257

Solution. We shall find out the reciprocals of the above values from the mathematical tables which are available, instead of manual calculation.

Calculation of Harmonic Mean

Values (X)	Reciprocal (1/X)
15	0.06667
250	0.00400
15.7	0.06369
157	0.00637
1.57	0.63690
105.7	0.00946
10.5	0.09524
1.06	0.94340
25.7	0.03891
0.257	3.89100
N = 10	5.75564 □(1

$$\text{H.M.} = \frac{N}{\sum\left(\frac{1}{X}\right)} = \frac{10}{5.75564} = 1.735$$

Calculation of H.M. in continuous series

As has been pointed out earlier, in a continuous series, the mid-values of class-interval represent the value of the variable. Once this is done, the continuous series becomes a discrete series and the HM is easily calculate.

Example 14. From the following data, calculate Harmonic Mean:

Class-interval :	10-20	20-30	30-40	40-50	50-60
Frequency :	30	75	70	135	220

Solution.

Calculation of Harmonic Mean

Class Interval (X)	**Frequency (f)**	**Mid-value (m.v.)=m**	**f/m**
10–20	30	15	2
20–30	75	25	3
30–40	70	35	2
40–50	135	45	3
50–60	220	55	4
	N = 530		$\sum$ (f/m) = 14

$$\text{H.M.} = \frac{N}{\sum\left(\frac{f}{m}\right)} = \frac{530}{14} = 37.86.$$

9.12 CONCLUSION

Measure of Central Tendency is a typical value around which other figures congregate. Thus it is clear that an average is single value which represents a whole series and is supposed to contain its measure characteristics.

9.13 FURTHER STUDY

1. Monga, G.S., Elementary Statistics.

2. Gupta, S.B., Principles of Statistics.

3. Alhance, D.N., Statistics.

UNIT-10 DISPERSION

Objectives

After going through this unit you should be able to know about the–

1. Range.
2. Quartile Deviation and Mean Deviation, Standard Deviation and Coefficient of Variation.
3. Lorenz Curve.

Structure

10.1 INTRODUCTION

Averages fail to reveal the full details of the distribution. Two or three distributions may have the same average but still they may differ from each other in many ways. In such cases, rather statistical analysis of the data is necessary so that these differences between various series can be studied and accounted for such analysis will make our results more accurate and we shall be more confident of our conclusions.

Suppose, there are three series of nine items each as follows:

Series A	Series B	Series C
40	36	1
40	37	9
40	38	20
40	39	30
40	40	40
40	41	50
40	42	60
40	43	70
40	44	80
Total 360	360	360
Mean 40	40	40

In the first series, the mean is 40 and the value of all the items is identical. The items are not at all scattered, and the mean fully discloses the characteristics of this distribution. However, in the second case, though the mean is 40 yet all the items of the series have different values. But the items are not very much scattered as the minimum value of the series is 36 and the maximum is 44 in the range. In this case also, mean is a good representative of the series because the difference between the mean and other items is not very significant. In the third series also, the mean is 40 and the values of different items are also different, but here the values are very widely scattered and the mean is 40 times of the smallest value of the series and half of the maximum value. Though the mean is the same in all the three series, yet the series differ widely from each other in their formation. Obviously, the average does not satisfactorily represent the individual items in this group and to know about the series completely, further analysis is essential. The scatter among the items in the first case is nil, in the second case it varies within a small range, while in the third case the values range between a very big span and they are widely scattered. It is evident from the above, that a study of the extent of the scatter around average should also be made to throw more light on the composition of a series. **The name given to this scatter is dispersion.**

10.2 DEFINITION

Some important definitions of dispersion are given below:

(i) "Dispersion or spread is the degree of the scatter or variation of the variable about a central value." – *Brooks and Dick*

(ii) "Dispersion is the measure of the variations of the items." – *A.L. Bowley*

(iii) "The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data." – *Spiegel*

From the above definitions, it is clear that in a general sense the term dispersion refers to the variability in the size of items. If the variation is substantial, dispersion is said to be considerable and if the variation is very little, dispersion is insignificant.

Usually, in a precise study of dispersion the deviations of size of items from a measure of central tendency are found out and then these deviations are averaged to give a single figure representing the dispersion of the series. This figure can be compared with similar figures representing other series. Such comparisons give a better idea about the formation of series than a mere comparison of their averages.

Averages of second order : For a precise study of dispersion, we have to average deviations of the values of the various items, from their average. We have seen earlier that arithmetic mean, median, mode, geometric mean and harmonic mean, etc., are all averages of the first order. Since in the calculation of measures of dispersion, the average values are derived by the use of the averages of the first order, the measures of dispersion are called averages of the second order.

12.3 OBJECTIVES OF MEASURING DISPERSION

Measures of variations are calculated to serve the following purposes:

(i) To judge the reliability of measures of central tendency.

(ii) To make a comparative study of the variability of two series.

(iii) To identify the causes of variability with a view to control it.

Spur and **Bonimi** have very rightly observed that, "in matters of health, variations in body temperature, pulse beats and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes

of which are sought through inspection and quality control programs."

In Social Sciences where we have to study problems relating to inequality in income and wealth, measures of dispersion are of immense help.

(iv) To serve as a basis for further statistical analysis.

10.4 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

The properties of a good measure of dispersion are the same as the properties of a good measure of central tendency. Precisely, they are:

(i) It should be rigidly defined.

(ii) It should be based on all the observations of the series.

(iii) It should be capable of further algebraic treatment.

(iv) It should be easy to calculate and simple to follow.

(v) It should not be affected by fluctuations of sampling.

10.5 DIFFERENT MEASURES OF DISPERSION

Absolute and relative "dispersion : Dispersion or variation can be expressed either in terms of the original units of a series or as an abstract figure like a ratio or percentage. If we calculate dispersion of a series relating to the income of a group of persons in absolute figures, it will have to be expressed in the unit in which the original data are, say, rupees. Thus, we can say that the income of a group of persons is Rs. 5000 per month and the dispersion is Rs. 500. This is called Absolute Dispersion. If, on the other hand, dispersion is measured as a percentage or ratio of the average, it is called Relative Dispersion. Since the relative dispersion is a ratio, it has no units. In the above case, the average income would be referred to as Rs. 5000 per month and the relative dispersion $\frac{500}{5000} = 0.1$ or 10%. In a comparison of the variability of two or more series, it is the relative dispersion that has to be taken into account as the absolute dispersion may be erroneous or unfit for comparison if the series are originally in different units.

10.6 RANGE

Range is the simplest possible measure of dispersion. It is the difference between the values of the extreme items of a series. Thus, if in a series relating to the weight measurements of a group of students, the lightest student has a weight of 40 kg. and the heaviest, of 110 kg. The

value of range would be 110 — 40 = 70 kg. This figure indicates the variability in the weight of students.

Symbolically,

$$\text{Range (R)} = L - S$$

where, L is the largest value and S the smallest value in a series.

Range as calculated above is an absolute measure of dispersion which is unfit for purposes of comparison if the distributions are in different units. For example, the range of the weights of students cannot be compared with the range of their height measurements as the range of weights would be in kg. and that of heights in centimetres. Sometimes, for purpose of comparison, a relative measure of range is calculated. If range is divided by the sum of the extreme items, the resulting figure is called "The Coefficient of the Range" or "The Coefficient of the Scatter."

Symbolically,

The Ratio of Range or the Coefficient of the scatter (or Range)

$$= \frac{\text{Max. value} - \text{Min. value}}{\text{Max. value} + \text{Min. value}} = \frac{L - S}{L + S}$$

$$= \frac{\text{Absolute range}}{\text{Sum of the extreme values}}$$

The following illustration would illustrates the use of the above formulae

Example 1. The profits of a company for the last 8 years are given below. Calculate the Range and its Coefficient:

Year	1975	1976	1977	1978	1979	1980	1981	1982
Profits (in '000 Rs.)	40	30	80	100	120	90	200	230

Solution.

Here, L = 230 and S = 30

Range = L – S = 230 – 30 = 200

Coefficient of Range = $\frac{L-S}{L+S}$ or $\frac{230-30}{230+30}$ or $\frac{200}{260} = 0.77$

Example 2. Calculate Co-efficient of Range from the following data :

Weekly Wages (Rs.)	No. of Labourers
50 – 60	50
60 – 70	45
70 – 80	45
80 – 90	40
90 – 100	35
100 – 110	30
110 – 120	30

Solution.

Coefficient of Range (first method) = $\frac{L-S}{L+S}$

Here, L = 120 and S = 50. such the

Coefficient of Range = $\frac{120-50}{120+50} = \frac{70}{170} = 0.41$

Coefficient of Range (second method) = $\frac{L-S}{L+S}$ Here, L = 115 and S = 55

Co-efficient of Range = $\frac{115-55}{115+55} = \frac{60}{70} = 0.35$

Example 3. Find the Range and the Co-efficient of range for the following observations 65, 70, 59, 81, 76,57,60,55, and 50.

Solution. Highest value = 82

Lowest value = 50

Range = 82 – 50 = 32

Coefficient of Range = $\frac{82-50}{82+50} = \frac{32}{132} = 0.2424$

Merits and Demerits of Range

As has been pointed out that' a good measure of dispersion should be rigidly defined, easily calculated, readily understood, should be capable of further mathematical treatment and should not be much affected by fluctuations of sampling.

Out of these the only merit possessed by Range is that it is easily calculated and readily understood.

As against this, the Range as a measure of dispersion has the following demerits :

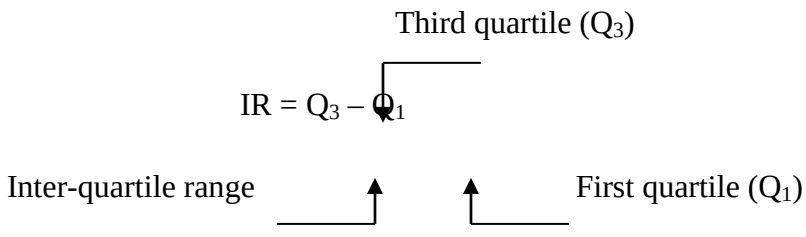
1. It is affected greatly by fluctuation of Sampling.
2. It is not based on all the observations of the series.
3. It cannot be used in case of open distributions.

Uses of Range

With all its limitations Range is commonly used in certain fields. For example:

1. Quality Control.
2. Variation in Money Sales, Share values, Exchange Rates and Gold Prices, etc.
3. Weather forecasting.

10.7. INTER-QUARTILE RANGE



Just as in a case of range the difference of extreme items is found, similarly, if the difference in the values of two quartiles is calculated, it would give us what is called the Inter-Quartile Range. Inter-Quartile range is also a measure of dispersion. It has an advantage over range, in as much

as, it is not affected by the values of the extreme items. In fact, 50% of the values of a variable are between the two quartiles and as such the inter-quartile range gives a fair measure of variability. However, the inter-quartile range suffers from the same defects from which range suffers. It is also affected by fluctuations of sampling and IS not based on all the observations of a series.

SEMI-INTER-QUARTILE RANGE

Semi-inter-quartile range, as the name suggests is the midpoint of the inter- quartile range. In other words, it is one-half of the difference between the third quartile and the first quartile. Symbolically,

$$\text{Semi-inter-quartile range or quartile deviation} = \frac{Q_3 - Q_1}{2}$$

where Q_3 and Q_1 are the upper and lower quartiles respectively.

In a symmetrical series median lies halfway on the scale form Q_1 to Q_3. In a symmetrical distribution Q_3 – median = Median – Q_1 or median = $\frac{Q_3 - Q_1}{2} = Q_3 - \left(\frac{Q_3 - Q_1}{2}\right) = Q_1 + \left(\frac{Q_3 - Q_1}{2}\right)$. If, therefore, the value of the quartile deviation is added to the lower quartile or subtracted from the upper quartile in a symmetrical series, the resulting figure would be the value of the median. But, generally series are not symmetrical and in a moderately asymmetrical series Q_1 + quartile deviation or Q_3 – quartile deviation would not give the value of the median. There would be a difference between the two figures and the greater the difference, the greater would be the extent of departure from normality.

Quartile deviation is an absolute measure of dispersion. If it is divided by the average value of the two quartiles, a relative measure of dispersion is obtained. It is called the Co-efficient of Quartile deviation.

$$\text{Co-efficient of a quartile deviation} = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 1. Find the Quartile Deviation and its Coefficient from the following data, relating to the weekly wages of seven labourers :

Weekly Wages (Rs.)	50	70	80	60	65	40	90

Solution

Calculation of Quartile Deviation

Wages arranged in ascending order would be as follows :

40 50 60 65 70 80 90

Q_1 = the value of $\left(\frac{N+1}{4}\right)^{th}$ or $\left(\frac{7+1}{4}\right)^{th}$ or 2^{nd} item = Rs. 50

Q_3 = the value of $3\left(\frac{N+1}{4}\right)^{th}$ or $\left(\frac{7+1}{4}\right)^{th}$ or 6^{th} item = Rs. 80

Quartile Deviation = $\frac{Q_3 - Q_1}{2} = \frac{80-50}{2}$ = Rs. 15

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{30}{130} = 0.23$

Discrete Series

Example 2. Calculate Quartile Deviation and its Coefficient from the following data:

Weight (in pounds)	120	122	124	126	130	140	150	160
No. of Students	1	3	5	7	10	3	1	1

Solution

Computation of Quartile Deviation and its Coefficient

Weight (in pounds)	120	122	124	126	130	140	150	160
Frequency	1	3	5	7	10	3	1	1
Cumu. Frequency	1	4	9	16	26	29	30	31

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{th} \text{ or } \left(\frac{31+1}{4}\right)^{th} \text{ or } 8^{th} \text{ item} = 124$$

$$Q_3 = \text{Size of } \left(\frac{31+1}{4}\right)^{th} \text{ or } 24^{th} \text{ item} = 130$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{130-124}{2} = 2 \text{ pounds.}$$

$$\text{Coefficients of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{130-124}{130+124} = \frac{6}{254} = 0.0236$$

Continuous Series (Exclusive)

Example 3. Calculate Semi-Inter-Quartile Range and its Coefficient of Q.D. from the following data:

Marks	No. of Students
0 – 10	11
10 – 20	18
20 – 30	25
30 – 40	28
40 – 50	30
50 – 60	33
60 – 70	22
70 – 80	15
80 – 90	22

Solution

Calculation of Quartile Deviation

Marks (X)	Frequency (f)	Cumulative Frequency (cf)
0 – 10	11	11
10 – 20	18	29
20 – 30	25	54
30 – 40	28	82
40 – 50	30	112
50 – 60	33	145
60 – 70	22	167
70 – 80	15	182
80 – 90	22	204

Q_1 = the value of $\left(\frac{204}{4}\right)^{th}$ or 51^{st} item which is in 20–30 group.

Q_3 = the value of $\left(\frac{204}{4}\right)^{th}$ or 153^{rd} item which is in 60–70 group.

The value of $Q_1 = l_1 + \frac{l_2 - l_1}{f_1}\left(\frac{N}{4} - c\right) = 20 + \frac{10}{25}(51 - 29) = 28.8$

The value of $Q_3 = l_1 + \frac{l_2 - l_1}{f_1}\left(\frac{3N}{4} - c\right) = 60 + \frac{10}{25}(153 - 145) = 63.64$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{63.64 - 28.8}{2} = 17.42 \text{ marks}$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.64 - 28.8}{63.64 + 28.8} = 0.37$$

Merits and Demerits of Quartile Deviation

Merits :

(i) It is simple to calculate and easy to understand.

(ii) It is useful in studying dispersion in open-ended series.

(iii) It is not affected by the values of extreme items, and, as such, in distributions which are highly asymmetrical or skew, quartile deviation is a better measure of dispersion than those which take into account the values of all the items of a distribution (like Mean Deviation or Standard Deviation).

Demerit

(i) It is not based on all the observations of the data as it ignores the first 25% and the last 25% of the items.

(ii) It is not capable of further algebraic treatment. It is in a way a positional average. It does not study variation of the values of a variable from any average. It merely indicates a distance on a scale.

(iii) It is affected considerably by fluctuations of sampling, A change in the value of a single item may, in many cases, affect its value considerably.

(iv) In such series where there is a considerable variation in the values of various items, the quartile deviation is not a suitable measure of dispersion. The sum of the absolute deviations from the median is less the Sum of the absolute deviation from the mean. Therefore, the value of the mean deviation from the median is always less than the value calculated from the mean.

10.8. MEAN DEVIATION

The range, the inter-quartile range and the quartile deviation suffer from a common defect, Le., they are calculated by taking into account only two values of a series—either the extreme values as in case of range, or the values of the quartiles as in case of quartile deviation. This method of studying dispersion (by location of limits) is also called the "Method of Limits". As we have seen earlier, such measures of dispersion suffer from many limitations and have many demerits.

It is, therefore, always better to have such a measure of dispersion which takes into account all the observations of a series and is calculated in relation to a central value. The method of calculating dispersion by calculating the deviations of all the values about a central value is called the method of averaging deviations. In this method, the deviations of items from a measure of central tendency are averaged to study the dispersion of the series. Mean deviation is such a measure of dispersion, "Mean deviation of a series is the arithmetic average of the absolute deviations of various items from a measure of central tendency (either mean, median or mode).

Theoretically, deviations can be taken from any of the three averages mentioned above but in actual practice, mean deviation is calculated either from mean or from median. Mode is usually not considered, as its value is indeterminate, and it gives erroneous conclusions. Between mean and median the median is supposed to be better than the mean, because the sum of the Absolute deviations from the medians is always less than the sum of the absolute deviations from the mean. Mean deviation is also known as the first moment of dispersion. Symbolically,

(i) $$\delta\overline{X} = \frac{\Sigma|d\overline{X}|}{N}$$

where $\delta\overline{X}$ stands for the mean deviation from mean, $d\overline{X}$ for the deviations from the mean, and N for the number of items.

(ii) $$\delta M = \frac{\Sigma|dM|}{N}$$

where δM stands for the mean deviation from median, dM for the deviations from the median, and N for the number of items.

(iii) $$\delta Z = \frac{\Sigma|dZ|}{N}$$

where δZ stands for the mean deviation from mode, dZ for deviations from the mode, and N for the number of items.

Mean deviation or first moment of dispersion, as calculated above would be an absolute measure of dispersion, expressed in the same units in which the original data are. In order to transform it into a relative measure, it is divided by the average from which it has been calculated. It is then known as the Mean coefficient of dispersion.

Thus, mean co-efficient of dispersion from mean, and median would be repetitively:

$$\frac{\delta\overline{X}}{X}, \frac{\delta M}{M} \text{ and } \frac{\delta Z}{Z}$$

Calculation of Mean Deviation and its Coefficient

1. **Series of Individual observations:** There are two methods of calculating mean deviation from a series of individual observation: one is the Direct method and the other is the Short-cut Method.

Direct method: In this method, the mean deviation is calculated by totalling the deviations from the mean, median or mode (plus, minus signs ignored) and dividing the total by the number of items. Thus,

$$\text{Mean Deviation} = \frac{\Sigma|d|}{N}$$

Short-cut method : In this method. mean or median is calculated and the total of the values of the items below the mean or median and above it are found out. The former is subtracted from the latter and divided by the number of items. The resulting figure is the Mean deviation:

Symbolically,

(i) $$\delta M = \frac{1}{N}(My - Mx)$$

where δM stands for the mean deviation from median, My for the total of the values above the actual median. and Mx for total of the values below the median and N for the number of items.

(ii) $$\delta\overline{X} = \frac{1}{N}(\overline{X}y - \overline{X}x)$$

where $\delta\overline{X}$ stands for the mean deviation from mean, $\overline{X}y$ stands for the total of the values above the actual arithmetic average and $\overline{X}x$ for the total of the values below the actual A.M. The following examples would illustrate these formulae :

Example 9. The following are the marks obtained by a batch of 9 students in a certain test:

Sl. No.	Marks (Out of 100)	Sl. No.	Marks (Out of 100)
1	68	6	38
2	49	7	59
3	32	8	66
4	21	9	41
5	54		

Calculate the mean deviation of the series.

Solution.

Direct method. Calculation of mean deviation of the series of marks of 9 students (arranged in ascending order of magnitude).

Students	Marks (X)	Deviations from median (49) (+and – signs ignored) (dM)
1	21	28
2	32	17
3	38	11
4	41	8
5	49	0
6	54	5
7	59	10
8	66	17
9	68	19
		Σ|dM| = 115

Median = value of $\left(\frac{N+1}{2}\right)^{th}$ item = 49 marks.

Mean deviation from the median or $\delta M = \frac{\Sigma|dM|}{N}$

where $\Sigma|dM|$ represents the summation of the deviations from the median, and N, the number of items.

$$\delta M = \frac{115}{9} \text{ marks} = 12.8 \text{ marks.}$$

Short-cut Method : Marks arranged in ascending order of magnitude.

Marks (X)	21	32	38	41	49	54	59	66	68

Sum of items above median (with values less than median) =

(21 + 32 + 38 + 41) = 132 (Mx)

Sum of items below median (with values more than median) =

$$(54 + 59 + 66 + 68) = 247 \text{ (My)}$$

Mean Deviation = $\frac{1}{N}(My - Mx) = \frac{1}{9}(247 - 132) = \frac{1}{9} \times 115 =$ 12.8 marks.

Example 13. Calculate the mean deviation (from mean) of the following data:

Marks	**No. of Students**
5	5
15	8
25	15
35	16
45	6

Solution

Calculation of mean deviation

Marks (X)	**Step Deviation from as av. (25) (d')**	**No. of students (f)**	**(fd')**	**Deviation from actual average (27) (dX)**	**(fd $\overline{X}$)**
5	– 2	5	– 10	22	110
15	– 1	8	– 8	12	96
25	0	15	0	2	30
35	+ 1	16	+ 16	8	128
45	+ 2	6	+ 12	18	108
	$\Sigma f = 50$	$\Sigma fd = + 10$		$\Sigma \lvert fd\,\overline{X} \rvert = 472$	

Arithmetic average of $\overline{X} = \left(\frac{10}{50} \times 10\right) = 27$.

Mean deviation = $\frac{\sum |fd\overline{X}|}{\sum f} = \frac{472}{50}$ marks = 9.44 marks.

2. **Calculation of mean deviation in continuous series:** The calculation of mean deviation in continuous series is done by the same procedure by which it is done in discrete series. In the short-cut method also the same procedure is followed provided the assumed mean or median is in the same class-interval in which the actual mean or median is. If the assumed average is in a different class interval, further adjustments are necessary. The following examples would illustrate these procedures:

Example 14. Calculate the mean deviation (from median) from the following data :

Class interval	Frequency	Class interval	Frequency
1–3	6	9–11	21
3–5	53	11–13	26
5–7	85	13–15	4
7–9	56	15–17	4

Solution. Direct and short-cut method : The median of the above series is 6.5.

Class Interval	Mid-points	Deviation from		Dev. x Freq-uency	Dev. from assumed median (g)	Dev. x Freq-uency
		Actual median (6.5)	Freq-uency (f)			
X	m	(dM)	(f)	(fdM)	(dM)	(fdM)

1–3	2	4.5	6	27.0	4	24
3–5	4	2.5	53	132.5	2	106
5–7	6	0.5	85	42.5	0	0
7–9	8	1.5	56	84.0	2	112
9–11	10	3.5	21	73.5	4	84
11–13	12	5.5	16	88.0	6	96
13–15	14	7.5	4	30.0	8	32
15–17	16	9.5	4	38.0	10	40
Total			**245**	**515.5**		**494**

Direct Method: Mean deviation = $\frac{\sum fdM}{N} = \frac{515.5}{245} = 2.1$

Short-cut Method : Total of deviations from assumed median = 494

No. of items with values less than the actual median (6.5) = (6+53+85) = 144

No. of items with values more than the actual median = (56+21+16+4+4) = 101

Difference between actual and assumed median = (6.5 – 6) = 0.5

Total deviation from actual median (when, actual and assumed medians are in the same class interval)

= 494 + (l44 × 0.5) – (101 × 0.5) = 494 + 72 – 50.5 = 515.5

Mean deviation = $\frac{515.5}{245} = 2.1$

Example 15. Calculate the mean deviation (from mean) from the following data :

Marks :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Students	6	5	8	15	7	6	3

Solution.

Calculation of Mean deviation direct and short-cut methods

Marks (X)	**Mid value (mv)**	**No. of students (f)**	**Dev. from assumed average dx (35)**	**Step Deviation dx/i**	**Total Dev. from a.av. fdx**	**Dev. from actual mean d $\overline{X}$ (+– signs ignored)**	**Total Dev. from actual fd $\overline{X}$**
(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**
0–10	5	6	– 30	– 3	– 18	28.4	170.4
10–20	15	5	– 20	– 2	– 10	18.4	92.0
20–30	25	8	– 10	– 1	– 8	8.4	67.2
30–40	35	15	0	0	0	1.6	24.0
40–50	45	7	+ 10	+ 1	+ 7	11.6	81.2
50–60	55	6	+ 20	+ 2	+ 12	21.6	129.6
60–70	65	3	+ 30	+ 3	+ 9	31.6	94.8
		$\Sigma f = 50$			**– 8**		**ΣfdX= 659.2**

Arithmetic Average = $A + \left(\frac{\Sigma dx}{N} \times i\right) = 35 + \left(\frac{-8}{50} \times 10\right) = 33.4$

Mean Deviation (Direct Method) = $\frac{\Sigma|fd\overline{X}|}{N} = \frac{659.2}{50} = 13.18$ marks.

10.9. STANDARD DEVIATION

The concept of standard deviation was first used by Karl Pearson in the year 1893. It is the most commonly used measure of dispersion. It satisfies most of the properties laid down for an ideal measure of dispersion.

Meaning: The technique of the calculation of mean deviation is mathematically illogical as in its calculation, the algebraic signs are ignored. This drawback is removed in the calculation of standard deviation, where squares of the deviations from the mean one used. Standard deviation is the square root of the arithmetic average of the squares of the deviations measured from the mean. The standard deviation is conventionally represented by the Greek letter Sigma σ.

Symbolically, $$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

Where σ stands for the standard deviation,

Calculation of Standard deviation

1. **Series of individual observation:** In such a series the standard deviation can be calculated in any of the following ways:

Direct method No. 1 : In this method, the following steps are involved:

(i) Find the arithmetic average of the series.

(ii) Find the deviations of each item from the arithmetic average and denote it by (d) i.e., find $(X - \overline{X})$ for each X.

(iii) Square these deviations and total them to find Σd^2.

(iv) Divide Σd^2 by the number of items to find $\frac{\Sigma d^2}{N}$. This figure is called the second moment about N the Mean.

(v) Standard Deviation = VN = $\sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$

Example 17. Calculate the standard deviation of the heights of 10 students given below :

Heights (in cms.) :160, 160, 161, 162, 163, 163, 163, 164,164, 170

Solution. Calculation of Standard Deviation of heights.

Height in centimeters	**Deviations from mean 163**	**Deviations squared (d^2)**
160	– 3	9
160	– 3	9
161	– 2	4
162	– 1	1
163	0	0
163	0	0
163	0	0
164	+ 1	1
164	+ 2	1
170	+ 7	49
		$\Sigma d^2 = 74$

Arithmetic average or $X = \frac{\Sigma X}{N} = \frac{1630}{10} = 163$ Cms.

Standard Deviation or $\sigma = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{74}{10}} = \sqrt{7.4} = 2.27$ Cms.

Second method for Calculating S.D.

Calculation of Standard deviation

Discrete series. In discrete series also standard deviation can be calculated by

(i) The direct method as well as by

(ii) The short-cut method. Further it is possible to have step deviations, both in the direct and short-cut methods :

(a) Direct method:

1. Calculate the arithmetic mean.

2. Find out the deviations (d) of the various values, from the mean value. Square these deviations (d^2).

3. Multiply d^2 with the respective frequencies (f) against various values and add all such values (Σfd^2).

4. Divide Σfd^2 by the number of items (N) and find out the square root of the figure so obtained i.e., find $\sqrt{\frac{\Sigma fd^2}{N}}$. This will be the value of the Standard Deviation.

(b) Short-cut Method

1. Assume a mean (A) and take deviations (dx) from it and square them up (dx^2).

2. Multiply dx^2 with the respective frequencies or f to get (fdx^2). Total them to get (Σfdx^2).

3. Divide (Σfdx^2) by the number of items or N to get $\frac{\left(\Sigma fdx^2\right)}{N}$.

4. From $\left(\frac{\Sigma fdx^2}{N}\right)$ subtract the square of the difference between actual and assumed average $\left(\overline{X} - A\right)^2$ to get $\left(\frac{\Sigma fdx^2}{N}\right) - \left(\overline{X} - A\right)^2$.

5. Find out the square root of the above value or $\sqrt{\frac{\left(\Sigma fdx^2\right)}{N} - \left(\overline{X} - A\right)^2}$ and it will be the value of the standard deviation. This formula can also be written as :

$$\sigma = \sqrt{\frac{\left(\Sigma fdx^2\right)}{N} - \left(\frac{\Sigma fdx}{N}\right)^2} \quad \text{as} \quad \overline{X} - A = \frac{\Sigma fdx}{N}$$

Example 22. Calculate standard deviation for the following distribution :

Values	10	20	30	40	50	60	70
Frequency	1	5	12	22	17	9	4

Solution

Calculation of standard deviation (Step Deviation Method)

Values (X)	**Freq-uency (f)**	**Dev. from assumed av. 40** $\frac{X-40}{10}=dx$	**Total Step Dev. fdx**	dx^2	fdx^2
10	1	– 3	3	9	9
20	5	– 2	– 10	4	20
30	12	– 1	– 12	1	12
40	22	0	0	0	0
50	17	+ 1	+ 17	1	17
60	9	+ 2	+ 18	4	36
70	4	+ 3	+ 12	9	36
	N = 70		$\Sigma fdx = +22$		$\Sigma fdx^2 = 130$

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2} \times i = \sqrt{\frac{130}{70} - \left(\frac{+22}{70}\right)^2} \times 10 = \sqrt{1.757} \times 10$$

$$= 1.326 \times 10 = 13.26$$

Thus, the standard deviation is 13.26.

Example 23. Calculate the standard deviation for the following table giving the age distribution of 542 members of the house of Common :

Age	No. of Members
20–30	3
30–40	61
40–50	132
50–60	153
60–70	140
70–80	51
80–90	2
Total	**542**

Solution. Calculation of the standard deviation of the age distribution of 542 members of the house of Commons :

Age Group (X)	**Mid-Value (mv)= x**	**Freq-uency (f)**	**Devia-tions from the assumed av. (55) dx**	**fdx**	**Square of devia-tions dx^2**	**Freq-uency × square of devia-tions fdx^2**
20–30	25	3	– 30	– 90	900	2700
30–40	35	61	– 20	– 1220	400	24400
40–50	45	132	– 10	– 1320	100	13200
50–60	55	153	0	0	0	0
60–70	65	140	+ 10	+ 1400	100	14000
70–80	75	51	+ 20	+ 1020	400	20400
80–90	85	2	+ 30	+ 60	900	1800
		N = 542		**Σfdx =150**		**Σfdx^2=76500**

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - \left(\frac{\sum fdx}{N}\right)^2} = \sqrt{\frac{76500}{542} - \left(\frac{-150}{542}\right)^2}$$

$= \sqrt{141.07} = 11.9$ years.

In the above example, step deviations were not taken, so the calculation become cumbersome. In the following example, step deviation method has been illustrated.

10.10 LORENZ CURVE

Dispersion can be studied graphically also with the help of what is called Lorenz Curve, after the name of Dr. Lorenz who first studied the dispersion of distribution of wealth by the graphic method. The technique of drawing Lorenz Curve is not very difficult. The technique is as follows:

(i) The size of items as well as frequencies are first cumulated.

(ii) Then taking the final cumulated figure as 100 percentages are calculated for all cumulative values.

(iii) On the X-axis begin from 100 to 0 and let it represent frequencies (This is not a hard and fast rule. X-axis can begin with 0 to 100 and can represent values instead of frequencies. However, the suggested procedure gives a more convenient curve).

(iv) On the y-axis begin from 0 to 100 and let it represent values. (It can also begin from 100 to 0 and represent frequencies).

(v) Draw a line from 0 of the X-axis to the 100 of y-axis. This is the line of equal distribution.

(vi) Plot the various points of x and y and draw the curve. If the distribution is not proportionately equal, the curve would be away from the line of equal distribution. The farther is the curve from the line of equal distribution, the greater is the variability in the series.

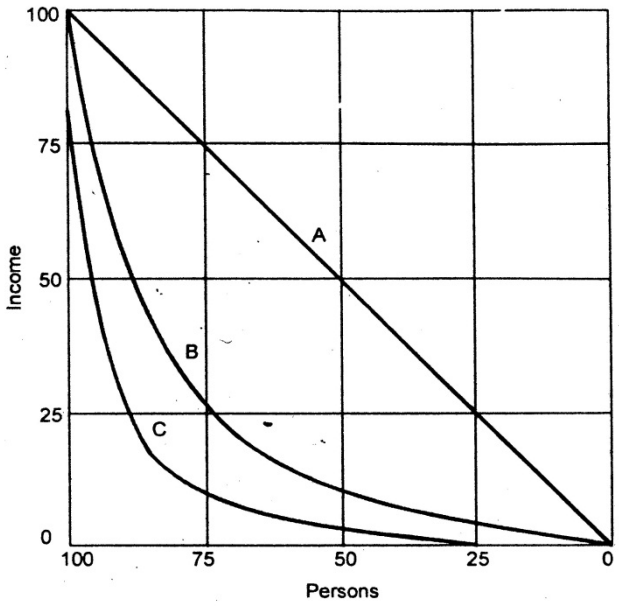
Example. Draw a Lorenz curve from the following data:

Income in thousand rupees	**Number of persons in thousands**		
	Group A	**Group B**	**Group C**
10	5	8	15
20	10	7	6
40	20	5	2
50	25	3	1
80	40	5	1

To draw the Lorenz curve from the above data, the size of the item and frequencies would have to be cumulated and then percentages would have to be calculated by taking the respective totals as 100. This has been done in the following table:

Rup-ees (000)	**Cumu-lated Income**	**Cum. %**	**No. of per-sons (000)**	**Cum-Num-ber**	**Cum. %**	**No. of per-sons (000)**	**Cum-Num-ber**	**Cum. %**	**No. of per-sons (000)**	**Cum-Num-ber**	**Cum. %**
10	10	5	5	5	5	8	8	32	15	15	60
20	30	15	10	15	15	7	15	60	6	21	84
40	70	35	20	35	35	5	20	80	2	23	92
50	120	60	25	60	60	3	23	92	1	24	96
80	200	100	40	100	100	2	25	100	1	25	100

Now, the cumulative percentages would be plotted on a graph paper. Percentages relating to the number of persons would be shown on the abscissa and from left to right the scale would begin with 100 and end with 0. The income percentages would be shown on the ordinate and here the scale will begin with 0 at the bottom and go upto 100 at the top. The above percentages would give the following type of curve.



From the above figure, it is clear that in the first group of persons, the distribution of income is proportionately equal, so that 5% of the income is shared by 5% of the population, 15% of the income by 15% of the population and so on. It gives the line of equal distribution. In the second group, the distribution is uneven so that 5% of the income is shared by 32% of the people and

15% of the income by 60% of the people. In the third group, the distribution is still more unequal so that 5% of that income is shared by 60% of the people and 15% of the income by 84% of the people. The variation in group C is, thus, greater tnan the variation in group B. Curve C is, thus, at a greater distance from the line of equal distributions, than curve B.

The Lorenz curve has a great drawback. It does not give any numerical value of the measure of dispersion. It merely gives a picture of the extent to which a series is pulled away from an equal distribution. It should be used along with some numerical measure of dispersion. It is very useful in the study of income distributions, distributions of land and wages, etc.

10.11 CONCLUSION

Dispersion is the degree of scatter or variation of variable about a central value. In simple words Dispersion refers to the variability in the size of the items.

10.12 FURTHER STUDY

1. Alhance, D.N., Fundamental of Statistics.

2. Gupta, S.B., Principles of Statistics.

3. Monga, G.S., Elementary Analysis.

UNIT-11 CORRELATION AND REGRESSION

Objectives

After going through this unit you should be able to know about the–

1. Karl Pearson's Coefficient of Correlation.
2. Spearman's Rank Correlation.
3. Regression.

Structure

11.1 INTRODUCTION

Meaning : In various types of analyses, we have confined ourselves to such series where various Items assumed different values of one variable. We have discussed how measures of central tendency and measures of dispersion and skewness are calculated in such cases for purposes of comparison and analysis. With the help of these measures, such data can be easily understood. There can, however, be such series also where each item assumes the values of two or more variables. For example, if the heights and weights of a group of persons are measured, we shall get such series where each member of the group would assume

two values—one relating to height and the other relating to weight. If, besides heights and weights, the chest measurements were also taken, each member of the group would assume three values relating to three different variables. In such cases, we can calculate averages, dispersion and skewness, etc., in accordance with the rules given in the previous chapters.

But, sometimes, it appears that the values of the various variables so obtained are interrelated. It is likely that such relationship may be obtained in two series relating to the heights and weights of a group of persons. It may be observed that weights increase with increase in heights so that tall people are heavier than short sized people. Similarly, if the data are collected about the prices of a commodity and the quantities sold at different prices, two series would be obtained. One variable would be the various prices of the commodity, and the other variable would be the quantities sold at these prices. In two such series, we are again likely to find some relationship. With increase in the price of the commodity, the quantity sold is bound to decrease. We can, thus, conclude that there is some relationship between price and demand. Such relationships can be found in many types of series, for example, prices and supply, heights and weights of persons, prices of sugar and sugarcane, ages of husbands and wives, etc.

The term correlation (or co-variation) indicates the relationship between two such variables in which, with changes in the values of one variable, the values of the other variable also change if the two variables move together.

11.2. DEFINITION

Some important definitions of correlation are given below:

(1) "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated". — *L.R. Connor*

(2) "Correlation means that between two series or groups of data, there exists some casual connection" . — *W.L. King*

(3) "Correlation analysis attempts to determine the 'degree of relationship' between variables. — *Ya Lun Chow*

(4) "When the relationship of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

The above definitions make it clear that the tern correlation refers to the study of relationship between two or more variables.

11.3 TYPES OF CORRELATION

Correlation can be :

(i) Positive or Negative;

(ii) Simple, Multiple or Partial;

(iii) Linear or Non-linear.

(i) Positive or Negative Correlation

Correlation can be either positive or negative. When the values of two variables move in the same direction, i.e., when an increase in the value of one variable is associated with an increase in the value of the other variable and a decrease in the value of one variable is associated with the decrease in the value of the other variable, correlation is to be positive.

If, on the other hand, the values of two variables move in opposite directions, so that with an increase in the values of one variable, the values of the other variable decrease, and with a decrease in the values of one variable the values of the other variable increase, correlation is said to be negative. There are some data in which correlation is generally positive while, in others, it is negative. Thus, generally, price and supply are positively correlated. When prices go up, supply also increases and with the fall in prices, supply also decreases. The correlation between price and demand is generally negative. With an increase in price, the demand goes down and with a decrease in price the demand generally goes up. Demand curve is downward sloping, whereas, supply curve is upward sloping.

(ii) Simple, Multiple and Partial Correlation

In simple correlation, we study only two variables say, price and demand. In multiple correlation, we study together the relationship between three or more factors like production, rainfall and use of fertilizers. In partial correlation, though more than two factors are involved but correlation is studied only between two factors and the other factors are assumed to be constant.

(iii) Linear and Non-linear (Curvi-linear) Correlation

The correlation between two variables is said to be linear if corresponding to a unit change in the value of one variable, there is a constant change in the value of the other variable, i.e., in case of linear correlation the relation between the variables x and y is of the type

$$y = a + bx.$$

If $a = 0$, the relation becomes $y = bx$ in such cases, the values of the variables are in constant ratio. The correlation between two

variables is said to be non-linear or curvilinear if corresponding to a unit change in the value of one variable the other variable does not change at a constant rate but at a fluctuating rate.

11.4 METHODS OF DETERMINING CORRELATION

The various methods by which correlation studies are made, are as follows:

(i) Scatter Diagram

(ii) Correlation Graph

(iii) Coefficient of Correlation

(iv) Coefficient of Correlation by Rank Differences

(v) Coefficient of Concurrent Deviation

(vi) Method of Least Square.

1. Coefficient of Correlation

Purpose of Calculation : Coefficient of correlation is calculated to study the extent or degree of correlation between two variables. As has been said, earlier, the fact that there is correlation between two variables does not mean that their relationship is functional or constant. If the value of a variable is known, it is not always possible to obtain the exact value of the other variable. This can be done only where there is linear relationship between the two variables. There are a few series in which linear relationship exists, e.g., natural numbers and their squares or square roots would always give a linear relationship. Similarly, linear relationship would be obtained between two series, one relating to radii of various. circles and the other relating to their areas. In economic data, such relationships are rarely found. No doubt, demand would fall with an increase in price, but the relationship is not functional. There is no constant ratio between the variation of the two series rating to price and demand.

Perfect Correlation: If the relationship between two variables is such that with an increase in the value of one, value of the other increases or decreases, in a fixed proportion, correlation between them is said to be perfect. If both the series move in the same direction and the variations are proportionate, there would be perfect positive correlation between them. If, on the other hand, the two series move in reverse directions, and the variations in their values are always proportionate, it is an example of perfect negative correlation. It is also likely that there may be no

relationship between the variations of the two series in which case there is said to be no correlation between them.

As has been said earlier, in economic data, perfect positive or negative correlation is usually not found, as the relationship between economic series is rarely functional. In such data, correlation is not perfect as the related series are not completely dependent on each other. Perfect correlation is obtained when there is complete mutual dependence between the two series.

It would be observed figs. 1 and 2 that all corresponding values of x and y are in a straight line. figure 1 indicates perfect positive correlation between x and y as the variation in the values of the two series are always in a fixed proportion and they move in the same direction. figure 2, on the other hand, shows a perfect negative correlation between x and y as the variations between their values are in a constant ratio and two series move in reverse directions.

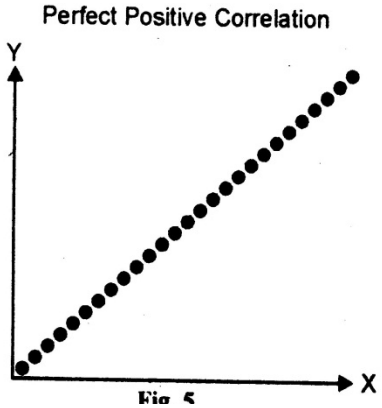


Fig. 1

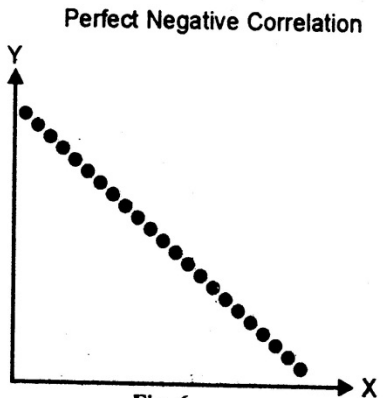


Fig. 2

After knowing this, it is necessary to obtain such a measure of correlation which can accurately indicate the degree of correlation in quantitative terms. The measure should be such that its extreme values represent perfect positive and perfect negative correlations and the value in the middle, absence of correlation. Such a measure is given by the coefficient of correlation.

The coefficient of correlation which we are going to discuss in the following pages always varies between the two limits of + 1 and – 1. When there is perfect positive correlation, its value is + 1 and when there is perfect negative correlation its value is – 1. Its midpoint is 0, which indicates absence of correlation. As the value of this coefficient decreases from the upper limit of + 1, the extent of positive correlation between the two variables also declines. When it reaches the value of 0, it indicates complete absence of correlation and when it goes further down in negative values (less than zero) it indicates negative correlation. When it reaches the other limit of – 1 there is evidence of perfect negative correlation between the two series.

The above mentioned points can be studied from the graphs which have been given so far. When the values of the variable are like those given in figure 1, there is perfect positive correlation or the value of the coefficient of correlation is + 1, when they are like those given in figure 1, there is positive correlation but it is not perfect, or the value of the coefficient of correlation is less than + 1 but more than 0. When the values are like those given in Figure 3, there is no correlation between the data or the value of the coefficient of correlation is 0; when the values are like those given in Figure 2, there is negative correlation though not perfect, which means that the value of the coefficient of correlation would be more than 0 (on the negative side) but less than – 1. If, however, the values of the variable are like those given in figure 2, there. is perfect negative correlation or in other words, the value of the coefficient of correlation would be – 1.

11.5 CALCULATION OF COEFFICIENT OF CORRELATION

(Karl Pearson's Formula)

Karl Pearson, the great biologist and statistician, has given a formula for the calculation of coefficient of correlation. According to it the coefficient of correlation of two variables is obtained by dividing the sum of the products of the corresponding deviations of the various items of the two series from their respective means by the product of their standard deviations and the number of pairs of observations.

If X_1, X_2, X_3, ... X_n are the values of the first variable and Y_1, Y_2, Y_3, Y_n are the values of the second variable, then:

$$\text{Correlation coefficient (r)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n\,\sigma_x\sigma_y} \quad \text{or} \quad r = \frac{\sum dx\,dy}{n\,\sigma_x\sigma_y}$$

Thus, if x_1, x_2, x_3, ... x_n are the deviations of various items of the first variable from their mean value and y_1, y_2, y_3, ... y_n are the corresponding deviations of the second variable from its mean value, the sum of the products of these corresponding deviations would be Σxy. If, further, the standard deviations of the two variables are respectively σx, σy and if n is the number of pairs of observations, Karl Pearson's coefficient of correlation represented by r would be :

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{n\,\sigma_x\sigma_y} \qquad \text{.....(i)}$$

It is clear from the above formula that if Σxy is positive, the coefficient of correlation would also be a positive figure indicating positive correlation between the two series. If, on the other hand, Σxy is negative, coefficient of correlation would also be negative, indicating that the 'correlation between the two series is negative, Σxy would be positive, if generally, positive and negative deviations in one series are associated with positive and negative deviations in the other series also, i.e., the deviations in both the series have the same sign. The value of Σxy would be negative, if the positive deviations of one variable are associated with the negative deviations in the other variable and vice versa. The deviations is both the series are of opposite sign. If positive and negative deviations of one variable are indifferently associated with the deviations of the other variable, the value of would be 0 or near it, indicating absence of correlation between the two series. The value of this coefficient of correlation always lies between + 1 and – 1. It cannot exceed unity.

The above formula of Karl Pearson is based on the study of covariance between two series. The covariance between two series is written as follows :

$$\text{Covariance (X, Y)} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n}$$

To study the correlation, the covariance of the two series is divided by the product of their standard deviations. Thus,

$$r = \frac{\text{covariance of the two series}}{\left(\text{variance of series 1}\right)\left(\text{variance of series 2}\right)} = \frac{\text{Covariance X, Y}}{\sigma_x \times \sigma_y}$$

Since the S.D. is independent of the change of origin $\sigma_x = \sigma_x$, $\sigma_y = \sigma_y$

$$\frac{\sum xy}{n\,\sigma_x \sigma_y} \quad \text{Where } x = X - \overline{X},\ y = Y - \overline{Y}$$

This formula is known as the **Product moment formula of Coefficient of Correlation.**

Calculation of the Pearson's Coefficient of Correlation

$$r = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt[n]{\frac{\Sigma\left(X - \overline{X}\right)^2}{n}}\sqrt{\frac{\Sigma\left(Y - \overline{Y}\right)^2}{n}}} = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}\sqrt{\frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2}}$$

Now, $$\frac{\sum(X-\overline{X})(Y-\overline{Y})}{n}=\frac{\sum XY-\sum X\overline{Y}-\sum\overline{X}Y+\sum\overline{XY}}{n}$$

$$=\frac{\sum XY}{n}-\overline{Y}\frac{\sum X}{n}-\overline{X}\frac{\sum Y}{n}+\frac{\sum\overline{XY}}{n}$$

$$=\frac{\sum XY}{n}-\overline{XY}-\overline{XY}+\overline{XY}=\frac{\sum XY}{n}-\left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)$$

$\because$ $$\overline{X}=\frac{\sum X}{n}\ ,\quad \overline{Y}=\frac{\sum Y}{n}$$

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

If the deviations of x-series are taken from any value A and that of Y-series taken from the value B, Then.

$$X-\overline{X}=X-\overline{X}+A-A=(X-A)-(\overline{X}-A)=dx-\overline{d}x$$

$$Y-\overline{X}=Y-\overline{Y}+B-B=(Y-B)-(\overline{Y}-B)=dy-\overline{d}y$$

By making the x substitution in A we get

$$r=\frac{\sum(dx-\overline{d}x)\ \sum(dy-\overline{d}y)}{n\sqrt{\frac{\sum(dx-\overline{d}x)^2}{n}}\sqrt{\frac{\sum(dy-\overline{d}y)^2}{n}}}$$

$$r=\frac{\frac{\sum d_x d_y}{n}-\left(\frac{\sum d_x}{n}\right)\left(\frac{\sum d_y}{n}\right)}{n\sqrt{\frac{\sum d_x^2}{n}-\left(\frac{\sum d_x}{n}\right)^2}\sqrt{\frac{\sum d_y^2}{n}-\left(\frac{\sum d_y}{n}\right)^2}}$$

If the deviations are taken from their respective actual means, then

$$r=\frac{\sum xy}{n\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

where $x = X - \overline{X}$ $y = Y - \overline{Y}$

Correlation Coefficient is independent of the change of origin as well as change of scale.

The following solved examples would illustrate the use of the above rules.

Example. Find out the coefficient of correlation between the sales and expenses of the following 10 firms (figure in '000 Rs.)

Firms :	1	2	3	4	5	6	7	8	9	10
Sales :	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution

Calculation of Coefficient of Correlation

Sales (X)	**Dev. from mean (58) (x)**	**Dev. square (x^2)**	**Expenses (Y)**	**Dev. from mean (14) (y)**	**Dev. square (y^2)**	**Product of deviation (xy)**
50	– 8	64	11	– 3	9	+ 24
50	– 8	64	13	– 1	1	+ 8
55	– 3	9	14	0	+ 0	0
60	+ 2	4	16	+ 2	4	+ 4
65	+ 7	49	16	+ 2	4	+ 14
65	+ 7	49	15	+ 1	1	+ 7
65	+ 7	49	15	+ 1	1	+ 7
60	+ 2	4	14	0	0	0
60	+ 2	4	13	– 1	1	– 2
50	– 8	64	13	– 1	1	+ 8
$\Sigma X = 580$ $n = 10$	$\Sigma x = 0$	$\Sigma x^2 = 360$	$\Sigma Y = 140$	$\Sigma y = 0$	$\Sigma y^2 = 22$	$\Sigma xy = 70$

$$\text{Mean of X} = \frac{\sum X}{n} = \frac{580}{10} = 58$$

$$\text{Mean of Y} = \frac{\sum Y}{n} = \frac{140}{10} = 14$$

$$r = \frac{\sum xy}{n\sqrt{\frac{\sum x^2}{n}}\sqrt{\frac{\sum y^2}{n}}} = \frac{+70}{10\sqrt{\frac{360}{10}}\sqrt{\frac{22}{10}}}$$

$$= \frac{+70}{10\sqrt{36}\sqrt{2.2}} = 0.786$$

Alternatively : $$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{+70}{\sqrt{360 \times 22}} = +0.786$$

Example. Calculate Coefficient of Correlation between the values of X and Y given below :

Values of X:	65	66	67	67	68	69	70	72
Values of Y:	67	68	65	68	72	72	69	71

Solution

Calculation of Coefficient of Correlation

(X)	**(Y)**	**(X^2)**	**(Y^2)**	**(XY)**
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\Sigma(X)$ = 544	$\Sigma(Y)$ = 552	$\Sigma(X)^2$ = 37028	$\Sigma(Y)^2$ = 38132	$\Sigma(XY)$ = 37560

Coefficient of correlation :

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

$$= \frac{\frac{37560}{8} - \frac{544}{8} \times \frac{552}{8}}{\sqrt{\frac{37028}{8} - \left(\frac{544}{8}\right)^2}\sqrt{\frac{38132}{8} - \left(\frac{552}{8}\right)^2}}$$

$$\frac{4695 - 4692}{\sqrt{4628.5 - 4624}\sqrt{4766.5 - 4761}} = \frac{3}{4.975} = 0.6030$$

Example. Calculate Karl Pearson's Coefficient Correlation between X and Y for the following information :

$n = 12$, $\Sigma X = 120$, $\Sigma Y = 130$, $\Sigma(X - 8)^2 = 150$,

$\Sigma(Y - 10)^2 = 200$ and $\Sigma(X - 8)(Y - 10) = 50$ [C.A. May, 1992]

Solution.

$\Sigma x = \Sigma(X - 8) = \Sigma X - \Sigma 8 = 120 - 8 \times 12 = 24$

$\Sigma y = \Sigma(Y - 10) = \Sigma Y - \Sigma 10 = 130 - 120 = 10$

$\Sigma xy = \Sigma(X - 8)(Y - 10) = 50$ (given)

$\Sigma x^2 = \Sigma(X - 8)^2 = 150$; $\Sigma y^2 = \Sigma(Y - 10)^2 = 200$

$$r_{xy} = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}}$$

$$= \frac{\frac{50}{12} - \frac{24}{12} \times \frac{10}{12}}{\sqrt{\frac{150}{12} - \left(\frac{24}{12}\right)^2}\sqrt{\frac{200}{12} - \left(\frac{10}{12}\right)^2}} = 0{,}2146 \textbf{ Ans.}$$

Assumption of Pearsonian Correlation

The Pearsonian coefficient of correlation rests on two assumptions.

The first is that a large number of independent contributory causes are operating in each of the two series correlated so as to produce normal or probability distribution. We know that such causes always operate in chance phenomena like tossing of coin or throw of a dice. They also operate in other types of data. For example, such forces are usually found operating in phenomena like indices of price and supply, ages of husbands and wives and heights of fathers and sons, etc.

The second assumption is that the forces so operating are not independent of each other but are related in a casual fashion. If the forces are entirely independent and unrelated, there cannot be any correlation between the two series. The forces must be common to both the series. The height of an individual during the last ten years may show an upward trend and his income during this period may also show a similar tendency but there cannot be any correlation between the two series because the forces affecting the two series are entirely unconnected with each other. If the coefficient of correlation in such series is calculated, it may even be +.8 indicating a very high degree of positive correlation, but such correlation is usually termed nonsense correlation because the two series are affected by such sets of forces which are entirely unconnected with each other.

In the words of Karl Pearson, "The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributing causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

11.6 RANK CORRELATION

Sometimes, such problems are faced where it is possible to arrange the various items of a series in serial order but the quantitative measurement of their value is difficult; for example, it is possible for a class teacher to arrange his students in ascending or descending order of intelligence, even though intelligence cannot be measured quantitatively. No doubt, the quantitative study about the intelligence of students can be made by holding an examination and assigning them marks, but this method can never be said to be infallible. There are many such attributes which are incapable of quantitative measurements; for example, honesty, character, morality, beauty etc.

If it is desired to have a study of association between two such attributes, say, intelligence and beauty, the Karl Pearson's Coefficient of Correlation cannot be calculated as these attributes cannot be assigned definite values. However, there is a method by which we can study correlation between such attributes. This method was developed by the British psychologist, Charles Edward Spearman, in the year 1904.

In this method, x and y variables denote the rank of the attributes A and B. In case of a study of correlation between intelligence and beauty, we can pick up 10 or 20 or any other number of individuals and first arrange them in order of their rank according to intelligence-beginning with the most intelligent person whose rank would be 1 and going down in order till the rank of the last person is indicated. Similarly, we can arrange these individuals again in order of rank according to beauty the most beautiful person getting Rank No.1 and going down to the last person who is the least beautiful.

In this way, we will have two sets of ranks for these two attributes. If 10 individuals are arranged like this, we will have ranks from 1 to 10 for each attribute.

Once this is done, we can find out a Coefficient of Correlation between these two series by the Spearman's Rank Correlation formula which is as under :

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \quad \text{or} \quad 1 - \frac{6\sum d^2}{n^3 - n} \quad \text{or}$$

$$r = 1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)}$$

where r_s denotes the Spearman's Rank Correlation and d denotes the difference of the corresponding ranks of the same individual in the two attributes, and n denotes the number of pairs.

The value of the rank correlation coefficient is interpreted in the same way as that of the Pearson's Coefficient of Correlation. Its value also ranges between + 1 and – 1. When r_s is + 1 it indicates complete agreement in the order of ranks between the two attributes (the most intelligent being the most beautiful also, and so on). If correlation is – 1 it indicates complete disagreement in the order of ranks (the most intelligent being the one who is the least beautiful).

The rank correlation coefficient will be equal to – 1 if the ranks assigned by the judges are exactly inverse i.e., an individual who gets the highest score from one judge, gets the lowest scores from the other and the individual getting the lowest score from one judge gets the highest from the other and so on.

There are two types of problems in calculating this coefficient :

(A) When actual ranks are given.

(B) When actual ranks are not given.

In each of these two types of problem, a difficulty arises when ranks of two individuals are the same. Such problems need a modification in the formula given above. (See example as Equal ranks).

(A) Where Ranks are Given

Where actual ranks of the items are given the steps that have to be taken to get the coefficient are as follows :

(i) Compute the difference of ranks (R_1 — R_2) and denote them by d.

(ii) Compute d^2 and total them to get Σd^2.

(iii) Use the formula given below to get the Coefficient:

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

The following examples will illustrate the above method.

Example. The values of the same 15 students in two subjects A and B are given below; the two numbers within the brackets denoting the ranks of the same student in A and B, respectively:

(1, 10) (2,7) (3,2) (4, 6) (5, 4) (6,8) (7,3) (8, 1) (9, 11) (10, 15) (11,9) (12,5) (13, 14) (14, 12) (15, 13)

Use Spearman's formula to find the rank Correlation Coefficient.

Solution.

Rank in A R_1	**Rank in B** R_2	$(R_1 - R_2)$ **d**	d^2
1	10	– 9	81
2	7	– 5	25
3	2	– 1	1
4	6	– 2	4
5	4	– 1	1
6	8	– 2	4
7	3	+ 4	16
8	1	– 7	49
9	11	– 2	4
10	15	– 5	25
11	9	– 2	4
12	5	– 7	49
13	14	– 1	1
14	12	– 2	4
15	13	– 2	4
n = 15		$\Sigma d = 0$	$\Sigma d^2 = 272$

Calculation of Rank Correlation Coefficient

Spearman's Coefficient of Correlation or

$$r_s = 1 - \frac{6\sum d^2}{n^2 - n}$$

Substituting the values, we get :

$$r_s = 1 - \frac{6 \times 272}{15^2 - 15} = 1 - \frac{1632}{3375 - 15} = 1 - \frac{1632}{3360} = 1 - \frac{17}{35} = \frac{18}{35} = +0.51$$

Example. Calculate the coefficient of rank correlation from the following data :

X	60	34	40	50	45	41	22	43	42	66	64	46
Y	75	32	34	40	45	33	12	30	36	72	41	57

Solution

Calculation of Coefficient of Rank Correlation.

X	**Rank 1**	**y**	**Rank 2 of Ranks (d)**	**Difference**	**d^2**
60	3	75	1	+ 2	4
34	11	32	10	+ 1	1
40	10	34	8	+ 2	4
50	4	40	6	– 2	4
45	6	45	4	+ 2	4
41	9	33	9	0	0
22	12	12	12	0	0
43	7	30	11	– 4	16
42	8	36	7	+ 1	1
66	1	72	2	– 1	1
64	2	41	5	– 3	9
46	5	57	3	+ 2	4
n = 12				0	$\Sigma d^2 = 48$

Coefficient of rank correlation or

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(48)}{12\left(12^2 = 1\right)} = 1 - \frac{288}{1716} = \frac{1428}{1716} = 0.82$$

Example. Calculate the coefficient of correlation from the following data by the Spearman's Rank Difference method :

Prices of Tea (Rs.)	Prices of Coffee (Rs.)	Prices of Tea (Rs.)	Prices of Coffee (Rs.)
75	120	60	110
88	134	80	140
95	150	81	142
70	115	50	100

Solution.

Calculation of Coefficient of Rank Correlation

Prices of Tea (X)	R_1	Prices of Coffee (Y)	R_2	$R_1 - R_2$ (d)	d^2
75	4	120	4	0	0
88	7	134	5	2	4
95	8	150	8	0	0
70	3	115	3	0	0
60	2	110	2	0	0
80	5	140	6	1	1
81	6	142	7	1	1
50	1	100	1	0	0
n = 8		**n = 80**		**0**	**$\Sigma d^2 = 6$**

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 6}{8^3 - 8} = 1 - \frac{36}{512 - 8}$$

$$= 1 - \frac{36}{504} = \frac{468}{504} = 0.921 \quad \textbf{Ans.}$$

Equal Ranks

Sometimes, where there is more than one item with the same value, a common rank is given to such items. This rank is the average of the ranks which these items would have got had they differed slightly from each other. When this is d(,me, the coefficient of rank correlation needs some correction, because the above formula is based on the supposition that no item is repeated, the ranks of various items are different and that no rank is given to more than one item.

If in a series, there are 'm' items whose ranks are common, then for correction of the coefficient of rank correlation 1/12 ($m^3 - m$) is added to the value of (Σd^2). If there are more than one such groups of items with common rank, this value is added as many times as the number of such groups.

The formula for the calculation of Rank Correlation is, thus, modified in the following manner:

$$r_s = 1 - \frac{6\left[\sum d^2 + \sum \frac{1}{12}\left(m^3 - m\right)\right]}{n^3 - n}$$

The following examples would illustrate the use of this modified formula :

Example. Calculate the coefficient of rank correlation from the following data:

X	48	33	40	9	16	16	65	24	46	57
Y	13	13	24	6	15	4	20	9	6	19

Solution.

Calculation of Coefficient of Rank Correlation

X	Rank 1	Y	Rank 2	Diff. of Ranks (d)	d^2
48	3	13	5.5	– 2.5	6.25
33	5	13	5.5	– 0.5	.25
40	4	24	1	– 3.0	9.00
9	10	6	8.5	– 1.5	2.25
16	8	15	4	+ 4.0	16.00
16	8	4	40	– 2.0	4.00
65	1	20	2	– 1.0	1.00
24	6	9	7	– 1.0	1.00
16	8	6	8.5	– 0.5	.25
57	2	19	3	– 1.0	1.00
n = 10		**n = 10**		**0**	**Σd^2 = 41.00**

In the above table, in X-series, figure 16 occurs three times. The rank of all these items is 8 which is the average of 7, 8, and 9 — the ranks which these items would had there been some difference between their values. In Y-series figures 13 and 6 both occur two times. Their ranks are respectively 5.5 and 8.5. Due to these common ranks, the coefficient of rank correlation would have to be corrected.

For correction, we shall add [1/12 (m^3 – m)] to the value of [Σd^2]. In X-series, this value would be equal to [1/12 (3^3 – 3)] as the value 16 has occurred three times in this series. In Y-series, there are two such groups of common ranks. In the first group, this correction would be [1/12 (2^3 – 2)] as the value 13 has occurred twice and for the second group also the correction value would be [1/12 (2^3 – 2)] as the value has also occurred twice in this series.

Spearman's Coefficient of Rank Correlation or

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[41 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right]}{10^3 - 10}$$

$$= 1 - \frac{6\left(41 + 2 + 0.5 + 0.5\right)}{990} = 1 - \frac{264}{990} = \frac{726}{990} = +0.73$$

Example. A firm, not sure of the response to its product in ten different colour shades, decides to produce them in those colour shades.

The two judges rank the 10 colours in the following order :

Colour No.	1	2	3	4	5	6	7	8	9	10
Ranking by Judge I	6	4	3	1	2	7	9	8	10	5
Judge II	4	1	6	7	5	8	10	9	3	2

Is there any agreement between the two judges, to allow the introduction of the product by the firm in the market?

D	6–4	4–1	3–6	1–7	2–5	7–8	9–10	8–9	10–3	5–2
|D|	2	3	3	6	3	1	1	1	7	3
D^2	4	9	9	36	9	1	1	1	49	9
$\Sigma D^2 = 128$										

Rank correlation = $1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)} = 1 - \frac{6 \times 128}{10\left(100 - 1\right)} = 1 - 0.78 = 0.22$

Ans.

Since the rank correlation is low, there is not much agreement between the two judges, and the firm may not produce the product to introduce it in the market.

Example. Calculate the Rank Coefficient from the sales and expenses of 10 firms as given below:

Sales (X)	50	50	55	60	65	65	65	60	60	50
Expenses (Y)	11	13	14	16	16	15	15	14	13	13

Solution.

Calculation of Coefficient of Rank Correlation

Firms	**Sales X**	**Expenses**	**Ranks**		**$R_1 - R_2$ = (d)**	**d^2**
			R_1	**R_2**		
1	50	11	9	10	– 1	1.00
2	50	13	9	8	1	1.00
3	55	14	7	5.5	1.5	2.25
4	60	16	5	1.5	3.5	12.25
5	65	16	2	1.5	0.5	0.25
6	65	16	2	3.5	– 1.5	2.25
7	65	15	2	3.5	– 1.5	2.25
8	60	14	5	55	– 0.5	0.25
9	60	13	5	8	– 3	9.00
10	50	13	9	8	1	1.00
						$\Sigma d^2 = 31.50$

In sales, largest figure is 65 and it occurs 3 times. So the rank corresponding to them are 1, 2 and 3. As all figures are same, their allotted

ranks should also be equal, i.e., $\frac{1+2+3}{3}=2$ for each figure. Allotting ranks in this manner, in sales, 3 figures get rank 2 each, 3 figures get rank 5 each and 3 figures get rank 9 each. In expenses 2 figures get the rank 1.5 each, 2 figures get 3.5 each, two figures get rank 5.5 each and three figures get ranks 8 each.

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \ldots\right]}{n^3 - n}$$

$$= 1 - \frac{6\left[31.50 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right]}{10^3 - 10}$$

$$= 1 - \frac{6(31.50 + 9.5)}{990} = 1 - \frac{6 \times 41}{990} = 1 - \frac{246}{990} = \frac{744}{990} = +0.75$$

Please Note: Sometimes, another case arises when the highest rank given in the series is more than the number of pairs. In such cases, the given ranks have to be treated as values and fresh ranks assigned.

Example. Explain the distinction between regression and correlation. Calculate the coefficient of correlation from the following data by the method of rank differences:

Rank of X	10	4	2	5	8	5	6	9
Rank of Y	10	6	2	5	8	4	5	9

[CA (Foundation) May 1994]

Solution. Though the given data is in the form of ranks but it cannot be used as ranks because the highest rank exceeds the number of pairs. In such cases, ranks are taken as values and fresh ranks are determined.

X [Given ranks of X treated as values]	Actual ranks	Y [Given ranks of Y treated as values]	Actual ranks	D Rank difference	D^2
10	1	10	1	0	0
4	7	6	4	3	9
3	8	2	8	0	0
5	5.5	5	5.5	0	0
8	3	8	3	0	0
5	5.5	4	7	– 1.5	2.25
6	4	5	5.5	– 1.5	2.25
9	2	9	2	0	0
					$\Sigma D^2 = 13.50$

Coefficient of correlation

$$r = 1 - 6\left\{\frac{\sum D^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{N\left(N^2 - 1\right)}\right\}$$

Here, m_l, m_2 ...., etc. denote the number of times values or the ranks are tied in the variables. The subscripts denote the first tie, second tie, ..., in both the variables.

$$r = 1 - 6\left\{\frac{13.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)}{8\left(8^2 - 1\right)}\right\} = 1 - \frac{6 \times 14.5}{8 \times 63} = 0.827 \quad \textbf{Ans.}$$

Merits and Demerits of the Rank Correlation

Merits

(1) Since in this method ΣD or the sum of the differences between R_1 and R_2 is always equal to zero, it provides a check on the calculation.

(2) Since Spearman's Rank Correlation is the same thing as Karl Pearson's Coefficient of Correlation between ranks, it can be interpreted in the same way as Karl Pearson's Coefficient of correlation.

(3) Rank correlation unlike Karl Pearson's Coefficient of Correlation does not assume normality in the universe from which the sample has been taken.

(4) Rank Correlation is very easy to understand and apply. However, Pearson's Coefficient is based on a set of full information while Spearman's Coefficient is based only on the ranks. The values of r obtained by these two methods would generally differ.

(5) Spearman's Rank method is the only way of studying correlation between qualitative data which cannot be measured in figures but can be arranged in serial order.

Demerits

(1) The method cannot be used in two-way frequency tables or bivariable frequency distribution.

(2) It can be conveniently used only when n is small, say, 30, otherwise calculation becomes tedious.

11.7 REGRESSION

Introduction

The dictionary meaning of the word regression is 'stepping back' or 'going back'. The use of this word dates back to the time of early studies made by **Francis Galton** in the latter half of the nineteenth century. He studied the relationship between the heights of fathers and their sons, and arrived at some very interesting conclusions. which are given below :

(i) Tall fathers have tall sons and short fathers have short sons.

(ii) The mean height of the sons of tall fathers is less than the mean height of their fathers.

(iii) The mean height of the sons of short fathers is more than the mean height of their fathers.

(iv) Galton found that the deviations in the mean height of the sons from the mean height of the race was less than the deviations in the mean height of the fathers from the mean height of the race. When the fathers move above the mean or below the mean, the sons tended to go back or regress towards the mean.

Regression, thus, implies going back or returning towards the mean. Galton studied the average relationship between these two variables graphically and called the line describing the relationship the line of regression.

Regression lines, thus, study the average relationship between two series and throw light on their Covariance. If the Coefficient of Correlation between the heights of fathers and sons is +0.7 it means that if a group of fathers have heights which are more than the average by x inches, their sons could have heights which would be more than average by 0.7 inches. Thus, the height of the sons regresses towards the mean. The study of this tendency is the subject matter of regression.

At this stage, we shall examine some definitions of this term.

1. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."

2. "Regression analysis attempts to establish the, nature of the relationship' between variables. that is, to study the functional relationship between the variables and thereby provide a mechanism for predicting, or forecasting."

11.8 UTILITY OF REGRESSION

The above definitions make it clear that regression analysis is done for estimating or predicting the unknown value of one variable from the known value of the other variable. This is a very useful statistical tool which is used both in natural and social sciences.

In the field of business, this tool of statistical analysis is very widely used. Businessmen are interested in predicting future production, consumption, investment, prices, profits, sales, etc. In fact, the success of a businessman depends on the correctness of the various estimates that he is required to make. In sociological studies and in the field of economic planning, projections of population, birth rates, death rates and other similar variables are of great use.

In our day to day life, we come across many variables which are interrelated. For example, with a rise in price, the demand of a commodity goes down. or with better monsoons the output of agricultural product increase, or the effect of expenditure on publicity may lead to a rise in the volume of sales. With the help of regression analysis, we can estimate or predict the effect of one variable on the other. e.g., we can predict the fall in demand when the price rises by a particular amount. However. in social sciences, there is multiple causation which means that a large number of factors affect various variables. The regression study which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression.

In the simple regression analysis, there are two variables—one of which is known as an 'independent variable' or 'regresser' or predictor or explanator. On the basis of the values of this variable, the values of the other variable are predicted. This other variable whose values are predicted is called the 'dependent' or 'regressed' or explained variable.

With the help of regression studies, we can also calculate the value of the Coefficient of correlation. The Coefficient of determination [square of coefficient of correlation] which measures the effect of the independent variable on the dependent variable gives us an indication about the predictive value of the regression studies.

11.9. COMPARISON OF CORRELATION AND REGRESSION STUDIES

Both the correlation and regression analysis helps us in studying the relation-ship between two variables yet they differ in their approach, and objectives.

1. Correlation studies are meant for studying the Co-variation of the two variables. They tell us whether the variables under study move in the same direction or in reverse directions. The degree of their co-variation is also reflected in the correlation coefficient, but the correlation study does not study the *nature of relationship.* It does not tell us about the relative movement in the variables under study and we cannot predict the value of one variable by taking into account the value of the other variable. This is possible through regression analysis, i.e., regression analysis can be used for prediction where correlation cannot be used for prediction.

2. Correlation between two series is not necessarily a cause and effect relationship. A high degree of positive correlation between price and supply does not mean that, supply is the effect of prices. There may be no cause and effect relationship between the variables under study and yet they may be correlated. Regression on the other hand presumes one variable as a cause and the other as its effect. The independent variable is supposed to be affecting the dependent variable and as such we can estimate the values of the dependent variable for a given value of the independent variable.

3. The Coefficient of correlation r varies between ± 1 i.e., $-1 \leq r \leq 1$. The regression coefficients have the same sings as the correlation coefficient. If r is positive regression coefficient would also be positive and if r is negative the regression coefficients would also be negative.

4. Further, whereas correlation coefficient cannot exceed unity, but one of the regression coefficients can have a value higher than unity but the product of the two regression coefficients can never exceed unity because correlation coefficient is the square root of the product of the two regression coefficients.

11.10 METHODS OF STUDYING REGRESSION

Broadly speaking, regression can be studied either :

(i) Graphically or

(ii) Algebraically

(i) Graphic Study of Regression

When regression is studied with the help of graphic methods, we have to draw a scatter diagram. A scatter diagram contains one point for one pair of values of X and Y variable. Usually, X variable is shown on the horizontal scale and Y variable on the vertical scale. When all related pairs of values have been plotted on a scatter diagram, we have to draw two regression lines to predict the values of X and Y variables. The regression line which is used to predict the values of Y for a value of X is called the Regression line of Y on X. Similarly, the regression line which is used to predict a value of X for a value of Y is called Regression line of X on Y. If the coefficient of correlation between X and Y is perfect, i.e., its value is either +1 or – 1, there will be only one regression line as the variations in the two series in such cases always increase or decrease by a constant figure. In other words, we say that the two regression lines will be identical if the correlation between the two variables is perfect.

If the graph between the values of a dependent variable and independent variable is a straight line, then the regression is called Linear Regression. If, however, the relationship between the two variables is not in the form of a straight tine but have some other functional relationship like $Y = X^2$, then regression between the variables is Non-linear regression. In this chapter, we are studying only linear regression.

How to draw linear regression lines

Regression lines can be drawn by :

(a) freehand curve method, or

(b) by the Method of Least Squares

Freehand Curve Method

In the freehand curve method, we first plot the pairs of the values of X and Y in the form of a scatter diagram-one point for one pair of values. After this, we draw two free hand straight lines. One of these lines is drawn in such a way that the positive deviations of Y-series from its mean are cancelled by the negative deviations. The sum of the deviations on one side of the line is equal to the sum of deviations on the other side. This will be the regression line of Y on X. The other regression line would be drawn in such a way that the positive deviations of X-series from its mean would cancel the negative deviations. This regression line would be called the regression line of X on Y. The two regression lines would cut each other at the point the coordinates of which represent the means of two

series. If there is perfect positive or negative correlations between the two variables, there will be only one regression line.

However, it is very difficult to draw regression lines by the freehand curve method. Usually, a piece of thread is repeatedly adjusted in such a manner in the scatter diagram that the positive and negative deviations cancel each other. Once these lines are drawn, we can predict or estimate the values of Y from the Regression line of Y on X and, similarly, the values of X can be predicted from the regression line X on Y.

Method of Least Squares

In order to avoid the difficulties associated with the drawing of regression lines by the freehand curve method. a mathematical relationship is established between the movements of X and Y series and algebraic equations are obtained to represents the relative movements of X and Y series.

In this method, we minimise the Sum of Squares of the deviations between the given values of a variable and its estimated values given by the line of the best fit. Line of Regression of Y on X is the line which gives the best estimate for the value of Y for a specified value of X and, similarly, the line of regression of X on Y is the line which gives the best estimate for the value of X for a specified value of Y.

If the values of Y are plotted on the Y axis (i.e., the vertical axis), then the regression line of Y on X will be such which minimises the sum of the squares of the vertical deviations. Similarly, if the values of X are plotted on the X axis (i.e., the horizontal axis) the regression line of X on Y will be such which minimises the sum of the squares of the horizontal deviations.

We have briefly discussed the method of least squares in the chapter on Correlation and pointed out that the line of the best fit is obtained by the equation of straight line $Y = a + bX$ and that in the method of least squares, this line is obtained with the help of the following two normal equations :

$$\Sigma Y = na + b\,(\Sigma X)$$

$$\Sigma XY = a\,\Sigma\,(X) + b\,\Sigma\,(X^2)$$

If the values of X and Y variables are substituted in the above equations, we get the values of a and b by solving these equations and thus, get the regression line of Y on X. Here, Y is the dependent variable and X the independent variable. To get the regression line of X on Y, we will have to assume X as the dependent variable and Y as the independent variable. We will then get the two equations for the two regression lines.

The following illustration will illustrate the above points.

Illustration

Plot the regression lines associated with the following data :

Values of X	1	2	3	4	5
Values of Y	166	184	142	180	338

Solution. To obtain the straight line equations of the given values, we will use the following normal equations.

$$\Sigma(Y) = na + b\,(\Sigma X)$$

$$\Sigma(XY) = a\Sigma X + b(\Sigma X^2)$$

These equations will give us the values of a and b which we will fit in the equation of the straight line,

$$Y = a + bX$$

This will give us the regression line of Y on X. The value of a in this problem would be a = 100 and b = 34. So the equation would be :

$$Y = 100 + 34X \text{ ..... regression equation of Y and X.}$$

To find the regression equation of X on Y, we have to find the values of a and b in the equation :

$$X = a + bY$$

and in this case, the two normal equations would be :

$$\Sigma(X) = na + b\,(\Sigma Y)$$

$$\Sigma(XY) = a\Sigma Y + b(\Sigma Y^2)$$

with these equations the equation, of the straight line for X on Y would be:

$$X = 0.172 + 0.014Y \text{ ..... regression equation of X and Y.}$$

as the value of a from the normal equations would be 0.172 and b = 0.014.

Thus, the equations of straight line or the lines of the best fit would be as follows:

$$Y = 100 + 34X \quad \text{..... (i)}$$

$$X = 0.172 + 0.014Y \quad \text{..... (ii)}$$

From the first equation, we will find any two values of Y for some values of X and plot them on the graph paper to get the Regression Line of Y on X.

From the second equation. we will find any two values of X for some values of Y and plot them on the graph paper to get the Regression Line on X on Y.

The two lines would cut each other at the point which gives, the average values of X and Y.

The following graph would emerge from these lines :

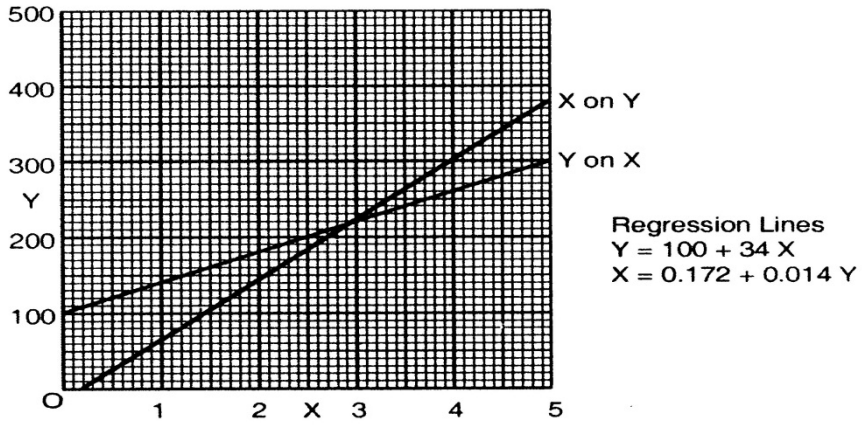


Fig. 1

It would be observed from the above graph that the two regression lines cut each other at the point of Arithmetic mean of the two series. If we have to find out any value of Y for a given value of X, we will draw a perpendicular from the X series (for the given value of X) and the point at which it cuts the regression line of Y on X will indicate the coupled value of Y which can be read on the Y-scale (by drawing a line parallel to X axis from the point where the perpendicular joins the regression line of Yon Y). Thus, when X = 2, Y would be 168. Similarly, values of Y series can be found by using the regression line of X on Y.

As was pointed out earlier, the regression line or the line of the best fit is one from which the square of the deviations between the given values of the variable and its estimated values is the least. In our problem, the Y series is on the Vertical scale and so the square of the deviations (measured on vertical scale) between' the original figures and the regression line would be the least. For the X series, the sum of squares on

the horizontal scale would be the least. The following two graphs illustrate these points:

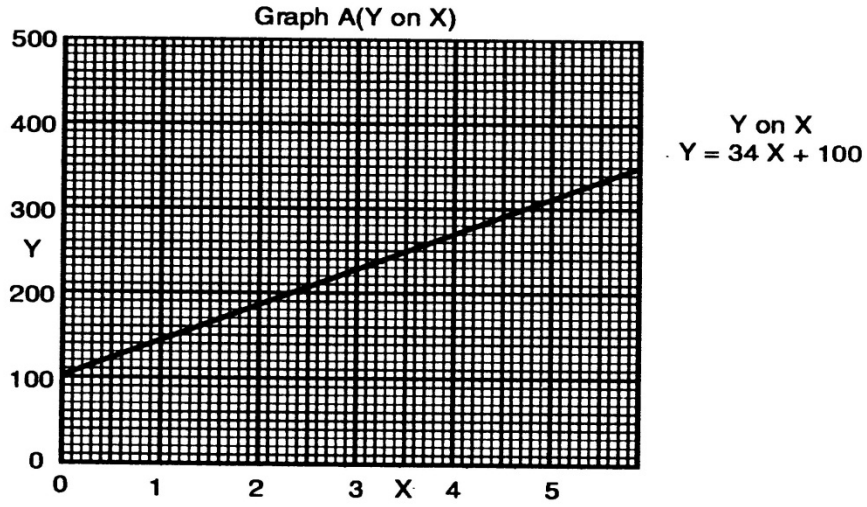


Fig. 2

In the above graph, besides the regression line of Y on X, we have plotted the original figures of Y and the perpendiculars show the deviation between original and computed figures. The sum of the square of these deviations would be the least when deviations are taken from the regression line values.

Fig. 3 shows similar deviations of the original values of X series with the values in the regression equation of X on Y.

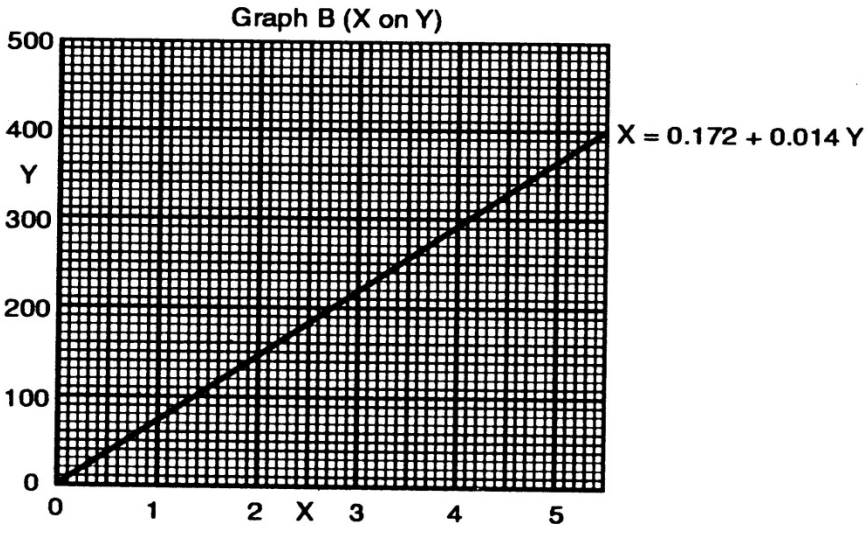


Fig. 3

From the graph, it is clear that the deviations of X series from the regression line are the least and so the sum of their squares would also be the least. In fact, the sum of the deviations is always zero and so the squares of the deviations have the least value.

Why Two Regression Lines

Very often, this question is asked as to why there should be two regression lines to obtain the values of Y and x : and why one regression line does not serve the purpose. The answer is simple and it is that one regression line cannot minimise the sum of squares of deviations for both the X and Y series unless the relationship between them indicates perfect positive or negative correlation. In case of perfect correlations, one regression line is enough because X and Y series have the same type of deviations. Ordinarily, in social sciences, perfect correlation is very rarely found. For this reason, one regression line minimises the sum of the squares of deviations of the X series and the other regression line takes care of the deviations of Y series. This point is obvious from Figs. 2 and 3. Fig. 2 has minimized the sum of the squares of deviations of Y series and Fig. 3, of the X series. In the first case, the deviations have been measured on the vertical scale and in the other on horizontal scale.

11.11 REGRESSION EQUATIONS

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations. Regression line of X on Y give the best possible mean values of X for given values of Y and, similarly, the regression line of Y on X gives the best possible mean values of Y for given values or X As such, regression equation of X on Y would be used to describe the variation in the values of X for given changes in the values of Y and, similarly, the regression equation of Y on X would be used to describe the variation in the value of Y for given changes in the values of X.

The regression equation of X on Y is $X = a + bY$ and the regression equation of Y on X is $Y = a + bX$. These are the equations of a straight line. In these equations, the values of a and b are constant which determine the positions of the lines of regression. The parameter a indicates the level of the line of regression (the distance of the line above or below the origin). The parameter b determines the slope of the line, i.e., the corresponding change in X in relation to per unit change in Y or vice versa.

The values of a and b in the above equations are found by the Method of Least Squares-reference to which was made earlier. The values of a and b are found with the help of normal equations given below :

(i) $\Sigma Y = na + b \Sigma X$

(ii) $\Sigma X = na + b \Sigma Y$

$$\Sigma XY = a \Sigma X + b \Sigma X^2 \qquad \Sigma XY = a \Sigma Y + b \Sigma Y^2$$

We shall take an example to illustrate this technique of finding out the values of a and b and with their help to obtain the regression equations.

Illustration

From the following data, obtain the two regression equations using the method of Least Squares.

X	2	4	6	8	10
Y	5	7	9	8	11

Solution

Computation of Regression Equations

X	**Y**	**XY**	**X^2**	**Y^2**
2	5	10	4	25
4	7	28	16	49
6	9	54	36	81
8	8	64	64	64
10	11	110	100	121
$\Sigma X = 30$	**$\Sigma Y = 40$**	**$\Sigma XY = 266$**	**$\Sigma X^2 = 220$**	**$\Sigma Y^2 = 340$**

Regression Equation Y on X is of the form $Y = a + bX$.

To find the values of a and b, the following two normal equations are used:

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

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

Substituting the values from the table we get:

$40 = 5a + 30b$ ..... (i)

$266 = 30a + 220b$ ..... (ii)

Multiplying equation (i) by 6 we get:

$240 – 30a + 180b$ ..... (iii)

Subtracting equation (iii) from equation (ii) we get:

$40\ b = 26$ or $b = 0.65$

Substituting the value of b in equation (i) we get:

$40 = 5a + 19.5$ or $5a = 20.5 \Rightarrow a = 4.1$

Substituting the values of a and b in the regression equation Yon X we get:

$Y = 4.1 + 0.65\ X \rightarrow$ This is the Regression Equation of Y on X

Regression Equation of X on Y is of the form $X = a + bY$

The two normal equation are :

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

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values in these equations we get:

$30 = 5a + 40b$ ..... (i)

$266 = 40a + 340b$ ..... (ii)

Multiplying equation (i) by 8 we get:

$240 = 40a + 320b$ ..... (iii)

Subtracting (iii) from (ii) we :

$20b = 26 \Rightarrow b = 1.3$

Substituting the value of b in equation (i) we get:

$30 = 5a + 52 \Rightarrow 5a = – 22$

or $a = – 4.4$

Substituting the values of a and b in the recession equation X on Y we get:

$X = – 4.4 + 1.3\ Y$. This is the regression equation of X on Y.

11.12 CONCLUSION

The term correlation indicates the relationship between two such variables in which with change of values of one variable the values of the other variable change if the two variable work together. Thus, correlation

analysis attempts to determine the degree of relationship between variables.

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

11.13 FURTHER STUDY

1. Sinha, V.C., Principles of Statistics.

2. Gupta, S.B., Principles of Statistics.

3. Monga, G.S., Elementary Statistics.

UNIT-12 PROBABILITY THEORY

Objectives

After going through this unit you should be able to know about the–

1. Methods of determining Correlation.
2. Concept and Types of Regression.
3. Distinction between Correlation and Regression.

Structure

12.1 MEANING AND DEFINITION OF PROBABILITY

One of the Primary reasons for the development of the Theory of Probability is the presence in almost every aspect of life, of random phenomena. A Phenomenon is random if change factors — determine its outcome. All the possible outcomes may be known in advance, but the particular outcome of a single trial in any experimental operation can't be pre-determined. Nevertheless, some regulatory is built into the process so that each of the possible outcomes can be assigned a probability fraction. Probability is especially important in statistics because of the many principles and Procedures that are based on this concept. Indeed Probability plays a special role in all our lives, and has an everyday meaning. Sometimes we hear phases like : 'You had better take an umbrella because it is likely to rain.' 'His changes of winning are preety small'. It is very likely that it will rain by the evening.' 'You are probably right.' Or 'There are fifty-fifty changes of his passing the examinations.' In each of these phrases an idea of uncertainty is acknowledged. Goeth remarked that, "There is nothing more frightful than action in ignorance."

12.2 PROBABILITY DEFINED

Ordinarily speaking, the probability of an event denotes the likelihood of its happening. The value of probability ranges between 0 and 1. If an event is certain to happen, its probability would be 1 and if it is certain that the event would not take place, then the probability of its happening is 0. Ordinarily, in social sciences, the probability of the happening of an event is rarely 1 or 0. The reason is that in social sciences we deal with situations where there is always an element of uncertainty about the happening or not happening of an event. For this reason, the probability of the events is somewhere between 0 and 1.

The general rule of the happening of an event is that if an event can happen in m ways and fail to happen in n ways, then the probability (p) of the happening of the event is given by :

$$p = \frac{m}{m+n}$$

or

$$p = \frac{\text{Number of cases favourable to the occurrence of the event, i.e., m}}{\text{Total number of mutually exclusive, equally likely and exhaustive cases, i.e., (m + n)}}$$

If $m = 0$, $p = 0$

If $n = 0$, $p = 0$

If m = 0, there is no case favourable to the occurrence of the event and the event is said to be impossible and if n = 0, there is no case unfavourable to the occurrence of the event and the event is said to be sure or certain.

odds in favour of the occurrence of the event

$$= \frac{n}{m} = \frac{\text{The number of cases favourable to the occurrence of the event}}{\text{The number of cases against the occurrence of the event}}$$

odds against the occurrence of the event

$$= \frac{m}{n} = \frac{\text{The number of cases against the occurrence of the event}}{\text{The number of cases favourable to the occurrence of the event}}$$

Notes : $P(A) + P(\overline{A}) = 1 \Rightarrow P(A) = 1 - P(\overline{A})$

i.e., the probability of the occurrence of an event, say, A.

= I – Probability of the occurrence of the event complementary to A.

e.g., if a universal set U = {1, 2, 3, 4, 5, 6, 7, 8, 9} and Set A = {1, 5, 6},

then $\overline{A}$ = {2, 3, 4, 7, 8, 9}

Probability of the occurrence of any event in the sample points of A

$$= P(A) = \frac{3}{9} = \frac{1}{3}, \text{ and } \quad P(\overline{A}) = 1 - \frac{1}{3} = \frac{2}{3}$$

(2) If the events are mutually exclusive and exhaustive, i.e., if the events are complementary the sum of their individual probabilities = 1.

(3) Odds in favour of the occurrence of the event and the odds against the occurrence are reciprocal of each other.

12.3 PROBABILITY THEOREMS

The solution to many problems involving. Probabilities requires a through understanding of some of basic rule which govern the manipulation of probabilities. They are generally called probability theorems. They are discussed below :

Addition Theorem : The theorem is stated as follows : "It two events are mutually exclusive and the Probability of the one is P_1 while that of other is P_2, the probability of either the one events or the other occurring is the sum $P_1 + P_2$".

For example, the Probability of getting spot (1) in a throw of a single die is 1/6 the Probability of getting spot (3) is also 1/6 and the Probability of getting spot (5) too is 1/6. The Probability of getting an old number (1, 3, 5) in a throw of a single die will be the addition of their respective Probabilities, that is —

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \quad \text{or} \quad \frac{1}{2}$$

The addition theorem will hold good only if :

(i) Item are mutually exclusive,

(ii) Mutually exclusive items belongs to some set.

Illustration

(a) A bag contains 4 white, 2 black, 3 yellow and 3 red balls. What is the Probability of getting a white or red ball at random in a single draw of one.

The Probability of getting one white balls = $\frac{4}{12}$

The Probability of getting one red ball = $\frac{3}{12}$

The Probability of getting one white or red ball

$$= \frac{4}{12} + \frac{3}{12} = \frac{7}{12} \quad \text{or} \quad \frac{7}{12} \times 100 = 58.3\%$$

= 58.3% **Ans.**

(b) Find out the Probability of getting a total of either 7 or 11 in a single throw with two dice.

A total of 7 cm come in 6 different ways —

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$$

A total of 11 can come 2 different ways —

$$\begin{pmatrix} 5 & 6 \\ 6 & 5 \end{pmatrix}$$

The Probability of getting a total of $7 = \frac{6}{36} = \frac{1}{6}$

The Probability of getting a total of $11 = \frac{2}{36} = \frac{1}{18}$

The Probability of getting either 7 or 11 = $\frac{1}{6} + \frac{1}{18} = \frac{4}{18} = \frac{2}{9}$

The addition theorem will hold good only if the events are mutually exclusive. If events contain no sample point in common, them some adjustment is necessary. Under such a case —

$$P[(A) \text{ or } (B)] = P(A) + P(B) - P(A \& B)$$

The following example will make it clear –

Example – A bag contain 25 balls, numbered from 1 to 25, one is to be drawn at random. Find the Probability that the number of the drawn ball will be multiple of 5 or of 7.

The Probability of number being multiple of

$$5\ (5, 10, 15, 20, 25) = \frac{5}{25} \text{ or } \frac{1}{5}$$

The Probability of the number being multiple of

$$7\ (7, 14, 21) = \frac{3}{25}$$

Thus the Probability of the number being a multiple of 5 or 7 will be –

$$\frac{5}{25} + \frac{3}{25} = \frac{8}{25}$$

in the above illustration, find the Probability that the number is multiple of 3 or 5 :

The Probability of the number being multiple of

$$3\ (3, 6, 9, 12, 15, 18, 21, 24) = \frac{8}{25}$$

The Probability of the number being multiple of

$$5\ (5, 10, 15, 20, 25) = \frac{5}{25}$$

Joint Probability $\frac{8}{25}+\frac{5}{25}=\frac{13}{25}$, but this answer is wrong, because item no. 15 is not mutually exclusive. Hence the correct Probability will be –

$$\frac{8}{25}+\frac{5}{25}-\frac{1}{25}=\frac{12}{25}$$

Hence, P(A + B) = P(A) + P(B) – P(AB).

Example – A card is drawn at random from an ordinary Pack of 52 playing cards. Find the Probability that a card drawn is either a spade or the ace of diamonds.

The Probability of drawing a spade = $\frac{13}{12}$.

The Probability of drawing and ace of diamonds = $\frac{1}{52}$.

Probability of drawing a spade on an ace of diamonds =

$$\frac{13}{52}+\frac{1}{52}=\frac{14}{52} \text{ or } \frac{7}{26}$$ **Ans.**

12.4 MULTIPLICATION THEOREM

According to this theorem, "If two events are mutually independent, and the Probability of the one is P_1 while that of the other is P_2, the Probability of the two events occurring simultaneously is the product of P_1 and P_2." For example, the Probability of head coming up in a toss of a coin is $\frac{1}{2}$ and the Probability of 4 coming in a throw of a die is $\frac{1}{6}$, if a coin and a die are thrown together, the Probability of head coming up in the toss of the coin and 4 coming up in the throw of a die will be $\frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$.

Illustration

(a) What is the Probability of throwing two 'fours' in two throws of a die ?

The Probability of a 'four' in first throw = $\frac{1}{6}$.

The Probability of a four in second throw = $\frac{1}{6}$.

The Possibility of two 'fours' = $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ **Ans.**

(b) What is the Probability of getting all the heads in four throws of a coin ?

The change of getting head in the 1^{st} throw = $\frac{1}{2}$.

The change of getting head in the 2^{nd} throw = $\frac{1}{2}$.

The change of getting head in the 3^{rd} throw = $\frac{1}{2}$.

The change of getting head in the 4^{th} throw = $\frac{1}{2}$.

Thus the Probability of getting heads in all the throws :

$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$ **Ans.**

(c) A Problem in statistics is given to three students A, B, C whose changes of solving it are 1/2, 1/3, 1/4 respectively. What is the Probability that the Problem will be solved ?

Probability that student A will fail to solve the Problem = $1-\frac{1}{2}=\frac{1}{2}$.

Probability that student B will fail to solve the Problem = $1-\frac{1}{3}=\frac{2}{3}$.

Probability that student C will fail to solve the Problem = $1-\frac{1}{4}=\frac{3}{4}$.

Since the events are independent, the Probability that all the students A, B, C will fail to solve the Problem –

$$\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}=\frac{1}{4}.$$

$\therefore$ The Probability that the Problem will be solved = $1-\frac{1}{4}=\frac{3}{4}$.

This Problem can also be solved in the following way :

	Condition	**Probability**
(i)	A Solve, B Solve, C Solve	$=\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$
(ii)	A Solve, B Solve, C fails to solves	$=\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}=\frac{3}{24}$
(iii)	A solve, B fails to solve, C solve	$=\frac{1}{2}\times\frac{2}{3}\times\frac{1}{4}=\frac{2}{24}$
(iv)	A fails to solve, B solve, C solve	$=\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}=\frac{1}{24}$
(v)	A solve, B fails to solve, C fails to solve	$=\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}=\frac{6}{24}$
(vi)	A fails to solve, B solve, C fails to solve	$=\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}=\frac{3}{24}$
(vii)	A fails to solve, B fails to solve, C solve	$=\frac{1}{2}\times\frac{2}{3}\times\frac{1}{4}=\frac{2}{24}$

(viii) A, B, C fails to solve $= \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{6}{24}$

The Problem is solved in all the conditions, except that of (viii). If the Probabilities of (i) to (vii) are added, that will give the Probability of Problem being solved. The total comes to $\frac{18}{24}$ or $\frac{3}{4}$.

The multiplication theorem will hold good only if the events belong to the same set. In order to show the importance of this fact. Moroney in his book "Facts from Figures" given an interesting example. His observe, "Consider the case of a man who demands the simultaneous occurrence of money virtue of an unrelated nature in his young lady. Let's suppose that he insists on a Grecian nose. Plantinum-blonde hair, eyes of odd colours one blue, one brown, and finally a first class knowledge of statistics. What is the Probability that the first lady he needs in the street will put ideas of marriage into his head? It is difficult to apply multiplication theorem in this case, because events do not belong to the same set.

12.5 PERMUTATION AND COMBINATION IN THE THEORY OF PROBABILITY

Sometimes we are interested in the total number of different ways in which items can be arranged so that the order of components is important, yet no two arrangement are similar. Arrangement of this short are called **Permutations.** For example if seven alphabets – A, B, C, D, E, F, G are to be arranged by taking two letters at a time, under no circumstances may be arranged contain the same 2 letters (like AA, BB) the following permutations are possible –

AB	AC	AD	AE	AF	AG	
BA	BC	BD	BE	BF	BG	
CA	CB	CD	CE	CF	CG	
DA	DB	DC	DE	DF	DG	7
EA	EB	EC	ED	EF	EG	
FA	FB	FC	FD	FE	FG	
GA	GB	GC	GD	GE	GF	

Hence there are $7 \times 6 = 42$ permutations. Thus following formula can give the number of permutations.

Permutation = n (n – 1)

The permutation can be show in a tree-diagram also. For example, three chairs, x, y and z can be arranged n (n – 1) (n – 2) or 3 (3 – 1) (3 – 2) = 6 ways.

Illustration

(a) If a person is given one cup of tea of each of 5 brands and asked to rank these according to preference. How many possible ranking can there be?

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

$$= \frac{5!}{(5-5)!} = \frac{5\times 4\times 3\times 2\times 1}{1} = 120$$

It can also be calculated —

Prem – n (n – 1) (n – 2) (n – 3)

5 (5 – 1) (5 – 2) (5 – 3)

$= 5 \times 4 \times 3 \times 2 = 120$ **Ans.**

(b) What is the Probability of getting all the heads in four throw of a coins ?

The change of getting head in the first throw = $\frac{1}{2}$

The change of getting head in the 2nd throw = $\frac{1}{2}$

The change of getting head in the 3rd throw = $\frac{1}{2}$

The change of getting head in the 4th throw = $\frac{1}{2}$

Thus the Probability of getting heads in all the throw :

$$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{16}$$ **Ans.**

Note : [This example is multiplication theorem]

(c) In how many ways first, second and third prizes can be distributed to three of 10 competitors ?

Solution – n = 10 r = 3

$$^{10}P_3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = \frac{10\times 9\times 8\times 7}{7}$$

= 720 ways. **Ans.**

Illustration –

(a) In how many ways can be letters of the word 'BASKET' be arranged ?

Solution – There are 6 letters.

Hence Perm = 6! = $6 \times 5 \times 4 \times 3 \times 2 \times 1$

= 720 ways. **Ans.**

(b) In how many ways 12 student be allotted to three tutorial group of 2, 4 and 6 respectively ?

Solution – n = 12 , P = 2 , q = 4, r = 6

$$^{12}P_{2,4,6} = \frac{n!}{P!\ q!\ r!} = \frac{12!}{2!\ 4!\ 6!}$$

$$= \frac{12\times 11\times 10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{(2\times 1)\times(4\times 3\times 2\times 1)\times(6\times 5\times 4\times 3\times 2\times 1)}$$

= 13860 **Ans.**

12.6 PERMUTATIONS AND COMBINATIONS

The word permutation refers to arrangements and the word combination refers to 'groups'. These terms are used in the calculation of probability. Some simple rules of permutations and combinations are given below :

(i) The number of permutations of n dissimilar things taken all at a time is n! Thus, if there are 3 letters A, Band C, the total number of ways in which they can be arranged is ABC, ACB, BAC, BCA, CAB and CBA, i.e., 3! = 3 × 2 × 1 = 6.

Factorial n (written as n!) is equal to the continued product of n natural numbers starting from 1, i.e.,

$$n! = 1 \times 2 \times 3 \ldots\ldots\ldots (n-1)n$$

$$= n(n-1)\ (n-2) \ldots\ldots\ldots 3 \cdot 2 \cdot 1$$

(ii) The number of permutations of n dissimilar things taken r at a time is ${}^{n}P_{r} = \frac{n!}{(n-r)!}$. Thus, if we are to make arrangements of any two letters out of the letters A, B, C, then the different arrangements will be AB, BA, AC, CA, BC, CB, i.e., 6 arrangements which in factorial notation can be represented as

$${}^{3}P_{2} = \frac{3!}{(3-2)!} = 3i = 3 \times 2 \times 1 = 6$$

$\Rightarrow$ Three letters taken 2 at a time can be arranged in 6 ways.

The number of arrangement of any two letters out of $4 = {}^{4}P_{2} = \frac{4!}{2!} = 4 \times 3 = 12$.

(iii) The number of permutations of n things when n_1 of them are of one kind and n_2 of another kind is $\frac{n!}{n_1! \times n_2!}$. Thus, if we have to find out the permutations of the letters of the word FARIDABAD (where A occurs 3 times and D occurs 2 times) the answer would be $\frac{9!}{3! \times 2!}$.

$$= \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} = 30,240$$

(iv) The fundamental rule of counting is that if an operation can be performed in 'm' ways and having been performed in anyone of these ways, a second operation can be performed in' n' ways, the total number of ways of performing the two operations together is $m \times n$.

Thus, if a journey between Jodhpur and Jaipur can be performed by rail is 4 ways and return journey from Jaipur to Jodhpur can be performed in 4 ways, the total number of ways of performing the two journeys is $4 \times 4 = 16$. However, if the return journey is not to be performed by the same train by which one went to Jaipur, then the number of ways of performing the return journey is only 3 and the total number of ways of performing both the journeys would be $4 \times 3 = 12$.

(v) The number of combinations of n different things taken r at a time is ${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$. Thus, if we have to pick up two alphabets out of three A, B and C, we can pick up AB or AC or BC, i.e., 3 ways or

$${}^{3}C_{2} = \frac{3!}{2!(3-2)!} = \frac{3!}{2!} = \frac{3\times 2\times 1}{2} = 3.$$

We have seen that the number of permutations in this case was 6 because each combination can be arranged in two ways as AB, BA, AC, CA, BC and CB. Thus, the number of permutations is equal to number of combinations multiplied by r!. In other words,

$${}^{n}P_{r} = {}^{n}C_{r} \times r! \quad \text{or} \quad {}^{n}C_{r} = \frac{1}{r!} \times {}^{n}P_{r}$$

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{1}{r!}\,{}^{n}P_{r}$$

$$\Rightarrow \qquad {}^{n}P_{r} = r!\,{}^{n}C_{r}$$

Note : 0! = 1 and 1! = 1.

12.7 CONCLUSION

The solution of many problems involving Probabilities requires a through under studying of some of the basic rule which govern manipulation of Probabilities. The basic reason of the development of this theory is the presence in almost every aspect of life of random phenomenon.

12.8 FURTHER STUDY

1. Alhance D.N. and Alhance Beena, Fundamental of Statistics.

2. Monga, G.S., Elementary Statistics.

3. Gupta, S.B., Principles of Statistics.

Uttar Pradesh Rajarshi Tandon
Open University

BBA-121

Research Methodology

BLOCK

4

STATISTICAL TEST

UNIT-13

CONCEPTUAL FRAMEWORK

UNIT-14

ANOVA AND OTHERS

UNIT-15

Z-TEST AND T-TEST

UNIT-16

USES OF ICT IN RESEARCH METHODOLOGY

परिशिष्ट–4

आन्तरिक कवर–दो का प्ररूप

Format of the II Inner Covers

लेखक Dr. Piyali Ghosh, Asst. Professor, School of Management, MNNIT, Allahabad

सम्पादक Prof. H.K. Singh, Professor, Department of Commerce, BHU Varansi.

परिमापक

सहयोगी टीम

संयोजक Dr. Gaurav Sankalp, SoMS, UPRTOU, Allahabd.

प्रूफ रीडर

©UPRTOU, Prayagraj-2020

ISBN : 978-93-83328-54-3

NODE-110/NODE-124/NFT-063

BLOCK INTRODUCTION

Unit – 13 Conceptual Framework of Hypothesis, Tests of Goodness of FIT Chi-Square.

Unit – 14 ANOVA and Others.

Unit – 15 Z Test and T Test.

Unit – 16 Uses of ICT in Research Methodology.

UNIT-13 CONCEPTUAL FRAMEWORK

Objectives

After going through this unit you should be able to know about the–

1. Conceptual Framework of Hypothesis
2. Test of Goodness of fit
3. Chi-Square

Structure

13.1 INTRODUCTION OR CONCEPTUAL FRAMEWORK OF HYPOTHESIS

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

A different meaning of the term hypothesis is used in formal logic, to denote the antecedent of a proposition; thus in the proposition "If P, then Q". P denotes the hypothesis (or antecedent); Q can be called a consequent. P is the assumption in a (possibly counterfactual) what if question.

The adjective hypothetical, meaning "having the nature of a hypothesis", or "being assumed to exist as an immediate consequence of a hypothesis", can refer to any of these meanings of the term "hypothesis".

13.2 USES OF HYPOTHESIS

In its ancient usage, hypothesis referred to a summary of the plot of a classical drama the english word hypothesis comes from the ancient Greek ύπόθεσις (hupothesis), meaning "to put under" or "to suppose".

In Plato's Meno (86e-87b), Socrates dissects virtue with a method used by mathematicians, that of "investigating from a hypothesis". In this sense, 'hypothesis' refers to a clever idea or to a convenient mathematical approach that simplifies cumbersome calculations. Cardinal Bellarmine gave a famous example of this usage in the warning issued to Galileo in the early 17th century that he must not treat the motion of the earth as a reality, but merely as a hypothesis.

In common usage in the 21st century. a hypothesis refers to a provisional idea whose merit requires evaluation. For proper evaluation, the framer of a hypothesis needs to define specifics in operational terms. A hypothesis requires more work by the researcher in order to either confirm or disprove it in due course, a confirmed hypothesis may become part of a theory or occasionally may grew to become a theory itself. Normally, scientific hypotheses have the form of a mathematical model. Sometimes, but not always, one can also formulate them as existential statements, stating that some particular instance of the phenomenon under examination has some characteristic and causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic.

Any useful hypothesis will enable predictions by reasoning (including deductive reasoning). It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction may also invoke statistics and only talk about probabilities. Karl Popper, following others, has argued that a hypothesis must be falsifiable, and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. Other philosophers of science have rejected the criterion of falsifiability or supplemented it with other criteria, such as verifiability (e.g., verificationism) or coherence (e.g., confirmation holism). The scientific method involves experimentation, to test the ability of some hypothesis to adequately answer the question under investigation. In contrast, unfettered observation is not as likely to raise unexplained issues or open questions in

science, as would the formulation of a crucial experiment to test the hypothesis. A thought experiment might also be used to test the hypothesis as well.

In framing a hypothesis, the investigator must not currently know the outcome of a test or that it remains reasonably under continuing investigation. Only in such cases does the experiment, test or study potentially increase the probability of showing the truth of a hypothesis. If the researcher already knows the outcome, it counts as a "consequence" and the researcher should have already considered this while formulating the hypothesis. If one cannot assess the predictions by observation or by experience, the hypothesis needs to be tested by others providing observations. For example, a new technology or theory might make the necessary experiments feasible.

13.3. SCIENTIFIC HYPOTHESIS

People refer to a trial solution to a problem as a hypothesis, often called an "educated guess" because it provides a suggested solution based on the evidence. However, some scientists reject the term "educated guess" as incorrect. Experimenters may test and reject several hypotheses before solving the problem.

Working Hypothesis

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

The provisional nature of working hypotheses make them useful as an organizing device in applied research. Here they act like a useful guide to address problems that are still in a formative phase.

In recent years, philosophers of science have tried to integrate the various approaches to evaluating hypotheses, and the scientific method in general, to form a more complete system that integrates the individual concerns of each approach. Notably, Imre Lakatos and Paul Feyerabend, Karl Popper's colleague and student, respectively, have produced novel attempts at such a synthesis.

13.4 MEASUREMENT OF HYPOTHESIS

Concepts in Hempel's deductive homological model play key role in the development and testing of hypothesis. Most formal hypothesis connect concepts by specifying the expected relationships between propositions. When a set of hypothesis are grouped together they become

a type of conceptual framework. When a conceptual framework is complex and incorporates causality or explanation it is generally referred to as a theory. According to noted philosopher of science Carl Gustav Hempel " an adequate empirical interpretation turns a theoretical system into a testable theory. The hypothesis whose constituent terms have been interpreted become capable of test by reference to observable phenomena. Frequently the interpreted hypothesis will be derivative hypotheses of the theory; but their confirmation or disconfirmation by empirical data will then immediately strengthen or weaken also the primitive hypotheses from which they were derived."

Hempel provides a useful metaphor that describes the relationship between a conceptual framework and the framework as it is observed and perhaps tested (interpreted framework). "The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretative connections, the network can function as a scientific theory". Hypotheses with concepts anchored in the plane of observation are ready to be tested. In "actual scientific practice the process of framing a theoretical structure and of interpreting it are not always sharply separated, since the intended interpretation usually guides the construction of the theoretician". It is, however, "possible and indeed desirable, for the purposes of logical clarification, to separate the two steps conceptually."

13.5 STATISTICAL HYPOTHESIS TESTING

When a possible correlation or similar relation between phenomena is investigated, such as whether a proposed remedy is effective in treating a disease, the hypothesis that a relation exists cannot be examined the same way one might examine a proposed new law of nature. In such an investigation, if the tested remedy shows no effect in a few cases, these do not necessarily falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall effect would be observed if the hypothesized relation does not exist. If that likelihood is sufficiently small (e.g., less than 1%), the existence of a relation may be assumed. Otherwise, any observed effect may be due to pure chance.

In statistical hypothesis testing, two hypotheses are compared. These are called the null hypothesis and the alternative hypothesis. The null hypothesis is the hypothesis that states that there is no relation between the phenomena whose relation is under investigation, or at least not of the form given by the alternative hypothesis. The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis. it states that there is some kind of relation. The alternative hypothesis may take several forms, depending on the nature of the hypothesized relation; in particular it can be two-sided (for example: there is some effect, in a yet

unknown direction) or one-sided (the direction of the hypothesized relation positive or negative, is fixed in advance).

Conventional significance levels for testing hypotheses (acceptable probabilities of wrongly rejecting a true null hypothesis) are .10, .05, and .01. Whether the null hypothesis is rejected and the alternative hypothesis is accepted, must be determined in advance, before the observations are collected or inspected. If these criteria are determined later, when the data to be tested are already known, the test is invalid.

The above procedure is actually dependent on the number of the participants (units or sample size) that is included in the study. For instance, the sample size may be too small to reject a null hypothesis and, therefore, it is recommended to specify the sample size from the beginning. It is advisable to define a small, medium and large effect size for each of a lumber of important statistical tests which are used to test the hypotheses.

13.6 TEST OF GOODNESS OF FIT

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, ego to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov-Smirnov test), or whether outcome frequencies follow a specified distribution (see Pearson's chi-squared test). In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.

13.7. FIT OF DISTRIBUTIONS

In assessing whether a given distribution is suited to a data-set, the following tests and their underlying measures of fit can be used :

- Kolmogorov-Smirnov test;
- Cramér-von Mises criterion;
- Anderson-Darling test;
- Shapiro-Wilk test,
- Chi Square test;
- Akaike Information criterion;
- Hosmer-Lemeshow test;

13.8. REGRESSION ANALYSIS

In regression analysis, the following relate to goodness of fit:

- Coefficient of determination (The R squared measure of goodness of fit);
- Lack-of-fit sum of squares.

Example

One way in which a measure of goodness of fit statistic can be constructed, in the case where the variance of the measurement error is known, is to construct a weighted sum of squared errors.

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

where σ^2 is the known variance of the observation, O is the observed data and E is the theoretical data. This definition is only useful when one has estimates for the error on the measurements, but it leads to a situation where a chi-squared distribution can be used to test goodness of fit, provided that the errors can be assumed to have a normal distribution.

The reduced chi-squared statistic is simply the chi-squared divided by the number of degrees of freedom:

$$X^2_{\text{red}} = \frac{X^2}{\nu} = \frac{1}{\nu} \sum \frac{(O - E)^2}{\sigma^2}$$

where ν is the number of degrees of freedom, usually given by N–n–1, where N is the number of observations, and n is the number of fitted parameters, assuming that the mean value is an additional fitted parameter. The advantage of the reduced chi-squared is that it already normalizes for the number of data points and model complexity. This is also known as the mean square weighted deviation.

As a rule of thumb (again valid only when the variance of the measurement error is known a prior rather than estimated from the data), $a\ X^2_{red} > 1$ indicates a poor model fit $\text{A}\ X^2_{\text{red}} > 1$ indicates that the fit has not fully captured the data (or that the error variance has been underestimated). In principle, a value of $X^2_{\text{red}} = 1$ indicates that the extent of the match between observation and estimates is in accord with the error variance $\text{A}\ X^2_{\text{red}} < 1$ indicates that the model is 'over-fitting' the data

either the model is improperly fitting noise, or the error variance has been overestimated.

Categorical Data

The following are examples that arise in the context of categorical data.

13.9. PEARSON'S CHI-SQUARED TEST

Pearson's chi-squared test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where :

O_i = an observed frequency (i.e. count) for bin i

E_i = an expected (theoretical) frequency for bin i, asserted by the null hypothesis.

The expected frequency is calculated by :

$$E_i = (F(Y_u) - F(Y_l))N$$

where :

F = the cumulative distribution function for the distribution being tested.

Y_u = the upper limit for class *i*, and

Y_l = the lower limit for class i, and

N = the sample size

The resulting value can be compared to the chi-squared distribution to determine the goodness of fit. In order to determine the degrees of freedom of the chi-squared distribution, one takes the total number of observed frequencies and subtracts the number of estimated parameters. The test statistic follows, approximately, a chi-square distribution with (k – c) degrees of freedom where k is the number of non-empty cells and c is the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution.

Example : Equal Frequencies of men and women

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to

the theoretical frequencies of 50 men and 50 women. If there were 44 men in the sample and 56 women, then

$$X^2 = \frac{(44-50)^2}{50} + \frac{(56-50)^2}{50} = 1.44$$

If the null hypothesis is true (i.e., men and women are chosen with equal probability in the sample), the test statistic will be drawn from a chi-squared distribution with one degree of freedom, though one might expect two degrees of freedom (one each for the men and women), we must take into account that the total number of men and women is constrained (100), and thus there is only one degree of freedom (2 – 1). Alternatively, if the male count is known the female count is determined, and vice-versa.

Consultation of the chi-squared distribution for 1 degree of freedom shows that the probability of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.23. This probability is higher than conventional criteria for statistical significance (.001 – .05), so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women (i.e. we would consider our sample within the range of what we'd expect for a 50/50 male/female ratio.)

13.10 BINOMIAL CASE

A binomial experiment is a sequence of independent trials in which the trials can result in one of two outcomes, success or failure. There are n trials each with probability of success, denoted by p. Provided that $pp_i \gg 1$ for every i (where i = 1, 2, ... , k), then

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

This has approximately a chi-squared distribution with k – 1 df. The fact that df = k – 1 is a consequence of the restriction $\sum N_i = n$. We know there are k observed cell counts, however, once any k – 1 are known, the remaining one is uniquely determined. Basically, one can say, there are only k – 1 freely determined cell counts, thus df = k – 1.

Other Measures of fit

The likelihood ratio test statistic is a measure of the goodness of fit of a model, judged by whether an expanded form of the model provides a substantially improved fit.

13.11 CHI-SQUARE TEST

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as χ^2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. As a non-parametric* test, it "can be used to determine if categorical data shows dependency or the two classifications are independent. It can also be used to make comparisons between theoretical populations and actual data when categories are used." Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance.

Chi-Square as a test for comparing variance

The chi-square value is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (μ) and with a specified variance (σ_p^2). The test is based on χ^2 – distribution. Such a distribution we encounter 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 distributions that are related to χ^2 – distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a χ^2 – distribution. Thus, $\frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\sigma_s^2}{\sigma_P^2}$ (d.f.) would have the same distribution as χ^2 –distribution with (n –1) degrees of freedom.

The χ^2 –distribution is not symmetrical and all the values are positive. For making use of this distribution, one is required to know the degrees of freedom since for different degrees of freedom we have different curves. The smaller the number of degrees of freedom, the more skewed is the distribution which is illustrated in Fig. 10.1:

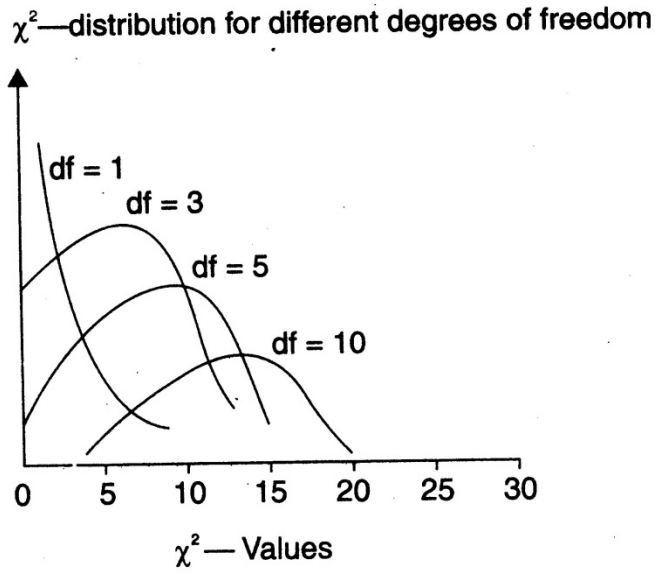


Fig. 10.1

Table given in the Appendix gives selected critical values of χ^2 for the different degrees of freedom. χ^2 -values are the quantities indicated on the x-axis of the above diagram and in the table are areas below that value.

In brief, when we have to use chi-square as a test of population variance, we have to work out the value of χ^2 to test the null hypothesis (viz., H_0 : $\sigma_s^2 = \sigma_p^2$) as under :

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

where

σ_s^2 = variance of the sample;

σ_p^2 = variance of the population;

(n–1) = degrees of freedom, n being the number of items in the sample.

Then by comparing the calculated value with the table value of χ^2 for (n – 1) degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis. If the calculated value of χ^2 is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected. All this can be made clear by an example.

Illustration 1

Weight of 10 students is as follows :

S.No.	1	2	3	4	5	6	7	8	9	10
Weight (kg.)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent and 1 per cent level of significance.

Solution

First of all we should work out the variance of the sample data or σ_s^2 and the same has been worked out as under:

S.No.	**X_i (Weight in kgs.)**	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	–9	81
2	40	–7	49
3	45	–2	04
4	53	+6	36
5	47	+0	00
6	43	–4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
n = 10	**$\Sigma X_i = 470$**		**$\sum (X_i - \bar{X})^2 = 280$**

$$\overline{X} = \frac{\sum X_i}{n} = \frac{470}{10} = 47 \text{ kgs.}$$

$$\therefore \qquad \sigma_s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}} = \sqrt{31.11}$$

or $$\sigma_s^2 = 31.11.$$

Let the null hypothesis be $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

$$= \frac{31.11}{20}(10-1) = 13.999$$

Degrees of freedom in the given case is (n – 1) = (10 – 1) = 9. At 5 per cent level of significance the table value of χ^2 = 16.92 and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both these values are greater than the calculated value of χ^2 which is 13.999. Hence we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 per cent as also at 1 per cent level of significance. In other words, the sample can be said to have been taken from a population with variance 20 kgs .

Illustration 2

A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

Solution.

n = 10

$$\sum (X_i - \overline{X})^2 = 50$$

$$\sigma_s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{50}{9}$$

Take the null hypothesis as $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis, we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\frac{50}{9}}{5}(10-1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = (10 – 1) = 9.

The table value of χ^2 at 5 per cent level for 9 d.f. is 16.92. The calculated value of χ^2 is less than this table value, so we accept the null hypothesis and conclude that the valiance of the population is 5 as given in the question.

13.12 CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test. As a non-parametric test, chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence.

As a test of goodness of fit, χ^2 test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data. When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of χ^2 is greater than its table value, the fit is not considered to be a good one.

As a test of independence, χ^2 test enables us to explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, χ^2 test will helps us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever. On this basis we first calculate the expected frequencies and then work out the value of χ^2. If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in controlling the fever). But if the calculated value of χ^2 is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e., the new medicine is effective in controlling the fever

and as such may be prescribed). It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

In order that we may apply the chi-square test either as a test of goodness of fit or as a test to judge the significance of association between attributes, it is necessary that the observed as well as theoretical or expected frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency as we find in case of observed distribution. χ^2 is then calculated as follows :

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where,

O_{ij} = observed frequency of the cell in ith row and jth column.

Eij = expected frequency of the cell in ith row and jth column.

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; but generally due to sampling errors, χ^2 is not equal to zero and as such we must know the sampling distribution of χ^2 so that we may find the probability of an observed χ^2 being given by a random sample from the hypothetical universe. Instead of working out the probabilities, we can use ready table which gives probabilities for given values of χ^2. Whether or not a calculated value of χ^2 is significant can be ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of significance. If the calculated value of χ^2 is equal to or exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference is considered as insignificant i.e., considered to have arisen as a result of chance and as such can be ignored.

As already stated, degrees of freedom play an important part in using the chi -square distribution and the test based on it, one must correctly determine the degrees of freedom. If there are 10 frequency classes and there is one independent constraint, then there are (10 – 1) = 9 degrees of freedom. Thus, if 'n' is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, the d.f. would be equal to (n – 1). In the case of a contingency table (i.e., a table with 2 columns and 2 rows or a table with two columns and more than two rows or a table with two rows but more than two columns or a table with more than two rows and more than two columns), the d.f. is worked out as follows:

$$\text{d.f.} = (c - 1)(r - 1)$$

where 'c' means the number of columns and 'r' means the number of rows.

The following conditions should be satisfied before χ^2 test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the items in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known are know as linear constraints.

13.13 CONCLUSION

A hypothesis is a proposed explanation for a phenomenon. Hypothesis comes from the Greek, Actually, hypothesis refers to a clever idea. But in recent century, it refers to a provisional idea whose merit requires evaluation.

13.14 FURTHER STUDY

1. Monga, G.S., Elementary Statistics.

2. Alhance, D.N., Alhance, Beena, Principles of Statistics.

3. Kirkman, T.W., Chi-Squared Curve Fitting.

4. Chartle Land and Tonya L. Kuhi, Chi-Squared Data Fitting.

UNIT-14 ANOVA AND OTHERS

Objectives

After going through this unit you should be able to know about the–

1. Conceptual Framework of ANOVA
2. Assumptions of ANOVA
3. Characteristics of ANOVA
4. Analysis of ANOVA

Structure

14.1 CONCEPTUAL FRAMEWORK OF ANOVA

Analysis of variance (ANOVA) is a collection of statistical models used in order to analyze the differences between group means and their associated procedures (such as "variation" among and between groups), developed by R.A. Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. As doing multiple two-sample t-tests would result in an increased chance of committing a

statistical type I error, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

14.2 ANALYSIS OF ANOVA

The analysis of variance can be used as an exploratory tool to explain observations A dog show provides an example. A dog show is not a random sampling of the breed: it is typically limited to dogs that are male, adult, pure-bred, and exemplary. A histogram of dog weights from a show might plausibly be rather complex, like the yellow-orange distribution shown in the illustrations. Suppose we wanted to predict the weight of a dog based on a certain set of characteristics of each dog. Before we could do that, we would need to explain the distribution of weights by dividing the dog population into groups based on those characteristics. A successful grouping will split dogs such that (a) each group has a low variance of dog weights (meaning the group is relatively homogeneous) and (b) the mean of each group is distinct (if two groups have the same mean, then it isn't reasonable to conclude that the groups are, in fact, separate in any meaningful way).

Groups are divided into two group ie. X_1 and X_2. In First group we divide the dogs according to the product (interaction) of two binary groupings: young vs old and short-haired vs long-haired (thus, group 1 is young, short-haired dogs, group 2 is young, long-haired dogs, etc.). Since the distributions of dog weight within each of the groups has a large variance, and since the means are very close across groups, grouping dogs by these characteristics does not produce an effective way to explain the variation in dog weights: knowing which group a dog is in does not allow us to make any reasonable statements as to what that dog's weight is likely to be. Thus, this grouping fails to fit the distribution we are trying to explain.

All attempt to explain the weight distribution by grouping dogs as (pet vs working breed) and (less athletic vs more athletic) would probably be somewhat more successful (fair fit). The heaviest dogs are likely to be big strong working breeds, while breeds kept as pets tend to be smaller and thus lighter. The distributions have variances that are considerably smaller than in the first case, and the means are more reasonably distinguishable. However, the significant overlap of distributions, for example, means that we cannot reliably say that X_1 and X_2 are truly distinct (i.e., it is perhaps reasonably likely that splitting dogs according to the flip of a coin-by pure chance-might produce distributions that look similar).

An attempt to explain weight by breed is likely to produce a very good fit. All Chihuahuas are light and all St Bernards are heavy. The difference in weights between two groups does not justify separate breeds. The analysis of variance provides the formal tools to justify these intuitive judgments. A common use of the method is the analysis of experimental

data or the development of models. The method has some advantages over correlation: not all of the data must be numeric and one result of the method is a judgment in the confidence in an explanatory relationship.

14.3. BACKGROUND AND TERMINOLOGY OF ANOVA

ANOVA is a particular form of statistical hypothesis testing heavily used in the analysis of experimental data. A statistical hypothesis test is a method of making decisions using data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, assuming the truth of the null hypothesis. A statistically significant result, when a probability (p-value) is less than a threshold (significance level), justifies the rejection of the null hypothesis, but only if the a priori probability of the null hypothesis is not high.

In the typical application of ANOVA, the null hypothesis is that all groups are simply random samples of the same population. For example, when studying the effect of different treatments on similar samples of patients, the null hypothesis would be that all treatments have the same effect (perhaps none). Rejecting the null hypothesis would imply that different treatments result in altered effects.

By construction, hypothesis testing limits the rate of Type I errors (false positive leading to false scientific claims) to a significance level. Experimenters also wish to limit Type II errors (false negatives resulting in missed scientific discoveries). The Type II error rate is a function of several things including sample size (positively correlated with experiment cost), significance level (when the standard of proof is high, the chances of overlooking a discovery are also high) and effect size (when the effect is obvious to the casual observer, Type II error rates are low).

The terminology of ANOVA is largely from the statistical design of experiments. The experimenter adjusts factors and measures responses in an attempt to determine an effect. Factors are assigned to experimental units by a combination of randomization and blocking to ensure the validity of the results. Blinding keeps the weighing impartial. Responses show a variability that is partially the result of the effect and is partially random error.

ANOVA is the synthesis of several ideas and it is used for multiple purposes. As a consequence, it is difficult to define concisely or precisely.

"Classical ANOVA for balanced data does three things at once :

1. As exploratory data analysis, an ANOVA is an organization of an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).

2. Comparisons of mean squares, along with F-tests ... allow testing of a nested sequence of models.

3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors."[1]

   In short, ANOVA is a statistical tool used in several ways to develop and confirm an explanation for the observed data.

Additionally :

4. It is computationally elegant and relatively robust against violations of its assumptions.

5. ANOVA provides industrial strength (multiple sample comparison) statistical analysis.

6. It has been adapted to the analysis of a variety of experimental designs.

As a result: ANOVA "has long enjoyed the status of being the most used (some would say abused) statistical technique in psychological research." ANOVA "is probably the most useful technique in the field of statistical inference."

ANOVA is difficult to teach, particularly for complex experiments, with split-plot designs being notorious. In some cases the proper application of the method is best determined by problem pattern recognition followed by the consultation of a classic authoritative test.

14.4 DESIGN OF EXPERIMENTS TERMS

(Condensed from the NIST Engineering Statistics handbook. Section 5.7. A Glossary of DOE Terminology.)

Balanced design

An experimental design where all cells (i.e. treatment combinations) have the same number of observations.

Blocking

A schedule for conducting treatment combinations in an experimental study such that any effects on the experimental results due to a known change in raw materials. operators, machines. etc., become concentrated in the levels of the blocking variable. The reason for blocking is to isolate a systematic effect and prevent it from obscuring the main effects. Blocking is achieved by restricting randomization.

Design

A set of experimental runs which allows the fit of a particular model and the estimate of effects.

DOE

Design of experiments. An approach to problem solving involving collection of data that will support valid, defensible, and supportable conclusions.

Effect

How changing the settings of a factor changes the response. The effect of a single factor is also called a main effect.

Error

Unexplained variation in a collection of observations. DOE's typically require understanding of both random error and lack of fit error.

Experimental unit

The entity to which a specific treatment combination is applied.

Factors

Process inputs an investigator manipulates to cause a change in the output.

Lack-of-fit error

Error that occurs when the analysis omits one or more important terms or factors from the process model. Including replication in a DOE allows separation of experimental error into its components: lack of fit and random (pure) error.

Model

Mathematical relationship which relates changes in a given response to changes in one or more factors.

Random error

Error that occurs due to natural variation in the process. Random error is typically assumed to be normally distributed with zero mean and a constant variance. Random error is also called experimental error.

Randomization

A schedule for allocating treatment material and for conducting treatment combinations in a DOE such that the conditions in one run neither depend on the conditions of the previous run nor predict the conditions in the subsequent runs.

Replication

Performing the same treatment combination more than once. Including repli-cation allows an estimate of the random error independent of any lack of fit error.

Responses

The output(s) of a process. Sometimes called dependent variable(s).

Treatment

A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.

14.5 ANALYSIS OF VARIANCE

There are three classes of models used in the analysis of variance, and these are outlined here.

Fixed-Effects Models

The fixed-effects model of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

Random-Effects Models

Random effects models are used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.

Mixed-Effects Models

A mixed-effects model contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

Example: Teaching experiments could be performed by a university department to find a good introductory textbook, with each text considered a treatment. The fixed-effects model would compare a list of candidate texts. The random-effects model would determine whether important differences exist among a list of randomly selected texts. The mixed-effects model would compare the (fixed) incumbent texts to randomly selected alternatives.

Defining fixed and random effects has proven elusive, with competing definitions arguably leading toward a linguistic quagmire.

14.6 Assumptions of ANOVA

The analysis of variance has been studied from several approaches, the most common of which uses a linear model that relates the response to the treatments and blocks. Note that the model is linear in parameters but

may be nonlinear across factor levels. Interpretation is easy when data is balanced across factors but much deeper understanding is needed for unbalanced data.

Textbook analysis using a normal distribution

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:

- Independence of observations — this is an assumption of the model that simplifies the statistical analysis.
- Normality – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called homoscedasticity – the variance of data in groups should be the same.

The separate assumptions of the textbook model imply that the errors are independently, identically, and normally distributed for fixed effects models, that is, that the errors (ε's) are independent and

$$\varepsilon \sim N\left(0, \sigma^2\right).$$

Randomization-Based Analysis

In a randomized controlled experiment, the treatments are randomly assigned to experimental units, following the experimental protocol. This randomization is objective and declared before the experiment is carried out. The objective random-assignment is used to test the significance of the null hypothesis, following the ideas of C.S. Peirce and Ronald A Fisher. This design-based analysis was discussed and developed by Francis J. Anscombe at Rothamsted Experimental Station and by Oscar Kempthorne at Iowa State University. Kempthorne and his students make an assumption of unit treatment additivity, which is discussed in the books of Kempthorne and David R Cox.

Unit-Treatment Additivity

In its simplest form; the assumption of unit-treatment additivity states that the observed response $y_{i,j}$ from experimental unit i when receiving treatment j can be written as the sum of the unit's response y_i and the treatment-effect $t_{j,}$ that is

$$y_{i,j} = y_i + t_j,$$

The assumption of unit-treatment additivity implies that, for every treatment j, the j^{th} treatment have exactly the same effect t_j on every experiment unit.

The assumption of unit treatment additivity usually cannot be directly falsified, according to Cox and Kempthorne. However, many

consequences of treatment-unit additivity can be falsified. For a randomized experiment, the assumption of unit-treatment additivity implies that the variance is constant for all treatments. Therefore, by contraposition, a necessary condition for Unit-treatment additivity is that the Variance is constant.

The use of unit treatment additivity and randomization is similar to the design-based inference that is standard in finite-population survey sampling.

Derived Linear Model

Kempthorne uses the randomization-distribution and the assumption of unit treatment additivity to produce a derived linear model, very similar to the textbook model discussed previously. The test statistics of this derived linear model are closely approximated by the test statistics of an appropriate normal linear model, according to approximation theorems and simulation studies. However, there are differences. For example, the randomization-based analysis results in a small but (strictly) negative correlation between the observations. In the randomization-based analysis there is no assumption of a normal distribution and certainly no assumption of independence. On the contrary, the observations are dependent.

The randomization-based analysis has the disadvantage that its exposition involves tedious algebra and extensive time. Since the randomization-based analysis is complicated and is closely approximated by the approach using a normal linear model, most teachers emphasize the normal linear model approach. Few statisticians object to model-based analysis of balanced randomized experiments.

Statistical Models for Observational Data

However, when applied to data from non-randomized experiments or observational studies, model-based analysis lacks the warrant of randomization. For observational data, the derivation of confidence intervals must use subjective models, as emphasized by Ronald A Fisher and his followers. In practice, the estimates of treatment-effects from observational studies generally are often inconsistent. In practice, "statistical models" and observational data are useful for suggesting hypotheses that should be treated very cautiously by the public.

Summary of Assumptions

The normal-model based ANOVA analysis assumes the independence, normality and homogeneity of the variances of the residuals. The randomization-based analysis assumes only the homogeneity of the variances of the residuals (as a consequence of unit-treatment additivity) and uses the randomization procedure of the experiment. Both these analyses require homoscedasticity, as an assumption for the normal-model

analysis and as a consequence of randomization and additivity for the randomization-based analysis.

However, studies of processes that change variances rather than means (called dispersion effects) have been successfully conducted using ANOVA. There are no necessary assumptions for ANOVA in its full generality, but the F-test used for ANOVA hypothesis testing has assumptions and practical limitations which are of continuing Interest.

Problems which do not satisfy the assumptions of ANOVA can often be transformed to satisfy the assumptions. The property of unit-treatment additivity is not invariant under a "change of scale", so statisticians often use transformations to achieve unit-treatment additivity. If the response variable is expected to follow a parametric family of probability distributions, then the statistician may specify (in the protocol for the experiment or observational study) that the responses be transformed to stabilize the variance. Also, a statistician may specify that logarithmic transforms be applied to the responses, which are believed to follow a multiplicative model. According to Cauchy's functional equation theorem the logarithm is the only continuous transformation that transforms real multiplication to addition.

14.7 CHARACTERISTICS OF ANOVA

ANOVA is used in the analysis of comparative experiments, those in which only the difference in outcomes is of interest. The statistical significance of the experiment is determined by a ratio of two variances. This ratio is independent of several possible alterations to the experimental observations. Adding a constant to all observations does not alter significance. Multiplying all observations by a constant does not alter significance. So ANOVA statistical significance results are independent of constant bias and scaling errors as well as the units used in expressing observations. In the era of mechanical calculation it was common to subtract a constant from all observations (when equivalent to dropping leading digits) to simplify data entry. This is an example of data coding.

Logic of ANOVA

The calculations of ANOVA can be characterized as computing a number of means and variances, dividing two variances and comparing the ratio to a handbook value to determine statistical Significance. Calculating a treatment effect is then trivial, "the effect of any treatment is estimated by taking the difference between the mean of the observations which receive the treatment and the general mean."

Partitioning of the Sum of Squares

ANOVA uses traditional standardized terminology. The definitional equation of sample variance is $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$ where the divisor is

called the degrees of freedom (DF), the summation is called the sum of squares (SS), the result is called the mean square (MS) and the squared terms are deviations from the sample mean. ANOVA estimates 3 sample variances : a total variance based on all the observation deviations from the grand mean, an error variance based on all the observation deviations from their appropriate treatment means and a treatment variance. The treatment variance is based on the deviations of treatment means from the grand mean, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model. For example, the model for a simplified ANOVA with one type of treatment at different levels.

$$SS_{Total} = SS_{Error} + SS_{Treatments}$$

The number of degrees of freedom DF can be partitioned in a similar way: one of these components (that for error) specifies a chi-squared distribution which describes the associated sum of squares, while the same is true for "treatments" if there is no treatment effect.

$$DF_{Total} = DF_{Error} + DF_{Treatments}$$

See also Lack-of-fit sum of squares.

14.8 THE F-TEST

The F-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F-test statistic

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{M\,S_{Treatments}}{M\,S_{Error}} = \frac{S\,S_{Treatments} / (I - 1)}{S\,S_{Error} / (n_T - I)}$$

where MS is mean square, I = number of treatments and n_T = total number of cases to the F-distribution with $I - 1$, $n_T - I$ degrees of freedom. Using the F-distribution is a natural candidate because the test statistic is the ratio of two scaled sums of squares each of which follows a scaled chi-squared distribution.

The expected value of F is $1+n\sigma^2_{\text{Treatment}}/\sigma^2_{\text{Error}}$, (where n is the treatment sample size) which is 1 for no treatment effect. As values of F increase above 1, the evidence is increasingly inconsistent with the null hypothesis. Two apparent experimental methods of increasing F are increasing the sample size and reducing the error variance by tight experimental controls.

There are two methods of concluding the ANOVA hypothesis test, both of which produce the same result:

- The textbook method is to compare the observed value of F with the critical value of F determined from tables. The critical value of F is a function of the degrees of freedom of the numerator and the denominator and the significance level (α). If $F \geq F_{\text{Critical}}$, the null hypothesis is rejected.
- The computer method calculates the probability (p-value) of a value of F greater than or equal to the observed value. The null hypothesis is rejected if this probability is less than or equal to the significance level (α).

The ANOVA F-test is known to be nearly optimal in the sense of minimizing false negative errors for a fixed rate of false positive errors (i e. maximizing power for a fixed significance level). For example, to test the hypothesis that various medical treatments have exactly the same effect, the F-test's p-values closely approximate the permutation test's p-values: The approximation is particularly close when the design is balanced. Such permutation tests characterize tests with maximum power against all alternative hypotheses, as observed by Rosenbaum. The ANOVA F-test (of the null-hypothesis that all treatments have exactly the same effect) is recommended as a practical test, because of its robustness against many alternative distributions.

Extended Logic

ANOVA consists of separable parts; partitioning sources of variance and hypothesis testing can be used individually. ANOVA is used to support other statistical tools. Regression is first used to fit more complex models to data, then ANOVA is used to compare models with the objective of selecting simple(r) models that adequately describe the data. "Such models could be fit Without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients. We think of the analysis of variance as a way of understanding and structuring multi-level models—not as an alternative to regression but as a too for summarizing complex high-dimensional Inferences...."

ANOVA for a Single Factor

The Simplest experiment suitable for ANOVA analysis is the completely randomized experiment with a single factor. More complex experiments

with a single factor involve constraints on randomization and include completely randomized blocks and Latin squares (and variants: Graeco-Latin squares, etc.) The more complex experiments share many of the complexities of multiple factors A relatively complete discussion of the analysis (models, data summaries, ANOVA table) of the completely randomized experiment is available.

ANOVA for Multiple Factors

ANOVA generalizes to the study of the effects of multiple factors. When the experiment includes observations at all combinations of levels of each factor, it is termed factorial. Factorial experiments are more efficient than a series of single factor experiments and the efficiency grows as the number of factors increases. Consequently, factorial designs are heavily used.

The use of ANOVA to study the effects of multiple factors has a complication. in a 3-way ANOVA with factors x, y and z, the ANOVA model includes terms for the main effects (x, y, z) and terms for interactions (xy, xz, yz, xyz). All terms require hypothesis tests. The proliferation of interaction terms increases the risk that some hypothesis test will produce a false positive by chance. Fortunately, experience says that high order interactions are rare. The ability to detect interactions is a major advantage of multiple factor ANOVA. Testing one factor at a time hides Interactions, but produces apparently inconsistent experimental results.

Caution is advised when encountering interactions; Test interaction terms first and expand the analysis beyond ANOVA if interactions are found. Texts vary in their recommendations regarding the continuation of the ANOVA procedure after encountering an interaction. Interactions complicate the interpretation of experimental data. Neither the calculations of significance nor the estimated treatment effects can be taken at face value. "A significant Interaction will often mask the significance of main effects. Graphical methods are recommended to enhance understanding. Regression is often useful. A lengthy discussion of interactions is available in Cox (1958). Some interactions can be removed (by transformations) while others cannot.

A variety of techniques are used with multiple factor ANOVA to reduce expense. One technique used in factorial designs is to minimize replication (possibly no replication with support of analytical trickery) and to combine groups when effects are found to be statistically (or practically) insignificant. An experiment with many insignificant factors may collapse into one with a few factors supported by many replications.

Worked Numeric Examples

Several fully worked numerical examples are available. A simple case uses one-way (a single factor) analysis. A more complex case uses two-way (two-factor) analysis.

Associated Analysis

Some analysis is required in support of the design of the experiment while other analysis is performed after changes in the factors are formally found to produce statistically significant changes in the responses. Because experimentation is iterative, the results of one experiment alter plans for following experiments.

Preparatory Analysis

In the design of an experiment, the number of experimental units is planned to satisfy the goals of the experiment. Experimentation is often sequential.

Early experiments are often designed to provide mean-unbiased estimates of treatment effects and of experimental error. Later experiments are often designed to test a hypothesis that a treatment effect has an important magnitude; in this case, the number of experimental units is chosen so that the experiment is within budget and has adequate power, among other goals.

Reporting sample size analysis is generally required in psychology. "Provide information on sample size and the process that lead to sample size decisions." The analysis, which is written in the experimental protocol before the experiment is conducted, is examined in grant applications and administrative review boards.

Besides the power analysis, there are less formal methods for selecting the number of experimental units. These include graphical methods based on limiting the probability of false negative errors, graphical methods based on an expected variation increase (above the residuals) and methods based on achieving a desired confident interval.

Power Analysis

Power analysis is often applied in the context of ANOVA in order to assess the probability of successfully rejecting the null hypothesis if we assume a certain ANOVA design, effect size in the population, sample size and significance level. Power analysis can assist in study design by determining what sample size would be required in order to have a reasonable chance of rejecting the null hypothesis when the alternative hypothesis is true.

Effect Size

Several standardized measures of effect have been proposed for ANOVA to summarize the strength of the association between a predictor(s) and the dependent variable (e.g., η^2, ω^2, or f^2) or the overall standardized difference (ψ) of the complete model. Standardized effect-size estimates

facilitate comparison of findings across studies and disciplines. However, while standardized effect sizes are commonly used in much of the professional literature, a non-standardized measure of effect size that has immediately "meaningful" units may be preferable for reporting purposes.

Follow-up Analysis

It is always appropriate to carefully consider outliers. They have a disproportionate impact on statistical conclusions and are often the result of errors.

Model Confirmation

It is prudent to verify that the assumptions of ANOVA have been met. Residuals are examined or analyzed to confirm homoscedasticity and gross normality. Residuals should have the appearance of (zero mean normal distribution) noise when plotted as a function of anything including time and modeled data values. Trends hint at interactions among factors or among observations. One rule of thumb: "If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations and our results will still be approximately correct."

Follow-up Tests

A statistically significant effect in ANOVA is often followed up with one or more different follow-up tests. This can be done in order to assess which groups are different from which other groups or to test various other focused hypotheses. Follow-up tests are often distinguished in terms of whether they are planned (a prior) or post hoc. Planned tests are determined before looking at the data and post hoc tests are performed after looking at the data.

Often one of the "treatments" is none, so the treatment group can act as a control. Dunnett's test (a modification of the t-test) tests whether each of the other treatment groups has the same mean as the control.

Post hoc tests such as Tukey's range test most commonly compare every group mean with every other group mean and typically incorporate some method of controlling for Type I errors. Comparisons, which are most commonly planned, can be either simple or compound. Simple comparisons compare one group mean with one other group mean. Compound comparisons typically compare two sets of groups means where one set has two or more groups (e.g., compare average group means of group A, B and C with group D). Comparisons can also look at tests of trend, such as linear and quadratic relationships, when the independent variable involves ordered levels.

Following ANOVA with pair-wise multiple-comparison tests has been criticized on several grounds. There are many such tests (10 in one table) and recommendations regarding their use are vague or conflicting.

14.9 STUDY DESIGNS AND ANOVAS

There are several types of ANOVA. Many statisticians base ANOVA on the design of the experiment, especially on the protocol that specifies the random assignment of treatments to subjects; the protocol's description of the assignment mechanism should include a specification of the structure of the treatments and of any blocking. It is also common to apply ANOVA to observational data using an appropriate statistical model.

Some popular designs use the following types of ANOVA:

- One-way ANOVA is used to test for differences among two or more independent groups (means) e.g. different levels of urea application in a crop. Typically, however, the one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$.
- Factorial ANOVA is used when the experimenter wants to study the interaction effects among the treatments.
- Repeated measures ANOVA is used when the same subjects are used for each treatment (e.g., in a longitudinal study).
- Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

14.10 ANOVA CAUTIONS

Balanced experiments (those with an equal sample size for each treatment) are relatively easy to interpret; Un balanced experiments offer more complexity. For single factor (one way) ANOVA, the adjustment for unbalanced data is easy, but the unbalanced analysis lacks both robustness and power. For more complex designs the lack of balance leads to further complications. "The orthogonality property of main effects and interactions present in balanced data does not carry over to the unbalanced case. This means that the usual analysis of variance techniques do not apply. Consequently, the analysis of unbalanced factorials is much more difficult than that for balanced designs." In the general case, "the analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and F-ratios will depend on the order in which the sources of variation are considered. The simplest techniques for handling unbalanced data restore balance by either throwing out data or by synthesizing missing data. More complex techniques use regression

Sir Ronald A Fisher introduced the term "variance" and proposed a formal analysis of variance in a 1918 article. The Correlation between relatives on the supposition of mendelian inheritance. His first application of the analysis of variance was published in 1921. Analysis of variance became widely known after being included in Fisher's 1925 book Statistical methods for Research workers.

Randomization models were developed by several researchers. The first was published in Polish by Neyman in 1923.

One of the attributes of ANOVA which ensured its early popularity was computational elegance. The structure of the additive model allows solution for the additive coefficients by simple algebra rather than by matrix calculations. In the era of mechanical calculators this simplicity was critical. The determination of statistical significance also required access to tables of the F function which were supplied by early statistics texts.

14.11 CONCLUSION

ANOVA provides a statistical tests of wheather or not the means of several groups are equal therefore, generalizes the T-Test to more than two groups. ANOVA is particular from the statistical hypothesis testing heavily used in the analysis of experimental data.

In short, ANOVA is statistical tool used in several ways to develop and confirm an explanation for a observed data.

14.12 FURTHER STUDY

1. Hilborn, Ray; Mangel, Marc (1997). *The ecological detective : confronting models with data.* Princeton University Press. p. 24. ISBN 978-0-03497-3.

2. Wilbur, R. Knorr, "Construction as existence proof in ancient geometry", p. 125, as selected by Jean Christianidis (ed.), *Classics in the history of Greek Mathematics*, Kluwer.

3. Gregory, Viastos, Myles Burnyeat (1994) *Socratic Studies*, Cambridge ISBN.

UNIT-15 Z-TEST AND T-TEST

Objectives

After going through this unit you should be able to know about the–

1. Concept and use of Z-Test.
2. Concept of T-Test and its history and uses.
3. Multivariate test

Structure

15.1 INTRODUCTION TO Z-TEST

A *Z*-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level, the *Z*-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's *t*-test which has separate critical values for each sample size. Therefore. many statistical tests can be conveniently performed as approximate *Z*-tests if the sample size is large or the population variance known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large ($n < 30$), the Student's *t*-test may be more appropriate.

If T is a statistic that is approximately normally distributed under the null hypothesis, the next step in performing a Z-test is to estimate the expected value θ of T under the null hypothesis, and then obtain an estimate s of the standard deviation of T. After that the standard score $Z = (T - \theta) / s$ is calculated, from which one-tailed and two-tailed p-values can be calculated as $\phi(-Z)$ (for upper-tailed tests), $\phi(-Z)$ (for lower-tailed tests) and $2\phi(-|Z|)$ (for two-tailed tests) where ϕ is the standard normal cumulative distribution function.

15.2 USE IN LOCATION TESTING

The term "Z-test" is often used to refer specifically to the one-sample location test comparing the mean of a set of measurements to a given constant. If the observed data X_1 ....., X_n are (i) uncorrelated, (ii) have a common mean μ, and (iii) have a common variance σ^2, then the sample average $\overline{X}$ has mean μ and variance σ^2 / n. If our null hypothesis is that the mean value of the population is a given number μ_0, we can use $\overline{X} - \mu_0$ as a test-statistic rejecting the null hypothesis if $\overline{X} - \mu_0$ is large.

To calculate the standardized statistic $Z = (\overline{X} - \mu_0) / s$, we need to either know or have an approximate value for σ^2, from which we can calculate $s^2 = \sigma^2 /n$. In some applications, σ^2 is known, but this is uncommon. If the sample size is moderate or large, we can substitute the sample variance for σ^2, giving a plug-in test. The resulting test will not be an exact Z-test since the uncertainty in the sample variance is not accounted for — however, It will be a good approximation unless the sample size is small. A t-test can be used to account for the uncertainty in the sample variance when the sample size is small and the data are exactly normal. There is no universal constant at which the sample size is generally considered large enough to justify use of the plug-in test. Typical rules of thumb range from 20 to 50 samples For larger sample sizes, the t-test procedure gives almost identical p-values as the Z-test procedure.

Other location tests that can be performed as Z-test are the two-sample location test and the paired difference test.

15.3 CONDITIONS FOR APPLICABILITY OF Z-TEST

For the Z-test to be applicable, certain conditions must be met.

- Nuisance parameters should be known, or estimated with high accuracy (an example of a nuisance parameter would be the standard deviation in a one-sample location test). Z-test focus on a Single parameter, and treat all other unknown parameters as being fixed at their true values. In practice, due to Slutsky's theorem,

"plugging in" consistent estimates of nuisance parameters can be justified. However if the sample size is not large enough for these estimates to be reasonably accurate, the Z-test may not perform well.

- The test statistic should follow a normal distribution. Generally, one appeals to the central limit theorem to justify assuming that a test statistic varies normally. There is a great deal of statistical research on the question of when a test statistic varies approximately normally. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

If estimates of nuisance parameters are plugged in as discussed above, it is important to use estimates appropriate for the way the data were sampled. In the special case of Z-test for the one or two sample location problem, the usual sample standard deviation is only appropriate if the data were collected as an independent sample.

In some Situations, It is possible to devise a test that properly accounts for the variation in plug-in estimates of nuisance parameters. !n the case of one and two sample location problems, a t-test does this.

Example

Suppose that in a particular geographic region, the mean and standard deviation of scores on a reading test are 100 points, and 12 points, respectively. Our Interest is in the scores of 55 students in a particular school who received a mean score of 96. We can ask whether this mean score is significantly lower than the regional mean — that is are the students in this school comparable to a simple random sample of 55 5 students from the region as a whole, or are their scores surprisingly low?

We begin by calculating the standard error of the mean :

$$\mathrm{SE} = \frac{\sigma}{\sqrt{\mathrm{n}}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

where σ is the population standard deviation.

Next we calculate the z-score, which is the distance from the sample mean to the population mean in units of the standard error:

$$\mathrm{z} = \frac{\mathrm{M} - \mu}{\mathrm{SE}} = \frac{96 - 100}{1.62} = -2.47$$

In this example, we treat the population mean and variance as known, which would be appropriate if all students in the region were tested. When population parameters are unknown, a t test should be conducted instead.

The classroom mean score is 96, which is – 2.47 standard error units from

the population mean of 100. Looking up the z-score in a table of the standard normal distribution, we find that the probability of observing a standard normal value below – 2.47 is approximately 0.5 – 0.4932 = 0.0068. This is the one-sided p-value for the null hypothesis that the 55 students are comparable to a simple random sample from the population of all test-takers. The two-sided p-value is approximately 0.014 (twice the one-sided p-value).

Another way of stating things is that with probability 1 – 0.014 = 0.986, a simple random sample of 55 students would have a mean test score within 4 units of the population mean. We could also say that with 98.6% confidence we reject the null hypothesis that the 55 test takers are comparable to a simple random sample from the population of test-takers.

The Z-test tells us that the 55 students of interest have an unusually low mean test score compared to most simple random samples of similar size from
the population of test-takers. A deficiency of this analysis is that it does not consider whether the effect size of 4 points is meaningful. If instead of a classroom, we considered a sub-region containing 900 students whose mean score was 99, nearly the same z-score and p-value would be observed. This shows that if the sample size is large enough, very small differences from the null value can be highly statistically significant. See statistical hypothesis testing for further discussion of this issue.

15.4. Z-TEST OTHER THAN LOCATION TESTS

Location tests are the most familiar Z-test. Another class of Z-test arises in maximum likelihood estimation of the parameters in a parametric statistical model. Maximum likelihood estimates are approximately normal under certain conditions, and their asymptotic variance can be calculated in terms of the Fisher information. The maximum likelihood estimate divided by its standard error can be used as a test statistic for the null hypothesis that the population value of the parameter equals zero. More generally, if $\hat{\theta}$ is the maximum likelihood estimate of a parameter θ, and θ_0 is the value of θ under the null hypothesis.

$$\left(\hat{\theta} - \theta_0\right) / \operatorname{SE}\left(\hat{\theta}\right)$$

can be used as a Z-test statistic.

When using a Z-test for maximum likelihood estimates, it is important to be aware that the normal approximation may be poor if the sample size is not sufficiently large. Although there is no simple, universal rule stating how large the sample size must be to use a Z-test, simulation can give a good idea as to whether a Z-test is appropriate in a given situation.

Z-test are employed whenever it can be argued that a test statistic follows a normal distribution under the null hypothesis of interest. Many non-

parametric test statistics, such as U statistics, are approximately normal for large enough sample sizes, and hence are often performed as Z-test.

15.5 T-TEST

A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution.

15.6 HISTORY OF T-TEST

The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin. Ireland ("Student" was his pen name). Gosset had been hired due to Claude Guinness's policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness's industrial processes. Gosset devised the t-test as a cheap way to monitor the quality of stout. The Student's Hest work was submitted to and accepted in the journal Biometrika and published in 1908. Company policy at Guinness forbade its chemists from publishing their findings, so Gosset published his mathematical work under the pseudonym "Student". Guinness had a policy of allowing technical staff leave for study (so-called "study !eave"), which Gosset used during the first two terms of the 1906-1907 academic year in Professor Karl Pearson's Biometric Laboratory at University College London. Gosset's identity was then known to fellow statisticians and to editor-in-chief Karl Pearson. It is not clear how much of the work Gosset performed while he was at Guinness and how much was done when he was on study leave at University College London.

15.7 USES OF T-TEST

Among the most frequently used t-test are :

- A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
- A two-sample location test of the null hypothesis such that the means of two populations are equal All such tests are usually called Student's t-test, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's t-test. These test are often referred to as "unpaired" or "Independent samples" t-

tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.

- A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired' or 'repeated measures" t-test: see paired difference test.
- A test of whether the slope of a regression line differs significantly from 0.

15.8 ASSUMPTIONS

Most t-test statistics have the form t = Z/s, where Z and s are functions of the data. Typically, Z designed to be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a scaling parameter that allows the distribution of t to be determined.

As an example, in the one-sample t-test $t = \frac{Z}{\left(s/\sqrt{n}\right)} = \frac{\left(\overline{X} - \mu\right)/\left(\sigma - \sqrt{n}\right)}{\left(s/\sqrt{n}\right)}$

where $\overline{X}$ is the sample mean of the data, n is the sample size, and s is the sample standard deviation. σ is the population standard deviation of the data.

The assumptions underlying a t-test are that

- $\left(\overline{X} - \mu\right)/\left(\sigma - \sqrt{n}\right)$ follows a standard normal distribution under the null hypothesis.
- S^2 follows a X^2 distribution with p degrees of freedom under the null hypothesis, where p is a positive constant.
- Z and s are independent.

In a specific type of t-test. these conditions are consequences of the population being studied, and of the way in which the data are sampled. For example, in the t-test comparing the means of two independent samples, the following assumptions should be met:

- Each of the two populations being compared should follow a normal distribution. This can be tested using a normality test, such as the Shapiro-Wilk or Kolmogorov-Smirnov test, or it can be assessed graphically using a normal quantile plot.

- If using Student's original definition of the t-test, the two populations being compared should have the same variance (testable using F-test, Levene's test, Bartlett's test, or the Brown-Forsythe test; or assessable graphically using a Q–Q plot). If the sample sizes in the two groups being compared are equal, Student's original t-test is highly robust to the presence of unequal variances. Welch's t-test is insensitive to equality of the variances regardless of whether the sample sizes are similar.
- The data used to carry out the test should be sampled independently from the two populations being compared. This is in general not testable from the data, but if the data are known to be dependently sampled (i.e., if they were sampled in clusters), then the classical t-tests discussed here may give misleading results.

15.9 UNPAIRED AND PAIRED TWO-SAMPLE T-TEST

Two-sample t-test for a difference in mean involve independent samples or paired samples. Paired t-test are a form of blocking, and have greater power than unpaired test when the paired units are similar with respect to "noise factors" that are independent of membership in the two groups being compared. In a different context, paired t-test can be used to reduce the effects of confounding factors in an observational study.

Independent (Unpaired) samples

The Independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained. one from each of the two populations being compared. For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomly assign 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the t-test The randomization is not essential here — if we contacted 100 people by phone and obtained each person's age and gender, and then used a two-sample t-test to see whether the mean ages differ by gender, this would also be an independent samples t-test, even though the data are observational.

Paired samples

Paired samples t-test typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" t-test).

A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control. That

way the correct rejection of the null hypothesis (here: of no difference made by the treatment) can become much more likely, with statistical power increasing simply because the random between-patient variation has now been eliminated. Note however that an increase of statistical power comes at a price: more tests are required, each subject having to be tested twice. Because half of the sample now depends on the other half, the paired version of Student's t-test has only "n/2–1" degrees of freedom (with n being the total number of observations). Pairs become Individual test units, and the sample has to be doubled to achieve the same number of degrees of freedom.

A paired samples t-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors.

Paired samples t-test are often referred to as "dependent samples t-test".

Calculations

Explicit expressions that can be used to carry out various t-test are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a t-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed or two-tailed test.

Once a t value is determined, a p-value can be found using a table of values from Student's t-distribution. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis is rejected in favour of the alternative hypothesis.

One-Sample T-Test

In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statistic.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test are n – 1. Although the parent population does not need to be normally distributed, the distribution of the population of sample means, $\overline{x}$, is assumed to be normal. By the central limit theorem, if the sampling of the

parent population is independent then the sample means will be approximately normal. (The degree of approximation will depend on how close the parent population is to a normal distribution and the sample Size, n.)

Slope of a regression line

Suppose one is fitting the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where x_i, i = 1, ..., n are known, α and β are unknown, and ε_i are independent identically normally distributed random errors with expected value 0 and unknown variance σ^2, and Y_i, i = 1, ..., n are observed. It is desired to test the null hypothesis that the slope β is equal to some specified value β_0 (often taken to be 0, in which case the hypothesis is that x and y are unrelated).

Let

$$\hat{\alpha}, \hat{\beta} = \text{least-squares estimators,}$$

$$SE_{\hat{\alpha}}, SE_{\hat{\beta}} = \text{the standard errors of least-squares estimators.}$$

Then,

$$t_{score} = \frac{\hat{\beta} - \beta_0}{SE_{\hat{\beta}}} \sim T_{n-2}$$

has a t-distribution with n – 2 degrees of freedom if the null hypothesis is true. The standard error of the slope coefficient:

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{y}_i\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}$$

can be written in terms of the residuals. Let

$$\hat{\varepsilon}_i = Y_i - \hat{y}_i = Y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right) = \text{residuals = estimated errors.}$$

$$SSR = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \text{sum of squares of residuals.}$$

Then t_{score} is given by :

$$\left(\hat{\beta} - \beta_0\right)\sqrt{n-2}$$

Independent Two-sample T-Test

Equal sample sizes, equal variance

This test is only used when both :

- The two sample sizes (that is, the number, n, of participants of each group) are equal;
- It can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{1}{2}\left(s_{X_1}^2 + s_{X_2}^2\right)}$$

Here $s_{X_1X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two, $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of t is the standard error of the difference between two means.

For significance testing. the degrees of freedom for this test is 2n – 2 where n is the number of participants in each group.

Equal or unequal sample sizes, equal variance

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute n for n_1, and n_2).

$s_{X_1X_2}$ is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, n = number of participants, 1 = group one. 2 = group two, n – 1 is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 – 2$) is the total number of degrees of freedom, which is used in significance testing.

Equal or unequal sample sizes, unequal variances

This test, also known as Welch's t-test, is used only when the two population variances are not assumed to be equal (the two sample sizes mayor may not be equal) and hence must be estimated separately. The t statistic to test whether the population means are different is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

s_{\overline{X}_1–\overline{X}_2} = \sqrt{{s_1^2 \over n_1}+{s_2^2 \over n_2}}.

Here s^2 is the unbiased estimator of the variance of the two samples, n_i = number of participants in group i, i=1 or 2. Note that in this case $s_{\overline{X}_1 - \overline{X}_2}$ is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as an ordinary Student's t distribution with the degrees of freedom calculated using

$$\text{d.f.} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1 - 1) + \left(s_2^2/n_2\right)^2/(n_2 - 1)}$$

This is known as the Welch-Satterthwaite equation. The true distribution of the test statistic actually depends (slightly) on the two unknown population variances (see Behrens-Fisher problem).

Dependent T-Test for paired samples

This test is used when the samples are dependent, that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test.

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

For this equation, the differences between all pairs must be calculated. The pairs are either one person's pre-test and post-test scores or between pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The average (X_D) and standard deviation (s_D) of those differences are used in the equation. The constant μ_0 is non-zero if you want to test whether the average of the difference is significantly different from μ_0. The degree of freedom used is n – 1.

Example of repeated measures			
Number	**Name**	**Test 1**	**Test 2**
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%
Example of matched pairs			
Pair	**Name**	**Age**	**Test**
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Worked examples

Let A_1 denote a set obtained by taking 6 random samples out of a larger set :

$A_1 = \{30.02,\ 29.99,\ 30.11,\ 29.97,\ 30.01,\ 29.99\}$

and let A_2 denote a second set obtained similarly :

$A_2 = \{29.89,\ 29.93,\ 29.72,\ 29.98,\ 30.02,\ 29.98\}$

These could be, for example. the weights of screws that were chosen out of a bucket.

We will carry out tests of the null hypothesis that the means of the populations from which the two samples were taken are equal.

The difference between the two sample means, each denoted by $\overline{X}_i$, which appears in the numerator for all the two-sample testing approaches discussed above, is

$$\overline{X}_i - \overline{X}_2 = 0.095$$

The sample standard deviations for the two samples are approximately 0.05 and 0.11, respectively. For such small samples, a test of equality between the two population variances would not be very powerful. Since the sample sizes are equal, the two forms of the two sample t-test will perform similarly in this example.

Unequal variances

If the approach for unequal variances (discussed above) is followed, the results are

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \approx 0.0485 \qquad \text{and} \qquad \text{df} \approx 7.03.$$

The test statistic is approximately 1.959. The two-tailed test p-value is approximately 0.091 and the one-tailed p-value is approximately 0.045.

Equal variances

If the approach for equal variances (discussed above) is followed, the results are

$$S_{X_1X_2} \approx 0.084$$

and

$$\text{df} = 10.$$

Since the sample sizes are equal (both are 6), the test statistic is again approximately equal to 1.959. Since the degrees of freedom is different from what it is in the unequal variances test, the p-values will differ slightly from what was found above. Here, the two-tailed test p-value is approximately 0.078, and the one-tailed p-value is approximately 0.039. Thus If there is good reason to believe that the population variances are equal, the results become somewhat more suggestive of a difference in the mean weights for the two populations of screws.

Alternatives to the T-Test for location problems

The t-test provides an exact test for the equality of the means of two normal populations with unknown, but equal, variances (The Welch's t-test is a nearly exact test for the case where the data are normal but the variances may differ,) For moderately large samples and a one tailed test, the t is relatively robust to moderate violations of the normality assumption.

For exactness, the t-test and Z-test require normality of the sample means, and the t-test additionally requires that the sample variance follows a scaled X^2 distribution, and that the sample mean and sample variance be statistically independent Normality of the individual data values is not required if these conditions are met. By the central limit theorem, sample means of moderately large samples are often well-approximated by a normal distribution even if the data are not normally distributed. For non-

normal data, the distribution of the sample variance may deviate substantially from a X^2 distribution. However, if the sample size is large, Slutsky's theorem implies that the distribution of the sample variance has little effect on the distribution of the test statistic. If the data are substantially non-normal and the sample size is small, the t-test can give misleading results. See Location test for Gaussian scale mixture distributions for some theory related to one particular family of non-normal distributions.

When the normality assumption does not hold, a non-parametric alternative to the t-test can often have better statistical power. For example, for two independent samples when the data distributions are asymmetric (that is, the distributions are skewed) or the distributions have large tails, then the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) can have three to four times higher power than the t-test. The nonparametric counterpart to the paired samples t-test is the Wilcoxon signed-rank test for paired samples. For a discussion on choosing between the t-test and nonparametric alternatives, see Sawilowsky (2005).

One-way analysis of variance generalizes the two-sample t-test when the data belong to more than two groups.

15.10. MULTIVARIATE TESTING

A generalization of Student's t statistic, called Hotelling's T-square statistic, allows for the testing of hypotheses on multiple (often correlated) measures within the same sample. For instance, a researcher might submit a number of subjects to a personality test consisting of multiple personality scales (e.g. the Minnesota Multiphasic Personality Inventory). Because measures of this type are usually positively correlated, it is not advisable to conduct separate univariate t-tests to test hypotheses, as these would neglect the covariance among measures and inflate the chance of falsely rejecting at least one hypothesis (Type I error). In this case a single multivariate test is preferable for hypothesis testing. Fisher's Method for combining multiple tests with alpha reduced for positive correlation among tests is one. Another is Hotelling's T^2 statistic follows a T^2 distribution. However, in practice the distribution is rarely used, since tabulated values for T^2 are hard to find. Usually, T^2 is converted instead to an F statistic.

One-Sample T^2 test

For a one sample multivariate test, the hypothesis is that the mean vector (μ) is equal to a given vector ({\mathbf\mu_0}). The test statistic is Hatelling's T^2:

$$T^2 = n(\overline{x} - \mu_0)' S^{-1} (\overline{x} - \mu_0)$$

where n is the sample size, $\overline{x}$ is the vector of column means and S is a $m \times m$ sample covariance matrix.

Two-Sample T^2 test

For a two-sample multivariate test, the hypothesis is that the mean vectors (μ_1, μ_2) of two samples are equal. The test statistic is Hotelling's 2-sample T^2:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{x}_1 - \bar{x}_2\right)' S_{spooled}^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$$

15.11. CONCLUSION

The explicit expressions that can be used to carry out various T-Tests and Z-test are analysed in detail in this chapter (unit) T-Test can be used to determine if two sets of data are significantly different from each other and is most commonly applied when the Test Statistic would follow a normal distribution if the value of a scaling term in the last scaling term in the test statistic were know. When the scaling term is unknown and is replied by any estimate based on the data that test statistic follows a student distribution.

While Z-Test is any statistical test for which the distribution of the test statistics under null hypothesis can be approximated by a normal distribution because of the central limit theorem. It has single critical value.

15.12 FURTHER STUDY

1. Sprinthal, R.C. (2011) : Basic Statistical Analysis, 9th Edition. Pearson Education Group.

UNIT-16 USES OF ICT IN RESEARCH METHODOLOGY

Objectives

After going through this unit you should be able to know about the–

16.1. Introduction

16.2. The Philosophy of qualitative data analysis

16.3. Qualitative data analysis

16.4. Computers in qualitative data analysis

16.5. Conclusion

16.6. Further study

16.1. INTRODUCTION

It has long been the custom to make use of new technological developments In easing the burden of complex or routine tasks. This is as true for research as it is for any other aspect of human activity. Thus one finds, for example, that over the years typewriters, word processors and computers generally have come to be adopted as part of the essential hardware of research.

By and large this is a process to be welcomed. If a labour or time saving technological artefact is available then there seems little to be gained by eschewing its use. Nevertheless, in the field of qualitative research, which for the purposes of this paper we are taking to mean research utilising linguistic data derived from interviews or similar conversational settings there are areas, we feel, where the untrammelled use of computer technology, specifically qualitative data analysis software, may do little to enhance the quality and value of the findings they produce.

In elaborating on this position we consider the philosophical foundations that underpin the practice of qualitative research. These, we argue, make use of a worldview that is contrary to the philosophical orientation of the positivistic science that has helped develop computer technology qualitative research aims to uncover meanings as they are apparent to an individual respondent; quantitative research relies on aggregation, quantification and categorisation as an adequate method to arrive at a scientific truth. In quantitative research there is a congruence between the underlying philosophies of the research and its analysis and the computer technology employed to assist with this. For example, statistical analysis in quantitative research has become a fast and routine process with many different pieces of software available to support this.

Software packages are now available to assist with me analysis of qualitative data which on the surface promise the same routinisation and speed benefits for the user as those available for quantitative analysis. Our argument is that qualitative data are derived from language and allow for the detailed exploration of feelings, drives, emotions and the subjective understanding a respondent had of a certain social situation at a particular point in time. They are indexical and context bound. The data are fuzzy, with slippery boundaries between meanings, and not ideally suited to categorisation and classification using digitally based software. Employing a digital tool of this type on qualitative data has the potential to distort any understanding arrived at.

16.2. THE PHILOSOPHY OF QUALITATIVE DATA

There are fundamental differences between the philosophies which on the one hand underpin information and communication technology (ICT) and on the other the philosophical thinking behind qualitative research. Computing technology assumes a positivistic approach to the natural world that sees it as being composed of objects that humans can study, understand and manipulate. It is a view that finds acceptance amongst quantitative researchers. Within sociology, generally, this positivistic orientation encompasses the idea that everything in society is amenable to being numbered, counted, measured or otherwise quantified (SPENCER 1971; SWINGEWOOD 1991) and that there is a particular process (copied largely from the natural sciences) that allows true understanding to be arrived at. When looked at from this perspective, society comes to be seen as something external to the people who inhabit it and who in turn find their behaviour controlled and influenced by it (LAYDER 1994). Human behaviour, the complex patterns of social interaction, then becomes a reflection of the macro level structure. All observed phenomena, when aggregated together and quantified, can be related back to the macro structure for analysis and understanding.

Qualitative research, and qualitative researchers, approach the world from a different perspective and set of understandings from quantitative researchers. Qualitative research is largely rooted in an understanding of the social world that sees human action as being the force that creates what we perceive to be society (STREUBERT & CARPENTER 1995), it is grounded in a humanist. phenomenological understanding of social action. The humanistic approach, common to much qualitative research, gives primacy to action over structure (LAYDER 1994). It becomes the goal of qualitative researchers therefore to try and see things from the perspective of the human actors. This is in direct contrast to the thinking of the positivistic schools where the external society is seen to shape human action.

Generally, in qualitative research there is less acceptance of the argument that it is the existence of an objectified society that constrains,

shapes and governs how people think and act. Because of this reduced emphasis on structure good understanding of the social world is not going to be achieved through the objective classifying and quantifying of observed phenomena; understanding social phenomena can only be achieved by accessing the meaning as it existed for the participants (PARKIN 1982, MORSE 1989; HOLLOWAY & WHEELER 1996). This is not necessarily to say that there is some kind of absolute prohibition on using qualitative methods if one takes the view that an external society is responsible for patterning and constraining actions and human behaviour. It is more that there is for those undertaking research an elective affinity between the adoption of a perspective on the location of the causal forces in society and the research paradigm to be employed in investigating them. For researchers of a phenomenological bent it follows more naturally to incline to qualitative research methods because of their focus on the individual. One consequence of this phenomenological approach is a greater sensitivity on the part of qualitative researchers to the ambiguities and subtle shades of interpretative meaning that social interaction can have for its participants (HOLLOWAY & WHEELER 1996). There is a recognition of the richness and complexity in human social interaction and an acceptance that this may not be amenable to quantification.

A qualitative approach may be used when little is known about a subject and the researcher may have few pre-conceived ideas about the subject or about the data, which will be gained. The aim is more likely to be inductive (that is moving towards theory) rather than testing theory.

Within the qualitative approach to social research and evaluation there are many different methods of collecting data resulting in many different types of qualitative data. With the focus on the lived experience of the individual, qualitative approaches are most suitable when the aim of the research is to understand and explore people's views, beliefs and experiences. To address such an aim, data are primarily linguistic; they may be textual or audio-visual and can be derived from, for example, interviews, observation, documents, diaries, field notes, which in turn may come from both primary and secondary sources interviews, of different levels of structure, are widely used methods and it is interview data and its analysis that this paper address. The discussion also has application to the more in-depth and less structured approaches of narrative and (audio) conversation analysis. Indeed, narrative and conversation analyses are approaches which illustrate the inductive, interpretive nature of qualitative data.

The importance to qualitative research of what Mead called the "significant symbol" (CRAIB 1984, P 73), or language, cannot be overstated Human languages are complex yet at the same time flexible, being capable of describing and representing a vast range of social situations and responses (GADAMER 1989, MACANN 1993). It is language that gives humans the experience of their "being-in-the-world" (GADAMER 1976, p. 3). Yet the complexity and ambiguity of language is

not given full recognition in quantitative research. There language is used uncritically, for example, on questionnaires, without thinking deeply about what it is or how It works or how it allows the world to be constituted and made use of (GADAMER 1976). So although both quantitative and qualitative researchers use data that are language based, for the quantitative researcher the use of language is not In Itself a problem or something that needs to be questioned. Quantitative researchers, arguably, tend to view language as a tool that can, with appropriate safeguards, be called upon to do a particular job in the same predictable and reliable way that a computer program might be calculate a statistical measure.

It is important for qualitative researchers to keep interview data in the context in which it was gathered and to preserve the respondents use of their own language to protect, as far as possible, the original meaning expressed through the data.

16.3. QUALITATIVE DATA ANALYSIS

The characteristics and heterogeneity of qualitative data translate Into challenges in analysis (LEE & FIELDING 1991, POUT & HUNGLER 1991 RICHARDS & RICHARDS 1998) particularly when viewed in stark contrast to the structured, numerical nature of quantitative data. That there are differing ontological and epistemological assumptions between qualitative and quantitative research does have profound implications for data analysis.

Quantitative data can be subjected to statistical analysis (contingent upon adequate knowledge of which tests to perform and how to Interpret the results) and clearly displayed in tabular or graphical form to address pre-determined hypotheses Contrast this with qualitative data analysis which is essentially although not entirely a hermeneutic enterprise, attempting to interpret the expressed experiences views and beliefs of people in social situations and then making that interpretation available to the research community. For both quantitative and qualitative researchers it is Important that the manner and techniques of analysis do not, to a greater extent than can be avoided, distort or corrupt the data. Although not addressed here, it is acknowledged that both qualitative and quantitative data can be collected in a single study.

One particular analytical challenge In qualitative research which involves the spoken word is posed by the centrality of language, its meaning and context. Making sense of a speech utterance is more than just effecting a mental translation of the words. In much of everyday social interaction and the speech that it generates there is a high degree of indexicality (LAYDER 1994, p.83), that is, a determination of the meaning given to speech utterances by the context in which they are uttered (GADAMER 1976, HOLLOWAY & WHEELER 1996). For a speech utterance to retain the meaning that it had at the time it was uttered (assuming that it is possible to ascribe a single meaning to a piece of speech with any degree

of absolute certainty) then it must be seen in the context of the surrounding speech and comments (and ideally the body language and non-verbal communication as well). Attempting to make sense of an utterance in isolation, without seeing it as part of a wider Whole, will be to lose an essential part of its meaning.

Whilst there is a multitude of data collection methods and sources of qualitative data, the focus here on the management and analysis of qualitative interview data can be simplified to a number of common activities and processes. A further key feature of qualitative research and evaluation is that rather than preceding analysis data collection is concurrent and interactive with data management and analysis (STRAUSS & CORBIN 1990, MILES & HUBERMAN 1994; BERG 2001). As such the generic activities and processes, summarised as follows, are not necessarily undertaken in a "linear" fashion.

- It is a reflective process with the researcher recording analytical notes and 'memos' (STRAUSS 1987; TESCH 1990; MILES & HUBERMAN 1994).
- Categories (themes) are derived from the data (BOULTON & HAMMERSLEY 1996; RICHARDS & RICHARDS 1998).
- Units of data are coded and annotated (STRAUSS 1987; STRAUSS & CORBIN 1990).
- The data coded are compared and contrasted (GLASER & STRAUSS 1967; STRAUSS & CORBIN 1990).
- Associations and patterns are identified and explored between categories (TESCH 1990; DEY 1993).
- The aim is a higher level synthesis (TESCH, 1990) perhaps moving towards theory generation and testing (MILES & HUBERMAN 1994).

These activities fit most accurately with the elements of grounded theory (GLASER & STRAUSS 1967) or theory building approaches to qualitative data analysis (MILES and HUBERMAN 1994) With some application to content analysis (CAVANAGH 1997; BERG 2001).

Whilst these common activities can be identified, qualitative data analysis is not prescriptive and precise details of how they are executed, for example, how to define categories and code data, identify relationships and explore theory cannot be specified. This is largely due to the variety of types of qualitative data and methods of data collection as well as the understanding that categories, and possibly hypotheses or theories, emerge from the data rather than being imposed upon it and that interpretation and creativity are required from the researcher.

The stark contrast between a purely quantitative and a purely qualitative approach illustrates the different approaches to collecting and analysing data. A study which required the use of solely quantitative data could

proceed in a more linear fashion and. although exploratory data analysis may take place before data collection is complete, any findings or reflections would not feed into data collection (STRAUSS 1987). Quantitative research, especially the questionnaire survey, is often likely to be deductive as opposed to inductive in approach and be focused on testing one or more pre-set hypotheses although this is obviously not always the case. Nevertheless, even when the research is not a typical example of the positivistic or experimental quantitative ideal, it still contains a high degree of pre-determined structure. For example, the areas to be explored during analysis will already have been determined and the main variables for analysis are defined through the questions.

HUBERMAN and MILES (1998) define (qualitative) data management as "the operations needed for a systematic, coherent process of data collection, storage and retrieval" necessary to enable the researcher to keep track of the volume of data, to flexibly access and use the data and to document the analytical process Data analysis can be defined as consisting of three concurrent elements data reduction, data display and conclusion drawing and verification (HUBERMAN &. MILES 1998).

The non-linear nature of data collection management and report writing in the qualitative tradition mean that all stages of a qualitative research project link into the exploration and interpretation of the data (WEITZMAN & MILES 1995).

16.4. COMPUTERS IN QUALITATIVE DATA ANALYSIS

The first and foremost point to make about the use of computers In qualitative analysis is that computers do not and cannot analyse qualitative data The fact that we have seen a development of computer-aided qualitative data analysis software (CAQOAS) should not be surprising given the widespread development and accessibility of ICT. However the use of ICT for the analysis of qualitative data remains a contentious issue and one which has not been universally and unquestioningly embraced (LEE & FIELDING 1991, MORISON & MOIR 1998). Computer techniques of logic and precise rules are not compatible with the unstructured, ambiguous nature of qualitative data and so it may distort or weaken data (BECKER 1993; KELLE 1995; RICHARDS & RICHARDS 1998) or stifle creativity. The nature of qualitative research in terms of the volume and complexity of unstructured data and the way in which findings and theory emerge from the data also makes software packages, developed to manage and analyse such data, difficult to become familiar with and use adequately.

The argument here is that it is not realistic, nor true to the purpose of qualitative research, to expect a social phenomenon described in language by the participants themselves, to be broken up quantified and analysed in a meaningful way by a tool based on a positivistic orientation

to the social and natural worlds. Of course, quantifying, categorising, and breaking up the data is possible and is a legitimate part of the analysis process at least insofar as some general high level sorting is concerned. The issue is more the extent to which the researcher is going to lose or distort the meaning that the social phenomenon had by attempting the interpretative process in the same way.

Computer technology and programs are we would argue, philosophically suited to analysing inanimate objects and matter, but not social phenomena expressed through the medium of language and uncovered by qualitative research techniques. If one takes technological artefacts, such as computers and computer programs, and then applies them to research and data analysis. this grounding in a positivistic philosophical background is going to fit them to certain tasks more than others. For research activities where measuring and counting are deemed to be essential to the analysis, as in quantitative research (itself an expression of a positivistic orientation to the social world), a device that can assist with that activity is going to be well matched.

It would be foolish and almost Luddite to argue that in the 21st century computers have no part to play in the process of qualitative data analysis. However, interpreting the complex meanings that social interactions and language can have, where they are coloured into many shades of grey is not going to be achieved by forcing the analysis into using pre-defined analytical categories. Qualitative data, i.e., the conversational/linguistic material that we are concerned with here has what could be described as almost an "analogue" feel to it inasmuch as it is, when first encountered by the researcher, essentially formless raw material. By subjecting it to a process of quantitative digitisation, to square off its shape and straighten its rounded edges through pushing it into a set of pre-defined categories it is inevitable that part of the original meaning is going to be either lost or changed.

This is not to say of course that an analytical approach that is not based on computers is going to leave the data in pristine condition and uncontaminated by the analyst. Any act of analysis is going to be influenced by the distance that a text stands from the original speaker or writer (GADAMER 1976). But, an approach that is not dependent on a digital logic system is going to be more sympathetic, to be more accepting of the quirks and inconsistencies inherent in any human social behaviour than one which is based on digital logic. To that extent an understanding that could present the lived experience of "the-being-in-the-world would be better achieved without the intervention of a computer.

The argument here is that ICT has definite application with many of the routine or mechanical tasks of qualitative research. However there remain difficulties and reservations regarding its widespread application in the stages of analysis which require understanding such as the development of themes assigning codes to the data and proposing and testing theoretical concepts. That is, although ICT can be of assistance in many of the data collection, management, storage and retrieval tasks, "the

central analytical task in qualitative research—understanding the meaning of texts—cannot be computerized" (KELLE 1995,). There are, as we have outlined, philosophical and methodological arguments against applying ICT to the analysis of qualitative data. Quantitative data analysis and the production of statistics, on the other hand, has been transformed by developments in ICT.

The most widely cited advantages of CAQDAS are that time may be saved (LEE & FIELDING 1991; MOSELEY, MEAD & MURPHY 1997), the analysis of larger data sets may be possible (KELLE & LAURIE 1995; BOWLING 1997, WEBB 1999) and that claims to making qualitative data more "scientific" can be made (CONRAD & REINHARZ 1984; RICHARDS & RICHARDS 1991; KELLE & LAURIE 1995; WEBB 1999). Some authors argue that as time can be saved and management of data is less cumbersome, the researcher concentrates more on the creative and interpretative tasks (RICHARDS & RICHARDS 1991; MORISON & MOIR 1998) thus leading to more substantive analySIS (MOSELEY et. al. 1997) or enhanced quality analysis (CONRAD & REINHARZ 1984; TESCH 1990; LEE & FIELDING 1991).

The debate surrounding the application of ICT to qualitative data analysis inevitably involves the discussion of both positive and negative effects. As stated by the developers of NUD*IST "the computer method can have dramatic implications for the research process and outcomes, from unacceptable restrictions on analysis to unexpected opening out of possibilities" (RICHARDS & RICHARDS 1998, The argument here addresses both restriction and opportunity.

Data reduction Incorporating the management storage and cataloguing of data can be made more efficient and more manageable by the use of ICT because of the speed and sophistication with which computer packages work Audio taped data, hand written notes and summaries can be typed up edited saved in different formats and reproduced, as well as made available to relevant others. quickly and easily using word processors. For large projects, the potential volume of data and other information can be organised and managed more efficiently and conveniently (ROBSON 1993, KELLE 1995) in order to prevent "data overload" (MILES & HUBERWN 1994). This includes enabling the researcher for the duration of the project, to record, store and retrieve empirical data field notes, emerging ideas, analytical memos and references whether using word processors or CAQDAS. Data overload, that is, "limitations on the amount of data that can be dealt with (too much to receive, process and remember)" (ROBSON 1993,) is suggested here to be one of the deficiencies of the human as analyst which can be addressed by ICT.

Whilst the mechanistic tasks or "routine elements" can be greatly supported by the use of ICT (BROWN, TAYLOR, BALDY, EDWARDS & OPPENHEIMER 1990), it is the activities which require human thought

processes, interpretation, creativity and reflection which are most difficult to reconcile with ICT.

Most analysis of qualitative data Involves the allocation of categories or themes to sections of data usually via coding to enable subsequent re1rieval exploration and theory building. Adopting a purely Inductive approach to data collection and analysis would mean all categories emerging from the data whilst a purely deductive approach would mean that all categories were pre-determined The reality of categorising qualitative data is likely to be that some categories are determined before data collection ("coding down") whilst most emerge during data collection and management ("coding up") (BERG 2001). The first stages of developing categories will result in a large number with the general rule being to include rather than exclude. As the project continues, categories may be modified, merged, deleted or renamed.

A crucial practice to enable themes to emerge from unstructured data, for memos to be recorded, for codes to be assigned and for patterns to be noted and explored is the researcher gaining closeness to the data (BOULTON & HAMMERSLEY 1996) by immersion in it (ABRAHAMSON 1983; HAMMERSLEY & ATKINSON 1995; STREU8ERT & CARPENTER 1995). It is because the analyst is a human, with the ability to relate to other humans that the complex blending of speech forms and context can be put back together in such a way that understanding results immersion in the data also allows the researcher to keep the data in their original context.

Word processors, like mechanised index cards, punched cards or filing systems greatly improve and make more efficient, the traditional "cut and paste" method of coding and retrieving information. That is, once data has been thoroughly coded manually by the researcher the word processor cut and paste functions can be used to create separate files for all data coded according to each category. Memos and notes can also be added as appropriate. This allows all data relevant to each code to be printed and examined or even pasted into published output it is also possible to use in-built word processor facilities to "find", "edit" "go to" for searching data for coding (MILES & HUBERMAN 1994; WEITZMAN & MILES 1995; COOMBES 2001) at a very basic level.

Using simple cutting and pasting does however pose a methodological problem for qualitative researchers in that the text is removed from its context (KELLE 1995) and different word processor files must be accessed to view the full data.

These basic functions supported by word processing packages will be sufficient for those qualitative researchers whose data set is small or whose research or evaluation simply requires a description or overview of key themes from the data. Data management, storage, assigning categories and recording memos can equally all be undertaken using CAQDAS (WEITZMAN & MILES 1995).

Automated coding can be undertaken by CAQDAS In a much more advanced and flexible way by using string searches. NUD*IST and NUD*IST Vivo (NVivo) permit the inclusion and exclusion of data in searches (GAHAN & HANNIBAL 1998; RICHARDS 1999; BAZELEY & RICHARDS 2000; GIBBS 2002) for example according to descriptive data or cooing already assigned to the data set. Text searches with different levels of specificity can be performed, wild cards (*) or searching for words with similar meaning or usage can be used to introduce flexibility, whilst Boolean searches ("and", "or" "not") and proximity searches (to find text near other text) allow more sophisticated and precise searching. It has been argued that new codes are easier to include when using ICT as the process is less time consuming and automated searches are easy to perform.

Maintaining the richness and in depth understanding and meaning of data in its original context are key features of qualitative research. It is acknowledged that results of searches can only be as good as the commands entered but real concerns exist around the true meanings of words and phrases and their being missed or coded incorrectly and the richness of experience and explanation being lost or taken out of context. This is largely because automated searching can only be based on lexical as opposed to semantic analysis of text (MOSELEY et al. 1997). Caution must be exercised around words having more than one meaning for example as a noun and verb as in "nurse" or "train" and each find must be checked for relevance before assignment to a category.

NVivo has the advantage over word processors in that it allows the researcher to easily view any coded data in its original context as well as alongside other data or memos coded in the same way. All data retrieved following coding contains an identifier of the original data source and retains links with the original data documents. In addition CAQDAS allow the modification of categories and coding (GAHAN & HANNIBAL 1998) much more easily than word processors and allow the list of categories to be readily viewed.

The argument here is that automated searching facilities using ICT should only be used to support, rattler than replace manual handling reading and re-reading and gaining familiarity with the data which is the essence of qualitative data analysis. Reading data on screen and not handling whole parts of the data set can be argued to distance (MORISON & MOIR 1998) or alienate (WEBB 1999) the researcher from their data. CAQDAS searching also risks overly mechanising the process and marginalising the reflection of the researcher (MORISON & MOIR 1998) thereby encouraging prescriptive analytical methods which inhibit interpretation and creativity (DEY 1993).

The centrality of coding to subsequent stages of analysis requires the thorough and accurate categorisation of all appropriate data. Getting to know the data thoroughly and coding according to human understanding are key elements of this process. Automated searching will not take the

place of additional searches and checking undertaken by another member of the research team.

The early CAQDAS concentrated on facilities to code text and search for occurrences of these codes (WEBB 1999) or to code and retrieve data (KELLE 1995). It is the subsequent stages of analysis, such as exploring patterns between categories and moving towards theory development, which underlie the true complexity and richness of qualitative data and one of the purposes of employing a qualitative approach. The aim is to interpret and draw meaning from the data.

Developments over the last decade in CAQDAS have seen these higher level functions also incorporated (WEITZMAN & MILES 1995, RICHARDS & RICHARDS 1998) to support the creative and interpretative activity of the researcher. Some advantages can be realised by CAQDAS but, as with searches for coding, the nature of language and the importance of context warn against over reliance on ICT. Familiarity with and closeness to the data are crucial for this higher level analysis and the same concerns exist around ICT distancing the researcher from the data and analysis becoming overly mechanised and prescriptive.

For qualitative researchers, a common activity as categories are identified and codes are assigned is for emerging patterns and relationships to be displayed graphically for example using tables, matrices or diagrams (STRAUSS & CORBIN 1990). That is, data are presented "as an organised compressed assembly of information that permits conclusion drawing and/or action taking" (HUBERMAN & MIL.ES 1998,) which may enable a new perspective on the emerging data. Displaying data in this manner may subsequently lead to further data collection or additional exploration of the data.

In the case of NVivo the software writers have opted for a hierarchical "tree" structure which displays the categories used for coding (RICHARDS 1999); a display which must be manipulated and explored to move towards theorising. The tree is modified by the researcher as analysis proceeds and It can function as a summary of the coding structure. Ways in which NUD*IST and NVivo can assist with theorising include exploring and testing the inter-relationships between categories through "index searching" (WEITZMAN & MILES 1995; GAHAN & HANNIBAL 1998). Patterns, associations and relationships can be suggested and explored in this way by using for example contextual (such as "followed by" or "near") or collation operators (such as "less", "overlap" or "union"). Such facilities however share similarities with the analysis of quantitative data with the emphasis on variables and causality which go against the purpose and value of quantitative, research.

NVivo can suggest areas for further exploration in this way, which may otherwise have been overlooked, but the researcher risks losing contact with the context and meaning of raw data by too much data manipulation by computer. The main concern is that the researcher may

not return to the original data with an open and questioning mind, or return as frequently as they may have done were they not using CAQDAS.

The restriction and opportunity posed to qualitative data analysis by ICT is apparent from this discussion. Qualitative data analysis is distinct from all other stages of the research process (both quantitative and qualitative) in that ICT also represents a restriction rather than just an opening of opportunity.

At all stages qualitative data can be organised, managed and manipulated effectively using ICT for exemple, storing and retrieving coded data and systematically searching patterns between categories. However, the emphasis on coding and the ease with which it can be undertaken pose a threat to the richness of qualitative data and the nuances of language and meaning. Coding data manually before using CAQDAS gains the advantage of applying human understanding to the raw data coupled with the efficiency of computer storage and retrieval. The problem with computer aided coding the ease and simplicity with which it can be undertaken, is the opportunities and temptations it offers to create more and more codes more discrete categories into which elements of the data are to be forced, without necessarily retaining sight of the larger whole. Creating and applying codes is not the same as analysis and indeed may only serve to break up and segment the data, fracturing the meaning that the integrated whole would have had. NVivo can also encourage and enable more complex manipulation and retrieval of data than is likely to be possible manually. Again, this is only the case once data has been thoroughly coded manually. However, it cannot give meaning to the data and is no substitute for gaining full familiarity with the data and for the researcher to adopt a questioning and exploratory approach.

Extending possibilities for example around larger data sets and more coding, should perhaps not be welcomed unquestioningly. The aim and purposes of the research must be the primary focus and the guide in data collection and analysis.

16.5. CONCLUSION

This paper has explored the applicability of ICT based analytical tools in qualitative research. It has been argued that given the philosophical differences between qualitative researchers and the science that develops these technologies this is not necessarily to the benefit of qualitative research. While there are some elements of the qualitative research process that can benefit from computer assistance the process of data analysis could be harmed by reliance on software packages. Such are the differences in the philosophies, we have argued that the original meaning inherent in the data could be distorted or lost. The employment of computer programs in qualitative data analysis is a practice that should be viewed with caution.

Analysing qualitative material that is based on speech or texts derived from interviews and conversations must have regard for the context and the integrated whole. Computer based systems to aid with analysis are, we would argue, based on the natural scientific view of the world that sees social phenomena as reflections of the higher level ordering of an objective social structure. The ideal data type here is one which is amenable to quantifying and segmentation into discrete categories as this allows for numerical manipulation and analysis. It is a world view that is not, we feel, sympathetic to the types of qualitative data that we are discussing here.

Speech derived data is rich data in the sense that it can encompass many meanings and requires careful reading with regard to the whole from which it is taken CAQDAS packages possess features that reflect their quantitative and positivistic heritage, particularly their facilities for creating and adding coding categories. Over-reliance on these features could lead to a fracturing of the data whole and a loss of meaning.

Researchers who make use of these packages must remain alert to the need to preserve the integrity and context of the original material and not lose sight of this during the process of coding and subsequent analysis.

16.6. FURTHER STUDY

1. Abrahamson, Mark (1983), Social research methods, New Jersey, Prentice Hall.

2. Berg, Bruce L. (2001), Qualitative research Methods for the Social Sciences (4th edition). Needham Heights, MA; Allyn & Bacon.

3. Bowling, Ann (1997). Research Methods in Health, Buckingham, Open University Press.

4. Coombes, Hilary (2001), Research using IT, Basingstoker, Palgrave.

5. Dey, Ian (1993), Qualitative Data Analysis : A user-friendly guide for social scientists, London : Routledge.